# AVERTISSEMENT



# LIENS

Code la Propriété Intellectuelle – Articles L. 122-4 et L. 335-1 à L. 335-10
Loi n°92-597 du 1er juillet 1992, publiée au *Journal Officiel* du 2 juillet 1992
http://www.cfcopies.com/V2/leg/leg-droi.php
http://www.culture.gouv.fr/culture/infos-pratiques/droits/protection.htm

Université de Toulouse

# THÈSE

**En vue de l'obtention du**

## DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

**Délivré par :** *l'Université Toulouse 1 Capitole (UT1 Capitole)*

---

**Présentée et soutenue le** *26/01/2018* **par :**

AURORE ARCHIMBAUD

---

**Méthodes statistiques de détection d'observations atypiques pour des données en grande dimension**

~

**Statistical methods for outlier detection for high-dimensional data**

---

**JURY**

| | | |
|---|---|---|
| BÉATRICE LAURENT-BONNEAU | Professeure | Présidente du Jury |
| JULIE JOSSE | Professeure | Membre du Jury |
| VALENTIN TODOROV | Senior Management Information Officer | Membre du Jury |
| ANNE RUIZ-GAZEN | Professeure | Directrice de thèse |
| KLAUS NORDHAUSEN | Assistant Professor | Co-Directeur de thèse |
| ANDREA CERIOLI | Professor | Rapporteur |
| JÉRÔME SARACCO | Professeur | Rapporteur |
| FRANÇOIS BERGERET | Dir. ippon innovation | Invité |
| CAROLE SOUAL | Ingénieure Statisticienne ippon innovation | Invitée |

---

**École doctorale et spécialité :**

*MITT : Domaine Mathématiques : Mathématiques appliquées*

**Unité de Recherche :**

*TSE-R, Toulouse School of Economics (UMR CNRS 5314 - INRA 1415)*

**Directeur(s) de Thèse :**

*Anne RUIZ-GAZEN* et *Klaus NORDHAUSEN*

**Rapporteurs :**

*Andrea CERIOLI* et *Jérôme SARACCO*

# Remerciements

Je tiens tout d'abord à remercier François Bergeret, dirigeant d'ippon innovation, sans qui cette thèse n'aurait jamais vu le jour. J'ai eu la chance de découvrir la recherche appliquée lors de mon stage de Master 1 au sein de son entreprise et c'est cette expérience qui m'a donné envie de poursuivre dans cette voie. Je le remercie de l'entière confiance qu'il m'a toujours accordée ainsi que de m'avoir donné l'opportunité de réaliser une thèse sur un projet de recherche appliquée dans le cadre d'une convention Cifre soutenue par l'ANRT.

Je suis également extrêmement reconnaissante à ma directrice de thèse, Anne Ruiz-Gazen, de m'avoir accordé sa confiance afin de prolonger ses propres recherches de thèse et de m'avoir fait partager ses idées très fructueuses tout en me laissant autonome. Je lui exprime toute ma gratitude pour son soutien et son investissement sans faille à mes côtés pendant ces trois années, et ce toujours dans la bonne humeur et la gentillesse, en me consacrant autant de temps que nécessaire. Je la remercie aussi de m'avoir accompagnée et guidée à chaque étape de la découverte de la vie très riche et diversifiée d'enseignante - chercheuse. Ce fut une véritable chance et un honneur pour moi que d'avoir pu bénéficier de la supervision d'une telle directrice de thèse.

Merci également à mon co-directeur de thèse, Klaus Nordhausen, qui a supervisé ce travail avec le soutien de l'Académie de Finlande et de COST. Son suivi avisé et sa collaboration régulière malgré la distance, notamment pour le développement des packages R, ont permis d'apporter et d'ancrer une ouverture encore plus internationale à ce travail de recherche. Je lui suis reconnaissante pour ces trois années de collaboration ainsi que pour son accueil lors de mes visites de travail en Finlande.

Je remercie Messieurs Andrea Cerioli et Jérôme Saracco qui ont accepté d'être les rapporteurs de cette thèse et qui ont relevé des points très pertinents. Je remercie aussi Mesdames Julie Josse et Béatrice Laurent-Bonneau ainsi que Monsieur Valentin Todorov de leur participation en tant qu'examinateurs à ce jury de thèse.

D'un point de vue industriel, je tiens à exprimer ma gratitude à toute l'équipe d'ippon innovation : à Carole Soual pour son soutien aussi bien technique que moral à tous les niveaux, notre collaboration a été très enrichissante pour moi ; à Bernard Azalbert pour son envie de partager son expertise des semi-conducteurs ; à Michel Gomez et à Christophe Delagarde pour leur précieuse aide en informatique ; à Laurence Bergeret pour son efficacité à se charger de la partie administrative et à tous les autres pour les bons moments. Merci également à l'équipe de Microchip-Atmel : à Sophie D'Alberto pour son implication au bon déroulement de ma thèse incluant la production de données exploitables, mais également la possibilité de rendre publics les résultats ; à Christian Bonin pour sa réactivité et son excellent travail d'intégration logicielle, et à tous les autres dont Thierry Thebault qui ont accepté de faire partie de cette expérience et de co-écrire un article ensemble. Enfin,

je remercie François Braud qui est l'initiateur de cette collaboration qui s'inscrit dans le projet européen RESIST.

J'en profite pour remercier tous les membres du laboratoire de TSE-R qui m'ont accueillie très chaleureusement et que j'ai appris à connaître au cours des séminaires et conférences. Merci à Thibault Laurent qui m'a introduite dans le groupe très sympathique des « Rencontres des ingénieurs statisticiens toulousains » et merci à tous ses membres pour les bons moments passés ensemble. Merci également au comité d'organisation des Rencontres R 2016 et de useR !2019, avec de superbes expériences toujours dans la bonne humeur. Et enfin mes remerciements vont au groupe des « Jeunes Statisticien.ne.s » de la SFdS pour m'avoir proposé d'embarquer avec eux pour de nouvelles activités.

Je remercie tous les chercheurs de la communauté ICS que j'ai rencontrés à l'occasion de conférences et de visites, dont les discussions et échanges ont été très intéressants et ont souvent inspiré mes recherches, notamment Henri Caussinus, précurseur de la méthode ICS avec Anne Ruiz-Gazen, ainsi que Hannu Oja, Frank Critchley et les membres de leurs équipes respectives : Sara Taskinen, Joni Virta, Markus Matilainen, Jari Miettinen, Radka Sabolova et Germain Van Bever. Merci également à Salvador Flores, pour sa collaboration et son aide pour proposer une nouvelle adaptation d'ICS. Enfin, je remercie Joris May qui, sous ma supervision, a developpé en grande partie l'application ICSShiny lors de son stage de Master 1 au sein du laboratoire TSE-R.

Au niveau des relations administratives, je remercie l'Université Toulouse 1 Capitole, le laboratoire TSE-R, l'Université de Turku en Finlande ainsi que COST qui ont financé plusieurs de mes déplacements en conférences et visites à l'étranger. Je suis également très reconnaissante à Christel Gilabert, Corinne Vella, Aline Soulie et Virginie Mangion, secrétaires à l'Université Toulouse 1 Capitole, ainsi que Martine Labruyère et Agnès Requis à l'EDMITT, qui ont accompagné très efficacement mes nombreuses démarches administratives en s'adaptant aux spécificités requises par ma convention Cifre.

Enfin, je suis extrêmement reconnaissante à tous mes ami(e)s, et à tous mes proches, qui sont restés auprès de moi pendant ces trois années. Leur soutien et leurs efforts de compréhension ont été précieux, notamment lorsque le temps me manquait. Merci pour tous ces encouragements et ces moments passés ensemble qui ont été indispensables à mon équilibre.

Une thèse est souvent vécue comme une expérience très personnelle, mais je reconnais avoir eu la chance d'avoir pu échanger avec beaucoup de personnes d'horizons très différents, et c'est cette diversité qui fait la richesse de ce travail de thèse.

# Table des matières

**Remerciements** **i**

**Notations** **vii**

**Acronyms** **ix**

**Introduction** **1**

**1 État de l'art : détection non-supervisée d'observations atypiques sur des données quantitatives avec le logiciel R** **9**
1.1 Introduction . . . 11
1.2 Contrôle de qualité . . . 12
1.2.1 Contexte des semi-conducteurs . . . 13
1.2.2 Les standards en univarié . . . 14
1.2.3 Les méthodes multivariées usuelles non-supervisées . . . 17
1.2.4 Évaluation des méthodes et contributeurs . . . 22
1.2.5 Exemple réel de l'industrie des semi-conducteurs et mise en œuvre en R . . . 24
1.3 Approches en dimension standard : $n > p$ . . . 27
1.3.1 Généralités . . . 27
1.3.2 Présentation succincte des méthodes . . . 28
1.4 Approches en grande dimension - faible taille d'échantillon (HDLSS) : $n < p$ 36
1.4.1 Le fléau de la dimension . . . 36
1.4.2 Les analyses en dimension globale . . . 38
1.4.3 Les analyses de sous-espaces de l'espace originel . . . 41
1.5 Conclusion et perspectives . . . 42

**2 Multivariate Outlier Detection with ICS** **45**
2.1 Introduction . . . 47
2.2 Behavior of the Mahalanobis distance in large dimension . . . 49
2.3 Invariant Coordinate Selection . . . 51
2.3.1 Scatter matrices . . . 51
2.3.2 ICS principle . . . 53
2.4 ICS implementation for outlier detection . . . 54
2.4.1 The choice of the scatter pair . . . 54
2.4.2 The invariant components selection . . . 55
2.4.3 Outlier identification . . . 56
2.5 Simulations . . . 56
2.5.1 Simulation framework . . . 56
2.5.2 Selecting the invariant components . . . 58

2.5.3 Detecting outliers with ICS . . . 59
2.5.4 Comparing ICS with the Mahalanobis distance and PCA . . . 60
2.6 Data Analysis . . . 62
2.6.1 Glass recycling . . . 63
2.6.2 Reliability Data . . . 64
2.6.3 High-tech parts . . . 65
2.7 Conclusion and perspectives . . . 66
2.8 Appendix . . . 68
2.8.1 Proof of Proposition 1 . . . 68
2.8.2 Derivation of the eigenvalues and eigenvectors of the simultaneous diagonalization of COV and $COV_4$ for particular mixtures . . . 71

**3 Unsupervised outlier detection with the R package ICSOutlier 79**
3.1 Introduction . . . 81
3.2 Invariant Coordinate Selection (ICS) for outlier detection . . . 82
3.2.1 Principle . . . 82
3.2.2 Invariant Coordinates Selection . . . 84
3.2.3 Measure of outlierness . . . 85
3.2.4 Outlier identification . . . 85
3.3 Using package ICSOutlier . . . 85
3.4 Examples . . . 87
3.4.1 Example with no outlier . . . 87
3.4.2 HTP data set . . . 88
3.4.3 Reliability data set . . . 92
3.4.4 HBK data set . . . 95
3.5 Conclusion and future developments . . . 97
3.6 Complements on Chapter 3: the ICSShiny package . . . 99

**4 ICS with positive semi-definite scatter matrices for data not in general position 101**
4.1 Introduction . . . 103
4.2 The classical ICS method . . . 104
4.2.1 Principle . . . 105
4.2.2 Interpretation . . . 105
4.2.3 Properties . . . 106
4.2.4 Other statistical methods solving a GEP of two scatter matrices . . . 107
4.3 ICS with semi-definite positive scatter matrices . . . 109
4.3.1 Challenges with singular scatter matrices: an example . . . 110
4.3.2 ICS with a Moore-Penrose pseudo-inverse . . . 112
4.3.3 A dimension reduction as preprocessing . . . 115
4.3.4 ICS with a Generalized Singular Value Decomposition . . . 119
4.4 GSVD on a practical case: an industrial example with collinearity . . . 125
4.4.1 Context . . . 126
4.4.2 Computation of the two scatter matrices COV and $COV_4$ . . . 126
4.4.3 Implementation of the GSVD procedure . . . 126
4.4.4 Some results . . . 127
4.5 Conclusion and perspectives . . . 129

**5 A new outlier detection solution for HDLSS data in an industrial context 131**
5.1 Context and objectives . . . 133

5.2 High dimensional outlier screening of small dice samples for Aerospace IC reliability . . . 134
5.2.1 Introduction . . . 134
5.2.2 Available Statistical tools . . . 135
5.2.3 Advanced Tools for High-Dimensional (HD) data . . . 137
5.2.4 Case study on space components . . . 138
5.2.5 Conclusion . . . 139
5.3 Challenges in the development . . . 140
5.3.1 Available data . . . 140
5.3.2 Numerical errors . . . 144
5.3.3 Methodological challenges . . . 148
5.4 Conclusion . . . 150

**Conclusion et Perspectives** **151**

**Bibliographie** **172**

# Notations

The transpose is denoted by a prime. The matrices and vectors are in bold.

| | |
|---|---|
| $\mathbf{X}$ | A $p$-multivariate real random vector. |
| $\mathbf{X}_n$ | A $p$-multivariate dataset containing $n$ observations ($n \times p$ matrix). |
| $\mathbf{x}_i$ | A $p$-multivariate vector associated to the observation $i$. |
| $n$ | Number of observations. |
| $p$ | Number of variables. |
| $\mu_n$ or $\bar{\mathbf{x}}$ | Empirical mean of the sample: $\mu_n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$ |
| $e_i$ | $i$-th column of the identity matrix. |
| $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ | multivariate Gaussian (normal) distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ |
| $P$ | Probability measure. |
| $\chi^2_k$ | Chi-square distribution with $k$ degrees of freedom. |
| COV | Variance-covariance statistics. |
| $\text{COV}_4$ | The so called scatter statistics of fourth moments. |
| MLC | The scatter statistics based on the maximum likelihood estimators of an elliptical Cauchy distribution. It belongs to the well-known class of M-estimators. |
| $\text{MCD}_\alpha$ | reweighted Minimum Covariance Determinant: reweighted empirical mean and covariance estimators of the MCD subset based on the $h \approx \alpha n$ observations whose covariance matrix has the smallest determinant. |
| rank | Rank of a matrix or an operator. |
| null | Null of matrix $\mathbf{A}$: vectors $\mathbf{x}$ such that $\mathbf{Ax} = 0$. |
| range | Range of a matrix: span of its column vectors. |
| dim | Dimension of a vector space. |
| $\propto$ | Proportional to. |
| $\oplus$ | Direct sum for vector spaces. |
| $\mathbb{E}$ | Expectation. |
| $\text{MD}^2_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\mathbf{x}_i)$ | Squared of the Mahalanobis distance of an observation $\mathbf{x}_i$ computed based on the location and scatter estimates $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. |

# Acronyms

**ABOD** Angle-based Outlier Detection. 29, 31, 38, 93, 94, 98

**ACP** French: *Analyse en Composantes Principales*. 2, 13, 20, 21, 23–26, 31–33, 38–40, 42

**AG** The Anscombe-Glynn test of kurtosis. 55, 85

**ANRT** Association Nationale Recherche Technologie. i

**BS** The Bonett-Seier test of Geary's kurtosis. 55, 85

**CATRENE** Cluster for Application and Technology Research in Europe on Nano-Electronics. 1, 5, 133, 135

**COST** European Cooperation in Science and Technology. i, ii

**CQI** Customer Quality Incident. 13, 22–27, 126, 127, 141, 144

**DA** The D'Agostino test of skewnes. 55, 58–62, 65–67, 85

**DR** Detection Rate. 23

**EDMITT** Ecole Doctorale Mathématiques, Informatique, Télécommunications de Toulouse. ii

**EM** Expectation-Maximisation. 30

**ESA** European Space Agency. 153, 155

**EVP** EigenValue Problem. 105, 108, 112, 117, 120, 124

**FAR** False Alarm Rate. 23, 25

**FDC** Fault Detection and Classification. 13

**FLDA** Fisher Linear Discriminant Analysis. 110

**FN** False Negative. 22, 23, 57

**FP** False Positive. 22, 23, 58–61

**FSRMCD** Finite-Sample Reweighted MCD. 90–92

**FT** Final Test. 14, 141, 143

**FWER** Family-Wise Error Rate. 18

**GAT** Good Average Testing. 41, 133–135, 137–140, 146, 149, 150

**GDBC** Good Die in Bad Cluster. 16, 136

**GEP** Generalized Eigenvalue Problem. iv, 102, 105, 107–110, 115, 117, 119, 121, 122, 124

**GM** Reference to the cut-off from Green and Martin (2017b). 61

**GPAT** Geographic Part Average Testing. 16, 136

**GSVD** Generalized Singular Value Decomposition. iv, 4, 7, 101, 102, 110, 121–129, 152, 155

**HDLSS** High Dimension - Low Sample Size. iii, iv, 2, 4, 6, 7, 9, 10, 12, 36, 38, 39, 41, 103, 104, 109, 110, 116, 131, 133, 142–144, 146, 148, 150–152, 154, 155

**HOS** High-dimensional Outlying Subspaces. 42

**HTP** High Tech Parts example. iv, 24, 26, 63, 66, 80, 88, 90–92, 98, 142

**IC** Invariant Component. 83, 106, 114–116, 121, 125

**ICA** Independent Component Analysis. 13, 108

**ICS** Invariant Coordinate Selection. iii, iv, 2, 3, 6, 7, 32, 33, 43, 45–51, 53–56, 58–67, 79, 80, 82–88, 91, 92, 95, 97–99, 101–104, 106–121, 125, 126, 129, 133, 149, 151, 152, 154, 155, 174

**IQR** Inter-quartile Range. 15

**JB** The Jarque-Bera test based on both skewness and kurtosis. 55, 85

**KPCA** Kernel Principal Component Analysis. 13

**LAPACK** Linear Algebra PACKage. 122

**LDA** Linear Discriminant Analysis. 107, 116, 122, 129

**LOF** Local Outlier Factor. 34, 35, 41, 81, 93, 94, 98

**LSL** Lower Specification Limit. 14, 15, 136

**MAD** Median Absolute Deviation. 22

**MCD** (reweighted) Minimum Covariant Determinant. 20, 22, 24–26, 39, 51, 53, 60, 64, 65, 82, 88, 90, 92, 95, 97, 103, 137, 141, 144

**MD** Mahalanobis Distance. 2, 5, 6, 9, 17, 30, 45, 60, 61, 81, 84, 151, 154

**MLE** Maximum Likelihood Estimator. 93

**MMR** Mathematical Methods in Reliability. 4, 7, 131

**MSP** French: *Maîtrise Statistique des Procédés*. 13, 17, 20

**NFP** Number of False Positive. 62, 64, 66

**NNR** Nearest Neighbor Residual. 16, 17, 136

**NTP** Number of True Positive. 62

**OD** Orthogonal Distance. 21–26, 39, 81, 95, 137, 139

**PA** Parallel Analysis. 55, 58–62, 65–67, 84

**PAT** Part Average Testing. 15, 16, 27, 135, 136, 138, 139

**PC** Principal Component. 20

**PCA** Principal Component Analysis. iv, 5, 6, 9, 31, 46–49, 53, 55, 60–62, 66, 67, 81, 84, 98, 107, 108, 115, 116, 137

**PCER** Per-Comparison Error Rate. 18

**PT** Parametric Test. 13, 14

**QSVD** Quotient Singular Value Decomposition. 122

**RD** Robust version of the Mahalanobis Distance. 60, 61

**RESIST** RESilient Integrated SysTems. ii, 1, 5, 133, 135, 140, 150, 151, 154

**ROBPCA** ROBust Principal Component Analysis. 32, 39, 40, 62, 95, 97, 134, 135, 137–140

**SD** Score Distance. 21–26, 39, 81, 95, 137, 139

**SFdS** Société Française de Statistique. ii

**SIR** Sliced Inverse Regression. 109

**SOD** Subspace Outlier Degree. 42

**SPC** Statistical Process Control. 13, 17, 20

**SPE** Square Projection Error. 20, 22

**SVD** Singular Value Decomposition. 40, 116, 146–148

**SW** The Shapiro-Wilk normality test. 55, 85

**TN** True Negative. 23

**TP** True Positive. 22, 23, 57–61

**TSE-R** Toulouse School of Economics Research. ii, 1, 5, 133, 174

**USL** Upper Specification Limit. 14, 15, 136

# Introduction

## Version française

Dans l'exercice de la statistique au sens large, il est nécessaire de prendre en compte la présence d'observations atypiques et ce pour différentes raisons. Tout d'abord, dans la pratique, il n'est pas rare d'avoir à analyser des données comportant des erreurs humaines ou de mesure, avec des valeurs qui n'entrent pas dans le spectre attendu. Dans cette situation, il est primordial d'exclure ces données « incohérentes » des analyses futures. Ensuite, la présence d'observations ne présentant pas le même comportement que la majorité des données peut influencer un grand nombre d'analyses statistiques. Deux réponses sont alors envisageables : (i) prétraiter les données initiales en identifiant ces observations et en les excluant ou, (ii) robustifier les méthodes statistiques afin que les résultats ne soient pas affectés par la possible présence de ces individus. Enfin, l'objectif premier de certaines méthodes est de détecter les observations atypiques ou anormales, par exemple pour identifier des patients malades, détecter des fraudes (bancaires ou autres) ou des défauts de construction dans l'industrie. Dans ce manuscrit, on investigue principalement ces anomalies de fabrication, qui ne sont pas détectables par les limites de tolérances des produits.

Ce travail de recherche est le fruit d'une collaboration de type Cifre entre le laboratoire TSE-R de l'Université Toulouse 1 Capitole et l'entreprise ippon innovation, dirigée par M. François Bergeret. Cette société, spécialiste du domaine des semi-conducteurs, propose entre autres des solutions de détection de pièces défaillantes basées sur des algorithmes statistiques avancés. Principalement experte du domaine automobile, celle-ci souhaitait élargir son champ d'action à l'aérospatial. Elle s'est associée avec l'entreprise Microchip-Atmel, une des leaders sur le marché des circuits intégrés. Cette collaboration a vu le jour dans le cadre du projet RESIST (*RESilient Integrated SysTems*), co-financé par le programme européen CATRENE (*Cluster for Application and Technology Research in Europe on NanoElectronics*), pour promouvoir la résilience des appareils électroniques dans les domaines de l'avionique, de l'automobile et de l'aérospatial. L'objectif de ce partenariat était de développer une solution non-supervisée de détection de défauts pour des produits électroniques complexes, caractérisés par un nombre de mesures très grand par rapport à l'ensemble des composants fabriqués. D'un point de vue statistique, ce travail de thèse est donc dédié à la détection non-supervisée d'observations atypiques en grande dimension.

Tout d'abord, dans le Chapitre 1, **« État de l'art : détection non-supervisée d'observations atypiques sur des données quantitatives avec le logiciel R »** , une revue de la littérature est dressée. Tout au long de ce travail, la notion d'anormalité retenue suit celle donnée par Hawkins (1980), à savoir qu'une observation est atypique si elle est générée par un mécanisme différent de celui de la majorité des données.
Une première section se focalise sur le contexte du contrôle de qualité dans l'industrie des composants électroniques destinés aux applications automobiles, afin d'établir un inventaire des différentes méthodes utilisées en pratique. Le constat est assez sévère car ce sont principalement des méthodes univariées qui sont intégrées aux différents processus de détection de défauts. Seules quelques méthodes multivariées de type distance de Mahalanobis (MD) ou Analyse en Composantes Principales (ACP) semblent connues de quelques industriels.
Les sections suivantes essaient de résumer l'ensemble de la palette de possibilités destinées à la détection d'observations atypiques de manière non-supervisée ainsi que leur mise en œuvre sous le logiciel R (R Core Team, 2017). Une distinction est faite entre les méthodes ne traitant que des données en dimension standard, i.e avec plus d'observations que de variables, et celles acceptant des données en grande dimension et avec une faible taille d'échantillon (HDLSS).

Le Chapitre 2, **« Multivariate Outlier Detection with ICS »** , est un article destiné à la publication, actuellement soumis, co-écrit avec mes directeurs de thèse Anne Ruiz-Gazen et Klaus Nordhausen.
La première partie de l'article est dédiée au comportement théorique de la distance de Mahalanobis (MD) quand le nombre de dimensions augmente alors que les atypiques sont contenus dans un sous-espace. Nous démontrons que dans cette situation la méthode est en difficulté pour retrouver les observations atypiques.
Dans le reste de l'article nous nous concentrons sur la méthode Invariant Coordinate Selection (ICS) que nous proposons comme alternative. Cette approche, introduite par Tyler et al. (2009), présente des propriétés remarquables pour révéler des structures de données, en se basant sur la décomposition spectrale simultanée de deux estimateurs de dispersion multivariés. Elle conduit à un nouveau système de coordonnées invariantes par transformation affine dans lequel la distance euclidienne correspond à une distance Mahalanobis (MD) dans le système d'origine. Toutefois, le véritable avantage de cette méthode comparée à la distance de Mahalanobis est qu'elle permet de ne sélectionner que les composantes pertinentes à la détection des observations atypiques.
L'objectif principal de cet article est de proposer une méthodologie pour identifier les observations anormales à l'aide de la méthode ICS, dans le cas où il n'y a qu'un faible pourcentage d'atypiques présents dans un sous-espace. Les principaux challenges sont le choix des estimateurs ainsi que la sélection des composantes d'intérêt. Une étude basée sur des simulations compare ces choix ainsi que les performances atteintes avec l'ACP et la MD.
Enfin, une note complémentaire se concentre sur certains des modèles étudiés dans l'article

et justifie que, dans ce contexte, la structure d'atypicité est contenue dans les premières composantes d'ICS. Elle précise également dans quelles conditions la méthode serait en difficulté pour détecter les anomalies.

Le Chapitre 3, « **Unsupervised outlier detection with the R package ICSOutlier** » , est un article destiné à la publication, actuellement accepté sous réserve de révisions, co-écrit avec mes directeurs de thèse Anne Ruiz-Gazen et Klaus Nordhausen. Il présente le package R ICSOutlier que nous avons développé suite à l'écriture de l'article présenté dans le chapitre précédent.
Ce package met donc en œuvre la méthodologie que nous avons proposée pour détecter des observations atypiques de manière non-supervisée. La fonction principale du package permet de sélectionner automatiquement les composantes les plus pertinentes, de calculer un index d'anormalité ainsi que d'identifier les individus atypiques. Chaque étape fait appel à des fonctions auxiliaires qui peuvent être utilisées indépendamment.
Afin de rendre cette nouvelle méthode la plus attractive possible, nous avons également développé une application shiny (ICSShiny), qui permet de faire de l'exploration de données à l'aide de la méthode ICS ainsi que de la détection d'atypiques. Un complément de chapitre lui est dédiée.

Le Chapitre 4, « **ICS with positive semi-definite scatter matrices for data not in general position** » , s'intéresse à la manière d'adapter la méthode ICS lorsque les estimateurs de dispersion considérés ne sont pas nécessairement définis positifs. En effet, dans la présentation classique de Tyler et al. (2009), les estimateurs de dispersion sont caractérisés par des matrices symétriques définies positives équivariantes par transformation affine.
Néanmoins, si les données à analyser contiennent des variables colinéaires, ou si le nombre de dimensions $p$ est supérieur au nombre d'observations $n$, alors même l'estimateur de variance-covariance empirique devient singulier. Or, ce type de situation est de plus en plus fréquent dans le contexte industriel des semi-conducteurs pour l'automobile et caractéristique des circuits intégrés destinés à l'aérospatiale. S'intéresser à ce type d'estimateurs est donc un enjeu d'actualité.
Toutefois, Tyler (2010) note que si les données sont en position générale [1], alors tous les estimateurs affines équivariants sont proportionnels à la matrice de variance-covariance. Dans ce cas, la méthode ICS ne peut révéler la structure d'atypicité des données, car elle cherche à diagonaliser une matrice proportionnelle à l'identité. Dans ce chapitre, nous supposons donc nécessairement que les données ne sont pas en position générale. Dans notre contexte industriel, les observations sont habituellement concentrées dans un espace de dimension inférieur, ce qui rend cette hypothèse réaliste.
Pour généraliser ICS à des estimateurs de dispersion singuliers, trois approches sont proposées : le recours à l'inverse généralisée pour rendre non singulier un des deux estimateurs, un pré-traitement des données par une réduction de dimension ou une décomposition en

1. Des données sont en position générale s'il n'existe aucun sous-ensemble formé de $k$ observations qui se concentrent dans un sous-espace de dimension $k-2$, avec $k \leq p+1$.

valeur singulière généralisée (GSVD). Ces trois solutions sont investiguées d'un point de vue théorique sous différents aspects. Le critère de la méthode est-il modifié ? Les scores sont-ils affines invariants ? Les estimateurs de dispersion jouent-ils un rôle symétrique ? La dernière méthode basée sur la décomposition en valeur singulière généralisée des deux estimateurs apporte une réponse positive aux trois questions posées. La propriété d'affine invariance suppose l'utilisation d'estimateurs semi-définis positifs affines équivariants. Or, en pratique dans le contexte de la grande dimension ou en présence de variables colinéaires, les estimateurs les plus utilisés, à l'exception de la matrice de variance-covariance, ne sont au mieux qu'orthogonalement équivariants.
Un exemple théorique permet d'illustrer ces propos alors qu'un cas réel se focalise sur la méthode GSVD.

Le Chapitre 5, **« A new outlier detection solution for HDLSS data in an industrial context »** , se consacre au contexte industriel dans lequel s'inscrit ce travail de thèse.
La mission principale était le développement d'une solution commercialisable de détection non-supervisée d'observations atypiques pour des données de type HDLSS (plus de variables que d'individus). Comme ce genre de données est particulièrement habituel dans le domaine de l'aérospatiale, une collaboration avec Microchip-Atmel, une entreprise de pointe pour les composants électroniques utilisés dans l'industrie spatiale, s'est mise en place dans le cadre d'un projet européen. Ce partenariat nous a permis entre autres d'avoir accès à des données réelles présentant de véritables problèmes de fiabilité à détecter à un stade précoce du processus de contrôle de la qualité.
Notre association avec Microchip-Atmel s'est également concrétisée par l'écriture d'un acte de conférence en 2017 pour la $10^{\text{ème}}$ conférence internationale sur les méthodes mathématiques dans la fiabilité (MMR). Cet acte intitulé « High dimensional outlier screening of small dice samples for aerospace IC reliability » est inséré dans ce chapitre. Il présente plus en détail le contexte et l'idée générale de l'algorithme ainsi que des résultats en termes de performance comparés à d'autres méthodes statistiques.
Bien que l'algorithme final soit confidentiel, ce chapitre décrit également les différents challenges rencontrés au cours de son développement ainsi que les avantages retirés par l'entreprise. A l'heure actuelle, l'algorithme est intégré dans l'outil de production de Microchip-Atmel afin d'identifier en avance de futurs problèmes de fiabilité. Cet outil permet de gagner en temps de cycle de détection et en fiabilité pour les composants spatiaux.

Pour information, dans la version pdf de ce manuscrit, des liens hypertextes sont actifs pour toutes les citations, acronymes et références.

## English version

In statistics, the presence of outlying observations is really noteworthy for different reasons. First, it is usual to analyze data containing human or measurement errors, i.e. values that do not fit into the expected spectrum. In this situation, such inconsistent data should be excluded from the future studies. Then, the observations which behave differently from the bulk of the data can influence the statistical analysis considerably. Two solutions are possible: (i) preprocess the data by deleting the identified outlying observations or, (ii) robustify the methods, so that the conclusions are not affected by the presence of these outliers. Finally, the main goal of some analysis might be to detect these outlying observations. For example, it is interesting to identify sick people, to detect fraud (in banks or others), or defects in the industry. In this manuscript, we mainly focus on these manufacturing anomalies, which are not identified by the tolerance limits of the products.

This research work comes from a collaboration within a "Cifre" agreement between the TSE-R laboratory of Toulouse 1 Capitole University and the ippon innovation company, headed by François Bergeret. This company is specialized in the semiconductor's domain, and proposes among other things, failing part detection solutions based on advanced statistical algorithms. Mainly an expert in the automotive field, ippon wanted to expand its scope to aerospace. With Microchip-Atmel, one of the leading firms in the integrated circuit market, they decided to partner. This collaboration was initiated within the RESIST (*RESilent Integrated SysTems*) project, co-funded by the European CATRENE (*Cluster for Application and Technology Research in Europe on NanoElectronics*), to promote the resilience of electronics in the avionic, automotive and aerospace industries. The objective of this cooperation was to develop an unsupervised fault detection solution for complex electronic products, characterized by a very large number of measurements compared to the number of manufactured components. From a statistical point of view, this work is dedicated to the unsupervised detection of outlying observations in high-dimensional data.

First, Chapter 1, **"État de l'art: détection non-supervisée d'observations atypiques sur des données quantitatives avec le logiciel R"**, is a review of the literature written in french. All along this work, the notion of outlierness follows the definition given by Hawkins (1980), namely that an observation is outlying if it is generated by a different mechanism than the one of the bulk of the data.
A first section focuses on the context of quality control for the electronic components for automotive applications. It reviews all the common methods used in practice. The conclusions are not very positive as mainly univariate methods are integrated into the fault detection processes. Only a few multivariate methods like the Mahalanobis distance (MD) or the Principal Components Analysis (PCA) are used by some manufacturers.
The next sections attempt to summarize all the Unsupervised methods for outlier detection as well as their implementation in the R software (R Core Team, 2017). A distinction

is made between methods designed for standard data, i.e. with more observations than variables, and those adapted to high dimensional data with a small sampling size (HDLSS).

Chapter 2, **"Multivariate Outlier Detection with ICS"**, is a reprint of a submitted paper, co-authored with my thesis supervisors Anne Ruiz-Gazen and Klaus Nordhausen. The first part analyzes the theoretical behavior of the Mahalanobis distance (MD) when the number of dimensions increases and the outliers lie in a subspace. We prove that the method has troubles to identify the outliers in this particular situation. In the following, we introduce the Invariant Coordinate Selection (ICS) method as an alternative solution. This approach, introduced by Tyler et al. (2009), shows remarkable properties for revealing data structures. Based on the simultaneous spectral decomposition of two scatter matrices, it leads to a new affine invariant coordinate system in which the Euclidian distance corresponds to a Mahalanobis Distance (MD) in the original system. However, unlike MD, ICS makes it possible to select relevant components. The objective of this paper is to propose a methodology for using ICS in case a small proportion of outliers lie in a lower dimensional subspace. The main challenges are the choice of scatter matrices together with the selection of relevant invariant components. A simulation study provides a comparison with PCA and MD.
Finally, a complementary note focuses on some of the models studied in the paper and it demonstrates that the structure of outlierness is always contained in the first components of ICS in our context. It also specifies under what conditions the method would encounter problems to detect the anomalies.

Chapter 3, **"Unsupervised outlier detection with the package ICSOutlier"**, is a preprint of an accepted (under revision) paper, co-authored with my thesis supervisors Anne Ruiz-Gazen and Klaus Nordhausen. It presents the R package ICSOutlier which we developed following the writing of the paper presented in the previous chapter.
This package implements the methodology we proposed for unsupervised outlier detection. The main function of the package allows to select automatically the most relevant components, to calculate an outlierness index as well as to identify the outlying observations. Each step calls an auxiliary function which could be also used independently.
In order to make this new method as user-friendly as possible, we have also developed a shiny application (ICSShiny), for data mining purposes as well as outlier detection using the ICS method. A complementary note briefly describes this new package.

Chapter 4, **"ICS with positive semi-definite scatter matrices for data not in general position"** focuses on a way to adapt the ICS method to the case when the scatter matrices are not necessarily positive definite. Indeed, in the classical presentation of ICS, Tyler et al. (2009) define a scatter matrix as a symmetric positive definite matrix, equivariant under affine transformations.
However, if the data contain collinear variables, or if the number of dimensions $p$ is greater than the number of observations $n$, then even the empirical variance-covariance matrix becomes singular. This case is therefore worth considering as these data are typical in the context of semiconductors for the automotive industry and characteristic of the aerospace

integrated circuits.
In this context, Tyler (2010) demonstrates that, as long as the data is in general position [2], any affine equivariant scatter matrices is proportional to the variance-covariance matrix. So, the ICS method cannot reveal the outlierness structure as it diagonalizes a matrix proportional to the identity. The paradigm taken in this chapter is however that the data is not in general position. In our industrial context, observations are usually lying on a subspace of smaller dimension, which ensures the assumption to be realistic.
For generalizing the ICS method to singular estimates, three approaches are proposed: computing the generalized inverse of one of the scatter matrices, pre-processing the data by a reduction of dimension or performing a Generalized Singular Value Decomposition (GSVD). These solutions are investigated from a theoretical point of view under various aspects. Is the criterion of the method modified? Are the scores still affine invariant? Do the scatter estimators play a symmetric role? It occurs that the nice properties of the classical ICS only remain valid for the method based on the generalized singular value decomposition. The affine invariance property only holds for affine equivariant semi-definite positive scatter matrices. But, in high dimension or in presence of collinear features, the most common scatter estimates are only orthogonally equivariant, except for the variance-covariance matrix.
A theoretical example illustrates the first part while a real case focuses on the GSVD method.

Finally, Chapter 5, **"A new outlier detection solution for HDLSS data in an industrial context"**, focuses on the industrial context considered throughout this research work.
The main objective was the development of a marketable solution for unsupervised outlier detection in HDLSS data (more variables than production items). Since this particular data is common in the field of aerospace, ippon innovation cooperated with Microchip-Atmel, a cutting-edge company for the electronic components used in the space industry, within the framework of an European project. This partnership allowed us to have access to real data with actual reliability issues to identify in an early stage of the quality control process.
In addition, as part of our collaboration, we wrote a conference paper for the $10^{th}$ International Conference on Mathematical Methods in Reliability (MMR) in 2017. This chapter includes a reprint of the paper, which is entitled "High dimensional outlier screening of small dice samples for aerospace IC reliability". It presents in more details the objectives as well as the general idea of the developed algorithm and some results in terms of efficiency compared to other statistical methods.
Although the final algorithm is confidential, this chapter describes the different challenges encountered during its development as well as the benefits for the company. Currently, the algorithm has been fully implemented in the production tool of Microchip-Atmel in

2. Data is in general position if there is no subset of $k$ observations lying on a subspace of dimension $k-2$, with $k \leq p+1$ and $p$ denotes the number of variables.

order to identify in advance future reliability problems. This tool saves time in detection cycle, and improves the reliability of the space components.

For your information, in the pdf version of this manuscript, the hypertext links are active for all citations, references and acronyms.

# Chapitre 1

# État de l'art : détection non-supervisée d'observations atypiques sur des données quantitatives avec le logiciel R

*This chapter is a review of the literature, written in french. All along this work, the notion of outlierness follows the definition given by Hawkins (1980), namely that an observation is outlying if it is generated by a different mechanism than the one of the bulk of the data. A first section focuses on the context of quality control for the electronic components for automotive applications. It reviews all the common methods used in practice. The conclusions are not very positive as mainly univariate methods are integrated into the fault detection processes. Only a few multivariate methods like the Mahalanobis distance (MD) or the Principal Components Analysis (PCA) are used by some manufacturers. The next sections attempt to summarize all the Unsupervised methods for outlier detection as well as their implementation in the R software (R Core Team, 2017). A distinction is made between methods designed for standard data, i.e. with more observations than variables, and those adapted to high dimensional data with a small sampling size (HDLSS).*

## Sommaire


**1.1 Introduction . . . . . 11**

**1.2 Contrôle de qualité . . . . . 12**

1.2.1 Contexte des semi-conducteurs . . . . . 13

1.2.2 Les standards en univarié . . . . . 14

1.2.3 Les méthodes multivariées usuelles non-supervisées . . . . . 17

1.2.4 Évaluation des méthodes et contributeurs . . . . . 22

1.2.5 Exemple réel de l'industrie des semi-conducteurs et mise en œuvre en R . . . . . 24

**1.3 Approches en dimension standard : $n > p$ . . . . . 27**

1.3.1 Généralités . . . . . 27

1.3.2 Présentation succincte des méthodes . . . . . 28

**1.4 Approches en grande dimension - faible taille d'échantillon (HDLSS) : $n < p$ . . . . . 36**

1.4.1 Le fléau de la dimension . . . . . 36

1.4.2 Les analyses en dimension globale . . . . . 38

1.4.3 Les analyses de sous-espaces de l'espace originel . . . . . 41

**1.5 Conclusion et perspectives . . . . . 42**

## 1.1 Introduction

La détection d'observations présentant un comportement atypique est un sujet de préoccupation depuis le XIXème siècle au moins. Les chercheurs écartaient les valeurs de l'échantillon qu'ils considéraient comme n'étant pas en accord avec la majorité de la population. Cette analyse n'étant basée que sur l'expertise de chacun à définir ce qu'est une anormalité, Peirce (1852) fut le premier à proposer un critère objectif. Il s'ensuivit de nombreuses recherches dont les avancées les plus pertinentes effectuées jusqu'en 1931 ont été consignées par Rider (1933). Toutefois, c'est véritablement l'article de Pearson and Sekar (1936) qui marque un tournant dans la discipline. En effet, ils mettent en évidence le problème désormais très connu d'effet de masque (*masking effect* en anglais), détaillé en Section 1.2.3, qui peut se manifester dès que plus d'une observation se comporte de manière anormale. Or, jusqu'à cette publication, les différents critères utilisés n'étaient adaptés qu'à l'identification d'une seule observation atypique. Notons que des solutions basées sur un processus itératif de détection se sont avérées être une stratégie infructueuse (Barnett and Lewis, 1994).

Dans le même temps que la recherche de critères pouvant définir et identifier des observations anormales se poursuivait, la problématique de savoir comment gérer ces valeurs s'est imposée. Dès 1977, Daniel Bernoulli remet en question la stratégie d'écarter ces mesures de toutes les analyses. Il est vrai que plusieurs cas de figure peuvent se présenter : il est possible que ces valeurs proviennent d'erreurs humaines de copie, ou véritablement qu'elles traduisent une réalité physique, physionomique, chimique ou autre. Comme l'analyste ne peut être certain que les mesures de l'échantillon ne contiennent aucune erreur, les recherches se sont portées sur une manière de pouvoir gérer la présence de telles valeurs en évitant qu'elles n'impactent trop les résultats des analyses statistiques. Le concept de Robustesse se développe alors rapidement et de nouvelles notions sont introduites pour appréhender quantitativement la robustesse des méthodes statistiques (voir Maronna et al. (2006), Droesbeke et al. (2015) pour une présentation détaillée). Pour pallier au problème du manque de robustesse des analyses classiques, de nombreuses méthodes sont proposées, notamment pour l'estimation de paramètres de position et d'échelle en univarié et multivarié.

En dépit de l'importance des challenges à relever et de la recherche effectuée, il faut attendre Hawkins (1980) pour un premier ouvrage dédié aux méthodes de détection ainsi qu'une définition formelle du concept d'observation atypique :
*« An outlier is an observation which deviates so much from the other observations as to arouse suspicious that is was generated by a different mechanism. »*
Cette intuition, qu'une observation est atypique par rapport au reste de la population parce qu'elle a été générée par un mécanisme différent, va s'instaurer comme un fondement du domaine. Le second ouvrage incontournable est la contribution de Barnett and Lewis (1994) qui présentent des outils pour pouvoir identifier ces observations anormales ou réussir à s'en accommoder sans trop impacter les analyses. Enfin, il faut attendre Ag-

garwal (2013) pour avoir de nouveau une revue de la littérature qui prend en compte les avancées des dernières décades. L'originalité de son ouvrage tient à ce qu'il regroupe les méthodes émergentes dans la communauté statistique aussi bien que celles de la communauté informatique.

Paradoxalement au (trop) petit nombre de livres publiés sur le sujet, à l'heure actuelle, tous les domaines sont concernés par la gestion des observations atypiques, seul leur but final diffère. Dans certains domaines, il est seulement nécessaire d'identifier et de supprimer ces individus, dans d'autres, il faut que leurs analyses ne soient pas trop impactées par la présence potentielle de ces observations, et enfin pour les derniers, l'objectif même de leurs études est de les détecter. Pour exemple, on peut vouloir mettre en place un système de détection d'intrusion en cybersécurité, de détection de fraudes aux cartes de crédit, traquer les variations de capteurs de tous genres pour anticiper des pannes, diagnostiquer des maladies, prévenir des ouragans, des tremblements de terre ou bien renforcer la fiabilité des composants électroniques.

C'est ce dernier objectif qui nous intéresse particulièrement dans ce travail de recherche collaboratif entre le monde industriel et le monde académique. En effet, afin d'assurer un niveau de qualité élevé des composants électroniques dans les domaines de l'automobile ou du spatial, la détection des composants potentiellement défectueux est un enjeu crucial. D'un point de vue statistique, plusieurs challenges se posent. Tout d'abord, en fonction de la complexité des composants électroniques, le nombre de tests à prendre en compte peut être très important. Ensuite, principalement dans le contexte spatial, le nombre de tests excède généralement le nombre de composants testés. Or les méthodes statistiques classiques multivariées de détection d'anomalies ne peuvent s'appliquer que dans les cas où la taille d'échantillon $n$ est plus grande que la dimension $p$ ($n > p$). La gestion de données de type HDLSS (*High Dimension - Low Sample Size*), i.e. en grande dimension et avec une faible taille d'échantillon ($n < p$), est donc également un enjeu critique.

L'ensemble de la thèse se focalise sur des méthodes de détection d'observations anormales dans le cas non-supervisé, i.e. sans avoir d'information à priori, afin de se placer dans le contexte industriel de détection d'une infime proportion de défauts de fabrication. La première partie de ce chapitre présente les objectifs et les méthodes utilisées dans le cadre du contrôle de qualité principalement des semi-conducteurs. La deuxième partie synthétise la revue de la littérature d'Aggarwal (2013) en présentant les principales caractéristiques des méthodes ainsi que leur possible mise en œuvre sous le logiciel R (R Core Team, 2017). Finalement, la troisième partie est dédiée aux approches s'appliquant aux échantillons de type HDLSS, qu'elles proviennent des communautés informatique ou statistique.

## 1.2 Contrôle de qualité

Dans les industries automobile, aéronautique ou spatiale, la recherche du zéro-défaut est la finalité à atteindre. L'émergence des marchés à bas prix et le nombre toujours

croissant de composants électroniques qui équipent les différents systèmes, comme par exemple pour la voiture autonome, maintiennent la pression sur les entreprises pour qu'elles livrent des produits de qualité toujours supérieure. Pour parvenir à cet objectif, un long processus de qualification des produits, décrit ci-dessous, est mis en place pendant toutes les étapes de fabrication.

Dans cette section, on présente brièvement le contexte des semi-conducteurs. Ensuite, on se focalise sur les méthodes statistiques non-supervisées univariées et multivariées utilisées couramment dans ce domaine pour valider la fiabilité des puces fabriquées. La section suivante propose des critères d'évaluation de la performance des méthodes employées dans ce contexte particulier et introduit le rôle important des contributeurs. Enfin, un exemple réel de l'industrie est présenté afin de pouvoir comparer certaines méthodes multivariées.

### 1.2.1 Contexte des semi-conducteurs

Dans le domaine des semi-conducteurs, Moreno-Lizaranzu and Cuesta (2013) décrivent en détail les quatre phases importantes du processus de création : la fabrication du wafer, le probe, l'assemblage et le test final. Précisions qu'un wafer est une plaque (généralement en silicium) sur laquelle les circuits intégrés sont fabriqués couche par couche. L'étape du probe consiste à effectuer des mesures électriques sur chaque circuit pour s'assurer qu'ils ne sont pas défectueux et pouvoir choisir lesquels envoyer en assemblage. Les circuits intégrés assemblés en paquets sont de nouveau soumis à une série de tests afin de filtrer toutes les pièces qui présenteraient des défauts. En dépit de toutes ces étapes de vérification, certains appareils contiennent des pièces présentant des défauts latents qui se manifesteront plus tard et qui pourront être à l'origine d'un incident de qualité chez le client (CQI). Afin de réduire ces incidents au taux le plus bas, cinq niveaux de vérification sont appliqués dans le cadre du contrôle global de la qualité :

— FDC (*Fault Detection and Classification*) : tout d'abord, les paramètres machines sont testés pour détecter des défauts et les classer. Cette première étape permet de s'assurer que les paramètres machine ne dérivent pas et permet d'intervenir si nécessaire. Entre autres, Lee et al. (2004b) proposent d'utiliser la méthode d'analyse en composantes indépendantes (ICA) pour identifier ces déviations. Quant à Lee et al. (2004a) et Taouali et al. (2016) ils suggèrent d'exploiter une variante de l'analyse en composantes principales (ACP) à noyau (KPCA).

— SPC (*Statistical Process Control*) ou MSP (Maîtrise Statistique des Procédés) en français : cette vérification a lieu au cours de la fabrication des wafers et repose sur des méthodes statistiques et graphiques comme les cartes de contrôle. Une littérature abondante existe sur le sujet, voir Entre autres Mercier and Bergeret (2011); Mnassri et al. (2008); Jensen et al. (2007); Harkat et al. (2002); Cinar and Undey (1999); Rocke (1992) pour une synthèse des principales approches utilisées.

— PT (Test Paramétrique) : de nombreux tests électriques sont mesurés sur les wafers en fin de fabrication, afin d'écarter les plaques qui présenteraient un comportement

électrique anormal. À titre d'exemple, les travaux de thèse d'Hassan (2014) visaient la mise en place d'un système de détection en temps réel pour l'entreprise STMicroelectronics.

— Probe : à ce stade ce sont les puces de chaque wafer qui sont testées pour détecter de potentiels défauts fonctionnels. Elles sont donc caractérisées en deux groupes : les « pass » ou les « fail » . Seules les premières, qui sont jugées comme « bonnes » sont mises en assemblage. En fonction du type de produit, les tests, généralement de tension, de courant, de temps de réponse, etc, peuvent varier en intensité et en nombre.

— FT (Test Final) : une fois les puces assemblées, elles sont testées dans différentes conditions pour s'assurer de leur performance. Ces insertions regroupent des températures froide, ambiante et chaude (*cold, room, hot* en anglais) souvent suivies par une phase de *burn-in* où les puces sont placées dans des étuves. Cette dernière étape est réservée aux produits finis les plus complexes qui nécessitent une rigueur supplémentaire dans la détection des possibles pièces défaillantes.

Idéalement, il est pertinent de détecter les éventuels défauts dans les premières étapes du processus de validation. Toutefois rejeter un wafer au niveau du Test Paramétrique (PT) est déjà très coûteux car on écarte de ce fait toutes les puces créées sur cette plaque. Dans cette section ainsi que dans le Chapitre 5, nous nous limitons à l'analyse non-supervisée du contrôle de qualité des puces, ce qui correspond aux niveaux du Probe et du Test Final (FT).

### 1.2.2 Les standards en univarié

Le comité du JEDEC (2009) (Joint Electron Device Engineering Council) développe et publie les standards dans l'industrie du semi-conducteur. Dans l'industrie automobile, les entreprises Chrysler, Ford et General Motors ont créé un autre comité, Automotive Electronic Council (2011), qui donne également des lignes directrices pour assurer une production fiable et de haute qualité. Moreno-Lizaranzu and Cuesta (2013) proposent une synthèse des méthodes recommandées et mises en œuvre dans l'entreprise de semi-conducteurs Freescale (maintenant NXP). Nous présentons ici brièvement les approches non-supervisées.

#### LSL et USL : les limites de spécification

Avant l'arrivée de la statistique dans le processus de qualification, les ingénieurs de tests vérifiaient uniquement si les valeurs de chaque mesure étaient comprises entre les limites de spécifications. Ces limites, notées LSL (*Lower Specification Limit*) et USL (*Upper Specification Limit*), sont déterminées théoriquement ou par l'expertise des clients. Par définition, une pièce est considérée comme « bonne » (*pass*, en anglais) si elle est comprise entre ces limites LSL et USL. Même si cette méthode s'avère incontournable pour répondre

au cahier des charges défini pour chaque produit, les limites sont généralement très larges et ne permettent pas de détecter toutes les pièces au comportement anormal.

### PAT : Part Average Testing

Afin de renforcer le filtrage des pièces potentiellement atypiques, un second niveau de détection est mis en place : le PAT. Il permet d'identifier et de rejeter les pièces qui se comportent de manière anormale sur le plan statistique. Sous l'hypothèse de normalité de l'échantillon, cela revient à identifier les valeurs qui ne semblent pas avoir été générées par une distribution normale. Les valeurs qui se trouvent à plus de $k$ écarts types $\sigma$ par rapport à la moyenne $\mu$ mais toujours dans les limites de spécifications LSL et USL, sont de potentielles pièces atypiques, comme illustré sur la Figure 1.1. Formellement les limites basses et hautes, $\text{PAT}_L$ et $\text{PAT}_U$, sont déterminées en fonction de $k$ par :

$$PAT_L = \mu - k\sigma \qquad \& \qquad PAT_U = \mu + k\sigma \tag{1.1}$$

Généralement, la valeur de $k = 6$ est préconisée.

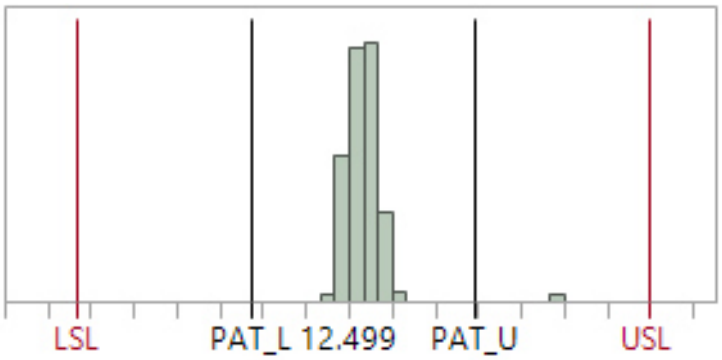



FIGURE 1.1 – PAT_L et PAT_U sont les limites de PAT calculées avec $k = 6$. LSL et USL sont les limites de spécifications.

Moreno-Lizaranzu and Cuesta (2013) présentent dans leur article différentes variantes du PAT. Dans la pratique, les limites peuvent être calculées de manière statique ou dynamique. Dans le premier cas, les paramètres statistiques $\sigma$ et $\mu$ sont calculés sur des pièces de référence et ensuite appliqués aux nouvelles pièces testées. Dans le deuxième cas, les paramètres $\sigma$ et $\mu$ sont estimés depuis l'échantillon par la moyenne empirique $\mu_n$ et l'écart type empirique $\sigma_n$. Enfin ces limites peuvent aussi être déterminées de manière robuste, c'est-à-dire en choisissant des estimateurs de position et de dispersion qui ne sont pas impactés par la présence de valeurs atypiques, comme la médiane et l'écart inter-quartiles (IQR, *Inter-quartile Range*) pondéré par exemple. Toutefois l'utilisation des estimateurs robustes peuvent induire le rejet de groupes entiers de pièces et non pas des pièces réellement atypiques.

### GPAT ou GDBC, *Geographic Part Average Testing* ou *Good Die in Bad Cluster*

Le GPAT est une variante du PAT qui prend en compte la dimension spatiale des pièces sur le wafer. Lors de la phase de Probe, toutes les pièces sont testées et si elles respectent les limites de spécification sur tous les tests, alors elles sont identifiées comme bonnes (« pass »), et sont caractérisées par une couleur blanche sur la cartographie du wafer représenté sur la Figure 1.2. Au contraire, elles sont considérées comme défectueuses (« fail »), et représentées par des couleurs si au moins un test est en dehors de ces limites. En fonction du type de tests hors limites, la couleur de la puce est différente. L'idée de la méthode est alors de rejeter les pièces « pass » situées dans un cluster de pièces « fail », comme par exemple la pièce blanche entourée par un carré noir en haut à gauche du wafer, qui est entourée de pièces qui vont être rejetées.

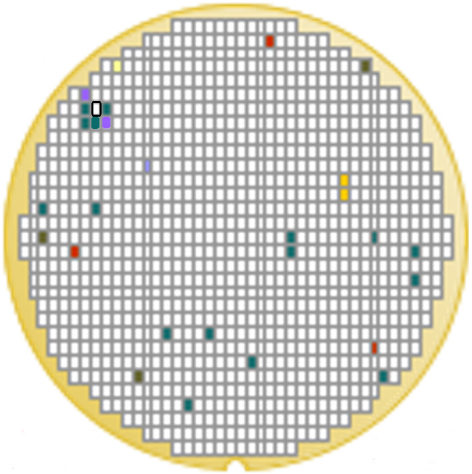


FIGURE 1.2 – Cartographie du wafer représentant la position géographique des puces électroniques. Les couleurs caractérisent la qualification de la puce : blanche pour « pass » et en couleur pour « fail ».

### NNR : *Nearest Neighbor Residual*

Cette approche prend en compte la notion de voisinage géographique par rapport à la position des pièces sur le wafer. L'idée est de rejeter la pièce si sa valeur mesurée est statistiquement différente de la valeur attendue, calculée comme une moyenne des pièces voisines, pour un test donné (voir Moreno-Lizaranzu and Cuesta (2013) pour une description complète de la méthode). La Figure 1.3 illustre l'idée de cette approche.

Toutes ces méthodes permettent une amélioration de la fiabilité, mais leur coût par rapport à la détection devient rapidement prohibitif. En effet, les méthodes sont appliquées sur chaque test de manière indépendante et donc le taux de fausses alarmes augmente de façon spectaculaire avec le nombre de tests. Par exemple, comme expliqué par Mercier and Bergeret (2011), si nous considérons $p$ tests électriques au niveau de risque de $\alpha\%$, nous

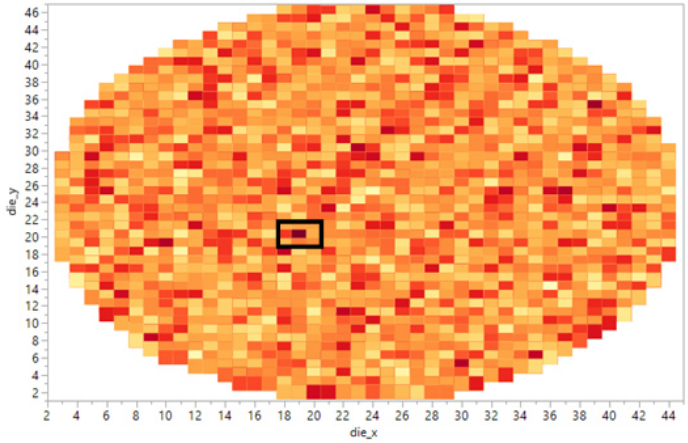



FIGURE 1.3 – Wafer dont les couleurs des puces représentent les différences entre les valeurs mesurées et attendues de chaque pièce sur un test électrique, par la méthode NNR. La puce en rouge foncée au centre du rectangle noir est différente de ses voisines sur ce test.

multiplions le taux de fausses alarmes par un facteur voisin de $p$ :

$$\begin{aligned} &P[\text{Au moins une mesure est hors contrôle}] \\ &= 1 - P[\text{Toutes les mesures sont sous contrôle}] = 1 - (1-\alpha)^p \approx p\alpha \end{aligned} \tag{1.2}$$

De plus, comme les tests de détection sont réalisés de manière indépendante, les corrélations entre les mesures ne sont pas prises en compte. Ces méthodes ne permettent que de détecter des anormalités univariées, ce qui est un véritable inconvénient pour l'analyse de produits complexes nécessitant une fiabilité extrême.

### 1.2.3 Les méthodes multivariées usuelles non-supervisées

#### La statistique du $T^2$ de Hotelling ou la distance de Mahalanobis (MD)

Une méthode parfois utilisée dans la maîtrise statistique des procédés (MSP ou SPC en anglais) est le calcul de la statistique du $T^2$ de Hotteling (1947), comme expliqué dans Jensen et al. (2007); Lafaye de Micheaux (2000); Mnassri et al. (2008). Cette statistique, équivalente à la distance de Mahalanobis (MD) au carré, peut également être appliquée pour renforcer le contrôle de qualité des puces (Aggarwal, 2013). Elle est souvent décrite comme une carte de contrôle multivariée.

Formellement, soient $\mathbf{x}_1, \ldots, \mathbf{x}_n$, $n$ observations caractérisées par $p$ variables quantitatives. Chaque $\mathbf{x}_i$ est supposé être un vecteur aléatoire réel $p$-multivarié généré par une distribution normale multivariée avec un paramètre de localisation $\boldsymbol{\mu}$ et une matrice de variance-covariance $\boldsymbol{\Sigma}$ : $\mathbf{x}_i \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. On définit la distance de chaque puce $\mathbf{x}_i$ au centre de la distribution $\boldsymbol{\mu}$ par :

$$T^2_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\mathbf{x}_i) = MD^2_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\mathbf{x}_i) = (\mathbf{x}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}) \tag{1.3}$$

Généralement les paramètres statistiques de position et de dispersion sont inconnus et doivent être estimés, le plus souvent à l'aide de la moyenne empirique $\boldsymbol{\mu}_n = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$ et de la matrice de variance-covariance empirique $\boldsymbol{\Sigma}_n = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu}_n)(\mathbf{x}_i - \boldsymbol{\mu}_n)'$. En fonction des estimateurs utilisés et sous l'hypothèse de normalité, la distribution des distances peut être déduite. Les distances peuvent donc être testées pour savoir si elles dépassent le quantile théorique et donc si les observations associées sont de potentielles anomalies ou non. Mardia et al. (1979) notamment rappellent que lorsque les paramètres $\boldsymbol{\mu}$ et $\boldsymbol{\Sigma}$ sont connus la distribution de $\mathrm{MD}^2_{\boldsymbol{\mu},\boldsymbol{\Sigma}}$ est une loi du $\chi^2$ à $p$ degrés de liberté. Wilks (1962) et Gnanadesikan and Kettenring (1972) démontrent que lorsque les paramètres sont estimés empiriquement par $\boldsymbol{\mu}_n$ et $\boldsymbol{\Sigma}_n$, la distribution exacte est une loi Beta. Plus précisément :

$$\frac{(n-1)^2}{n}\,\mathrm{MD}^2_{\boldsymbol{\mu}_n,\boldsymbol{\Sigma}_n}(\mathbf{y}_i) \sim \mathrm{Beta}\left(\frac{p}{2}, \frac{n-p-1}{2}\right),$$

qui peut être approximée par une loi de $\chi^2_p$ pour $n$ grand. C'est cette dernière approximation qui est le plus largement utilisée en pratique, même si son exactitude dépend de la dimension considérée. Une observation est donc identifiée comme une anomalie au niveau $\alpha\%$ si :

$$\mathrm{T}^2_{\boldsymbol{\mu}_n,\boldsymbol{\Sigma}_n}(\mathbf{x}_i) = \mathrm{MD}^2_{\boldsymbol{\mu}_n,\boldsymbol{\Sigma}_n}(\mathbf{x}_i) > c_{p,1-\alpha} \tag{1.4}$$

avec $c_{p,\alpha}$ le quantile d'ordre $1-\alpha$ d'une loi du $\chi^2$ à $p$ degrés de liberté.

Cette méthode est l'une des plus répandues pour la détection d'anomalies, principalement parce qu'elle est affine invariante tant que les estimateurs de position et de dispersion choisis sont affine équivariants, i.e. que par une transformation affine des données, ils vérifient : $\boldsymbol{\mu}_n(\mathbf{A}\mathbf{x}_i + \mathbf{b}) = \mathbf{A}\boldsymbol{\mu}_n(\mathbf{x}_i) + \mathbf{b}$ et $\boldsymbol{\Sigma}_n(\mathbf{A}\mathbf{x}_i + \mathbf{b}) = \mathbf{A}\boldsymbol{\Sigma}_n(\mathbf{x}_i)\mathbf{A}'$ avec $\mathbf{A}$ une $p \times p$ matrice non singulière et $\mathbf{b}$ un $p$-vecteur.

Toutefois, la méthode précédente soulève également certaines critiques. Tout d'abord le choix du niveau du test est encore un point délicat car il influence le type de taux d'erreurs qui va être contrôlé : le PCER (*per-comparison error rate*) ou le FWER (*family-wise error rate*), comme discuté dans Cerioli (2010). L'approche la plus courante consiste à contrôler le PCER qui garantit d'identifier $\alpha\%$ d'observations atypiques même s'il n'y a pas de valeur aberrante dans les données considérées. Becker and Gather (1999) et Cerioli et al. (2009) conseillent donc de commencer par tester de manière simultanée l'absence de valeurs aberrantes en contrôlant le niveau du test par des ajustements de type Bonferroni.

Ensuite, l'utilisation de paramètres empiriques tels que $\boldsymbol{\mu}_n$ et $\boldsymbol{\Sigma}_n$, sensibles à la présence de valeurs atypiques, est critiquée. En effet, ce choix de paramètres peut entraîner les effets bien connus de masque ou de débordement (*swamping*, en anglais), comme illustré sur la Figure 1.4 (Filzmoser and Todorov, 2011). Ce diagramme de dispersion représente les deux dimensions d'un jeu de données contenant 15 observations atypiques, représentées par les symboles « x » et « * ». Une observation est atypique si elle présente un comportement différent de la majorité des données, mais ne prend pas forcément des valeurs extrêmes sur chaque dimension. Les observations identifiées par des « + » ne sont pas considérées comme atypiques car bien qu'elles prennent des valeurs basses sur les deux

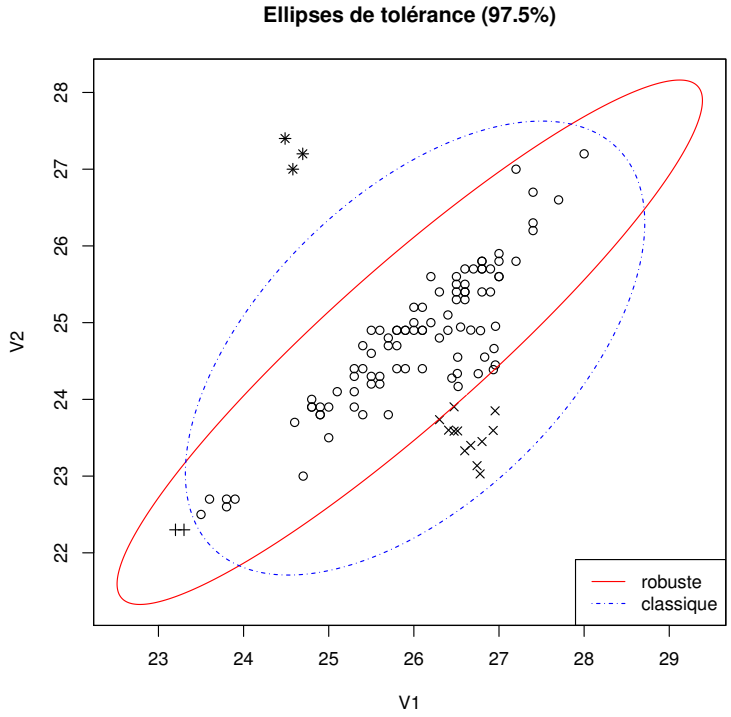



FIGURE 1.4 – Jeu de données bi-dimensionnel contenant plusieurs observations anormales. Deux ellipses de tolérance à 97.5% sont représentées pour essayer de révéler ces anomalies.

dimensions, leur comportement est similaire à la majorité des données. De plus, deux ellipses de tolérance à 97,5% sont tracées, une en pointillé, calculée avec les estimateurs non robustes usuels et une en trait plein, obtenue avec les estimateurs robustes du $\text{MCD}_{0.8}$ [1] (Rousseeuw, 1986). Ces ellipses sont censées contenir 95% de la population la plus « centrale » d'une distribution normale et donc révéler les observations atypiques comme des points extérieurs. En deux dimensions, elles illustrent parfaitement le fonctionnement de la distance de Mahalanobis calculée avec des estimateurs robustes ou non. Il apparaît ici que l'ellipse robuste est beaucoup plus plate que l'ellipse classique et qu'elles n'identifient donc pas les mêmes observations comme anormales. Dans le cas robuste, les 15 anomalies sont bien identifiées comme telles, alors que dans le cas classique, seules les trois « * » le sont et deux autres observations (les « + ») sont identifiées à tort comme atypiques. Ce phénomène est dû aux effets de masque (pour « x ») et de débordement (pour « + »).
L'effet de masque se produit lorsqu'un groupe de points distants de la majorité des observations attire les estimations de moyenne et de covariance, comme le font les observations « x ». La distance résultante des points périphériques par rapport à la moyenne devient alors petite et les observations ne peuvent pas être détectées comme atypiques, ce qui entraîne des faux-négatifs (atypiques détectés comme non-atypiques). En fait, les anomalies « x » masquent leur atypicité en influençant les valeurs des estimateurs de position et de dispersion.

1. $\text{MCD}_\alpha$ : estimateurs repondérés du déterminant minimum qui correspondent aux estimateurs empiriques non robustes usuels du sous espace contenant $\alpha n$ observations qui minimise le déterminant de la matrice de variance-covariance.

L'effet de débordement, survient lorsqu'un groupe de points périphériques fléchit les estimations de moyenne et de covariance dans leur direction et les éloigne des autres points. La distance résultante entre les points initiaux et la moyenne est grande. Dans ce cas, le nombre de faux-positifs augmente, i.e. des observations non atypiques sont détectées comme atypiques comme l'illustrent les observations « + ».
Utiliser des estimateurs de position et de dispersion robustes permet de prévenir ces effets.

Pour obtenir une distance de Mahalanobis robuste, il suffit de considérer des estimateurs de position et de dispersion robustes $\boldsymbol{\mu}_R$ et $\boldsymbol{\Sigma}_R$, comme illustré précédemment :

$$\mathrm{T}^2_{\boldsymbol{\mu}_R,\boldsymbol{\Sigma}_R}(\mathbf{x}_i) = \mathrm{RD}^2_{\boldsymbol{\mu}_R,\boldsymbol{\Sigma}_R}(\mathbf{x}_i) = (\mathbf{x}_i - \boldsymbol{\mu}_R)'\boldsymbol{\Sigma}_R^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_R) \tag{1.5}$$

Comme pour leur version non robuste, ces distances restent affines invariantes tant que les estimateurs considérés sont affines équivariants. Toutefois, il est important de noter qu'elles ne suivent pas exactement la même distribution.

La version robuste de la distance de Mahalanobis s'est particulièrement répandue suite à la démocratisation de l'utilisation des estimateurs du déterminant minimum (MCD) par Rousseeuw and Van Zomeren (1990). Cette approche est même parfois utilisée dans le SPC (Jensen et al., 2007). Toutefois, avec des estimateurs MCD, les distributions des distances présentées précédemment ne sont plus valides. Sous l'hypothèse de normalité, Hardin and Rocke (2005) donnent des arguments pour approximer la queue de distribution des distances de Mahalanobis calculées avec des estimateurs MCD par une loi de Fisher. Cerioli et al. (2009), Cerioli (2010) et Green and Martin (2017b) proposent d'ajuster le nombre de degrés de liberté de cette distribution de Fisher, dans le cas où le point de rupture n'est plus de 50% ou bien lorsque le MCD est repondéré.

Enfin, même si la distance de Mahalanobis, dans sa version robuste ou non, a fait ses preuves au cours des années, elle a l'inconvénient de prendre en considération tous les tests. Or, d'après notre expérience dans le domaine industriel, il est possible que les pièces défectueuses ne se comportent anormalement que sur un sous-ensemble de tests, ce qui implique de rechercher le sous-espace dans lequel elles se révèlent être atypiques. Une méthode très utilisée pour réduire l'espace initial est l'ACP.

#### ACP : Analyse en Composantes Principales

L'Analyse en Composantes Principales (ACP) est un procédé bien connu pour réduire la dimension d'un ensemble de données corrélées. Elle est couramment appliquée en MSP multivariée comme l'expliquent Lafaye de Micheaux and Vieux (2005); Lafaye de Micheaux et al. (2007). L'idée clé est de transformer les variables initiales en des composantes principales (PC) non corrélées et ordonnées de sorte que les premières $k$ composantes principales expliquent la plus grande partie de la variation des données initiales. Ce nouvel hyperplan $k$-dimensionnel est obtenu en minimisant l'erreur de projection au carré (SPE). Formellement, comme expliqué dans l'ouvrage de Jolliffe (2002), on diagonalise la matrice de

variance-covariance $\mathbf{\Sigma}$ de dimension $p \times p$ :

$$\mathbf{\Sigma} = \mathbf{PDP}' \tag{1.6}$$

où $\mathbf{D}$ est une matrice diagonale contenant les valeurs propres $\gamma_1 \geq \cdots \geq \gamma_p$ de $\mathbf{\Sigma}$, tandis que les colonnes de la matrice (orthonormée) $\mathbf{P}$ contiennent les vecteurs propres correspondants. Les nouvelles coordonnées des observations sont obtenues en projetant le tableau de données initiales $\mathbf{X}$ dans le sous-espace formé par les $k$ premiers axes :

$$\mathbf{C} = \mathbf{XP}_k \tag{1.7}$$

Nonobstant la simplicité apparente de la méthode, plusieurs difficultés se présentent pour identifier formellement les possibles observations atypiques.

Tout d'abord, le nombre $k$ de composantes à retenir n'est pas un choix aisé. L'idée est de trouver le sous-espace qui réduit significativement le nombre d'axes à prendre en compte tout en expliquant une part importante de la variation des données initiales. Pour sélectionner au mieux ce nombre de composantes, plusieurs critères existent. Le premier consiste à retenir le nombre d'axes nécessaires pour expliquer au moins une fraction $\phi$ de l'inertie totale. Formellement les valeurs propres permettent de déterminer la part d'inertie ou de variance expliquée. Avec $k$ composantes, le pourcentage cumulé de variance est donc :

$$PCV(k) = \frac{\sum_{l=1}^{k} \gamma_l}{\sum_{l=1}^{p} \gamma_l} \tag{1.8}$$

Il suffit donc de chercher le plus petit nombre $k$ qui permette d'expliquer la fraction $\phi$ de l'inertie totale souhaitée :

$$k = \arg\min_{l} \{PCV(l) \geq \phi\} \tag{1.9}$$

Un autre critère consiste à analyser le graphique des valeurs propres, appelé « éboulis de valeurs propres », qui représente les valeurs propres $\gamma_l$ par ordre décroissant. L'idée est de rechercher un « coude » dans cet éboulis de valeurs, qui serait suivi par une décroissance régulière. Cette décroissance est le signe que les axes associés à ces faibles valeurs ne sont pas pertinents pour l'analyse et que les axes associés ne permettent d'expliquer qu'une très faible part de la variation initiale.

Ensuite, utiliser une ACP pour détecter des anomalies n'est pas une tâche facile car les valeurs aberrantes peuvent être révélées sur les premiers et/ou les derniers axes. Ainsi, Hubert et al. (2005) ont introduit l'idée d'analyser un diagramme de dispersion qui représente deux scores : SD (*Score Distance*) et OD (*Orthogonal Distance*).
Pour chaque observation $\mathbf{x}_i$, la distance SD est calculée dans l'espace formé par les $k$ premières composantes sélectionnées et la distance OD dans l'espace orthogonal :

$$\begin{aligned} \mathrm{SD}^2_{\boldsymbol{\mu}_n, \mathbf{\Sigma}_n}(\mathbf{x}_i, k) &= ||\mathrm{diag}(\frac{1}{\sqrt{\gamma_1}}, \ldots, \frac{1}{\sqrt{\gamma_k}})\mathbf{P}'_k(\mathbf{x}_i - \boldsymbol{\mu}_n)||^2 \\ \mathrm{OD}^2_{\boldsymbol{\mu}_n, \mathbf{\Sigma}_n}(\mathbf{x}_i, k) &= ||(\mathbf{I}_d - \mathbf{P}_k\mathbf{P}'_k)(\mathbf{x}_i - \boldsymbol{\mu}_n)||^2 \end{aligned} \tag{1.10}$$

*Remarque* 1. Si $k = p$ alors la distance $\mathrm{SD}^2_{\mathbf{\Sigma}_n}(\mathbf{x}_i, k = p)$ est équivalente à la distance de Mahalanobis $\mathrm{MD}^2_{\boldsymbol{\mu}_n, \mathbf{\Sigma}_n}(\mathbf{x}_i)$.

Puis, une observation $\mathbf{x}_i$ est identifiée comme atypique si sa distance SD et/ou OD dépasse l'un des quantiles théoriques dérivés par Hubert et al. (2005) :

$$\mathrm{SD}(\mathbf{x}_i) > c_{p,1-\alpha} \text{ et/ou } \mathrm{OD}(\mathbf{x}_i) > (\mathrm{med}_{\mathrm{OD}} + \mathrm{MAD}_{\mathrm{OD}}\ z_{1-\alpha})^{3/2} \tag{1.11}$$

où $c_{p,1-\alpha}$ est le quantile d'une distribution du $\chi^2$ à $p$ degrés de liberté d'ordre $1-\alpha$, $z_{1-\alpha}$ est le quantile d'une distribution gaussienne d'ordre $1-\alpha$, $\mathrm{med}_{\mathrm{OD}}$ représente la médiane des distances OD et $\mathrm{MAD}_{\mathrm{OD}}$ la médiane des valeurs absolues des écarts à la médiane des distances OD (*Median Absolute Deviation*, en anglais).

*Remarque* 2. Avec cette méthode, il est nécessaire d'analyser les valeurs des distances SD et/ou OD d'une observation $\mathbf{x}_i$ car il n'est pas possible de déterminer en amont si l'observation va être atypique seulement dans l'espace des premières composantes principales (SD), dans l'espace orthogonal (OD) ou bien dans les deux.

*Remarque* 3. Cette méthode présente l'avantage de prendre en compte l'espace orthogonal aux premières composantes principales. Toutefois, contrairement à la distance de Mahalanobis, les distances SD et OD sont uniquement orthogonales équivariantes, c'est-à-dire que les résultats diffèrent selon que les données sont préalablement standardisées par un changement d'échelle (division par l'écart-type notamment) ou pas.

*Remarque* 4. La distance SD est également référencée comme étant le calcul de la statistique du $\mathrm{T}^2$ de Hotelling sur les $k$ premiers axes principaux (cf Harkat et al. (2002); Hassan (2014)). La distance OD correspond quant à elle à la SPE, soit l'erreur de projection au carré.

*Remarque* 5. Comme pour la distance de Mahalanobis, il est possible de robustifier la méthode en diagonalisant un estimateur de dispersion robuste comme le MCD par exemple, au lieu de la matrice de la matrice de variance-covariance.

### 1.2.4 Évaluation des méthodes et contributeurs

#### Évaluation des méthodes

Afin de pouvoir comparer les différentes approches, il est important de définir certains critères de performance qui mesurent l'efficacité des méthodes. L'objectif est d'identifier les « vraies » valeurs aberrantes, c'est-à-dire les incidents de qualité du client (CQI) avérés dans le domaine industriel, tout en minimisant les fausses détections. Les résultats d'identification peuvent être de quatre types, résumés dans le Tableau 1.1, avec :

- TP : le nombre de vrais positifs, c'est-à-dire le nombre de CQI détectés comme valeurs aberrantes.
- FN : le nombre de faux négatifs, c'est-à-dire le nombre de CQI détectés comme valeurs non aberrantes.
- FP : le nombre de faux positifs, c'est-à-dire le nombre d'observations non CQI, détectées comme valeurs aberrantes.

— TN : le nombre de vrais négatifs, c'est-à-dire le nombre d'observations non CQI, détectées comme valeurs non aberrantes.

| Réalité \ Identification | Valeurs aberrantes | Valeurs non aberrantes |
|---|---|---|
| CQI | TP | FN |
| Non CQI | FP | TN |

TABLE 1.1 – Classification des résultats des méthodes de détection de valeurs aberrantes.

Généralement les critères suivants, qui sont utilisés dans la Section 2.5, sont également calculés afin de synthétiser les résultats de la classification :

$$\begin{aligned} \text{Sensibilité} &= \frac{TP}{TP+FN} \\ \text{Spécificité} &= \frac{TN}{FP+TN} \\ \text{1-Spécificité} &= 1-\frac{TN}{FP+TN}=\frac{FP}{FP+TN} \end{aligned} \tag{1.12}$$

D'un point de vue industriel, la sensibilité est considérée comme le taux de bonne détection (DR) et 1−Spécificité comme le taux de fausses alarmes (FAR). De façon optimale, le DR devrait être de 100 % et le FAR de 0 %. Dans la pratique, en fonction des industries, le taux acceptable de fausses alarmes est variable. Il est compris entre 1 à 2% dans l'automobile et il peut être un peu plus élevé dans l'industrie spatiale s'il permet d'augmenter significativement le taux de bonne détection.

#### Contributeurs ou signature des défauts

Un autre point très important dans l'industrie est la nécessité de pouvoir remonter à la cause du défaut. Les ingénieurs de tests ont besoin de savoir sur quels tests la puce s'est comportée de manière anormale. Dans le contexte présenté ici, seules des méthodes de type non-supervisé sont introduites, ce qui signifie qu'il n'est pas possible d'apprendre des données. Les observations détectées comme atypiques le sont donc seulement pour des raisons statistiques. Or ces méthodes vont identifier sans distinction des observations prenant des valeurs extrêmes sur certains tests bruités et de « vraies » anomalies. L'ingénieur de test, en s'appuyant sur les tests qui expliquent l'anormalité des observations, est quant à lui capable de faire la différence entre ces deux types d'atypiques. La combinaison de la puissance de l'analyse exploratoire des méthodes statistiques et de l'expertise des ingénieurs permet donc de maximiser le taux de détection tout en minimisant le taux de fausses alarmes.

À titre d'illustration, pour l'ACP, Mnassri et al. (2008) ont cherché à déterminer la contribution des mesures initiales aux distances SD et OD résultantes de l'analyse de données réelles issues du processus de fabrication de l'entreprise STMicroelectronics. Pour ce faire il ont cherché à déterminer dans quelle mesure chaque test influence les distances SD et OD (voir aussi Harkat et al. (2002) pour une approche alternative).

Dans la communauté informatique, la phase d'interprétabilité des défauts est connue sous le nom *Intensional Knowledge* (en anglais), comme expliqué dans Aggarwal (2013, 2017). Un des travaux les plus notables dans le domaine est celui de Knorr and Ng (1999). Toutefois, la communauté statistique s'intéresse également à cette problématique comme le montrent Debruyne et al. (2017) dans leurs recherches.

### 1.2.5 Exemple réel de l'industrie des semi-conducteurs et mise en œuvre en R

Dans cette section nous souhaitons comparer les différentes méthodes de détection d'observations atypiques sur un exemple réel de l'industrie des semi-conducteurs à l'aide du logiciel R (R Core Team, 2017). Pour information, le domaine des semi-conducteurs fabrique des produits critiques et est donc soumis à une sévère politique de confidentialité. Réussir à avoir des données en accès libre est une véritable opportunité.

#### Mise en œuvre en R

De nombreux logiciels propriétaires mettent en œuvre les méthodes présentées précédemment. Les entreprises créent généralement leur propre logiciel afin de faciliter le traitement et la traçabilité des données des puces testées. Toutefois ces méthodes sont également disponibles sous le logiciel R. La fonction de base *mahalanobis* permet de calculer la distance de Mahalanobis dans ses versions robustes ou non, en précisant les estimateurs de position et de dispersion que l'on souhaite utiliser. Les quantiles des différentes distributions sont également disponibles afin de pouvoir déterminer les valeurs seuils permettant l'identification des observations anormales. Des ajustements de ces limites sont disponibles dans le package mvoutlier pour identifier des valeurs aberrantes seulement dans les queues de distribution. Enfin le package CerioliOutlierDetection permet de tester simultanément si des observations sont atypiques en se basant sur les distances de Mahalanobis calculées avec les estimateurs MCD. En ce qui concerne l'ACP, nous avons choisi de travailler avec le package rrcov qui calcule les distances SD et OD, ainsi que le package robustbase pour le calcul des estimateurs robustes.

#### Exemple réel : HTP

L'exemple choisi est un jeu de données de l'industrie des semi-conducteurs, nommé HTP, qui est disponible dans le package ICSOutlier. Il contient 902 pièces high-tech conçues pour des produits de consommation. 88 tests sont effectués pour assurer une haute qualité de production. Seules les pièces considérées comme fonctionnelles et qui ont été vendues sont présentes ici. Deux pièces se sont avérées être des incidents de qualité client (CQIs). Par conséquent, ces deux éléments peuvent être considérés comme des anomalies. Le but de l'analyse est de pouvoir déterminer si une méthode statistique aurait pu les identifier comme telles avant la vente.

Dans cette étude, on compare les résultats obtenus avec la distance de Mahalanobis et l'ACP. Les variantes robustes de ces méthodes multivariées sont également analysées en considérant les estimateurs MCD repondérés avec un point de rupture de 25%. L'ACP n'étant pas affine invariante, la question de la standardisation des données est pertinente. Toutefois, ici les 88 tests relèvent du même type de mesures et sont donc dans des unités similaires, ce qui ne rend pas la standardisation nécessaire. En ce qui concerne le choix du nombre de composantes à retenir, les fonctions du package rrcov proposent une décision automatique basée sur deux critères : ne garder que des valeurs propres assez grandes par rapport à la première, $\gamma_l/\gamma_1 \geq 10^{-3}$ puis parmi celles-ci expliquer 80% de la variance totale soit $k = \underset{l}{argmin}\{PCV(l) \geq 0.8\}$. L'identification des observations atypiques est réalisée au niveau de 2% pour pouvoir déduire le taux de fausses alarmes (FAR). Le cut-off des distances de Mahalanobis robustes a été ajusté par la technique de Green and Martin (2017b). Pour l'ACP, la détection basée sur les SD et les OD est effectuée au niveau $\alpha/2$ pour chacune des deux distances.

Les résultats sont illustrés sur la Figure 1.5. La première ligne fait référence aux méthodes non robustes et la deuxième ligne concerne les méthodes robustes. Les observations représentées en noir sont celles qui sont identifiées comme atypiques par la méthode considérée. Les symboles de type triangle permettent de repérer les deux observations de type CQI qu'il faut détecter.

Tout d'abord on s'aperçoit que toutes les méthodes permettent de détecter les deux CQIs avec un test à 2%. Toutefois le taux de fausses alarmes n'est clairement pas comparable entre les méthodes. Si on analyse plus en détail le graphique en haut à gauche représentant la distance de Mahalanobis de chaque observation, le taux de fausses alarmes est de 13.22%. Le CQI n°2 est clairement identifié comme observation anormale avec la distance la plus éloignée, alors que CQI n°1 est mélangé dans le reste des observations détectées comme atypiques. On aurait pu penser qu'utiliser des estimateurs robustes du MCD permettrait d'améliorer la détection puisqu'ils prémunissent contre les effets de masque ou de débordement, mais sur notre exemple, on constate l'effet inverse. Désormais le CQI n°2 n'est plus l'observation avec la distance la plus élevée et n'est donc plus aussi clairement identifié comme atypique. De plus, bien que le cut-off utilisé pour l'identification des observations anormales soit ajusté, le taux de fausses alarmes a presque doublé par rapport à la version non robuste. Cet exemple illustre les limites de l'utilisation de la distance de Mahalanobis ou du $T^2$ de Hotelling et leurs versions robustes dans le domaine des semi-conducteurs. Ce phénomène peut s'expliquer d'une part par la nature même des observations atypiques, qui ne se comportent anormalement que dans un sous-espace et non pas forcément sur tous les tests. D'autre part, les analyses basées sur des estimateurs robustes de position et de dispersion se focalisent sur la partie la plus centrale des données (composée ici d'environ 75% des observations). Or dans cet exemple, la proportion d'anomalies est très faible (bien inférieure à 2%), il est donc peu probable que des effets de masque ou de débordement apparaissent. Les résultats de l'ACP en haut à droite semblent confirmer cette hypothèse. Avec seulement 3 composantes parmi les 88, les deux CQIs ap-

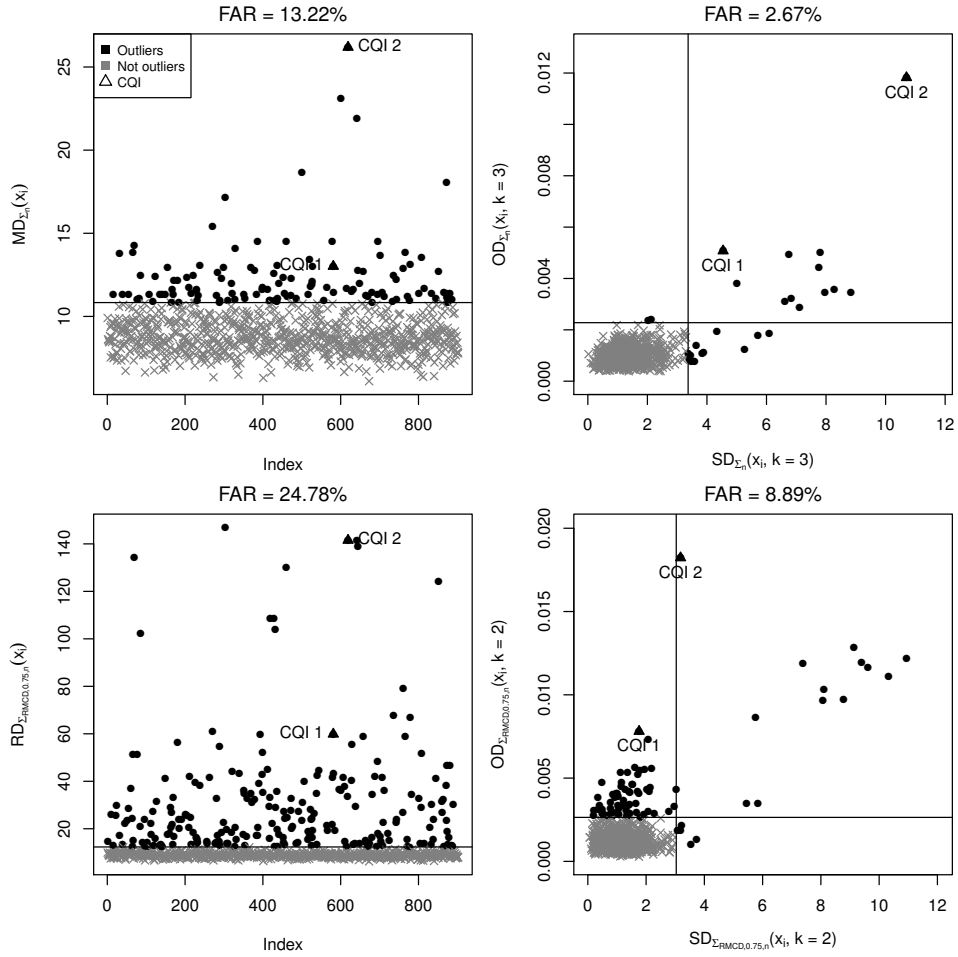



FIGURE 1.5 – Comparaison des résultats de détection des CQIs des données HTP. La première colonne représente les distances de Mahalanobis calculées de manière non robuste puis avec les estimateurs du MCD. La deuxième colonne donne les distances SD et/ou OD (cf Remarque 2) obtenues après une ACP non robuste puis robuste.

paraissent comme atypiques à la fois sur les distances OD et SD, tout en gardant un taux de fausses alarmes relativement faible, inférieur à 3%. Toutefois, ici aussi l'utilisation de la version robuste ne permet plus d'identifier aussi aisément les deux CQIs et amènent à considérer presque le triple de fausses alarmes.

Avec ce petit exemple illustratif des phénomènes propres aux semi-conducteurs, on s'aperçoit que les méthodes les plus robustes ne sont pas forcément celles les plus performantes dans ce domaine. En effet, dans ce secteur, seul un faible pourcentage d'anomalies est présent et il n'est donc pas nécessaire de se prémunir des effets de masque ou de débordement qui sont peu probables. Toutefois, il est difficile de pouvoir généraliser ces résultats car très peu de données sont publiées pour des raisons de confidentialité ou de disponibilité de l'information des incidents de qualité client. En ce qui concerne les méthodes univariées, Moreno-Lizaranzu and Cuesta (2013) ont réussi à mener une des premières analyses globales. En se basant sur des données de production de l'entreprise Freescale (maintenant NXP), ils ont testé différentes méthodes sur plus de 205 671 puces dont 26 CQIs. Ils montrent que, quand le nombre de tests effectués est de l'ordre de 1500, utiliser

la méthode PAT mène à détecter 34.6% des CQIs mais en éliminant 40% de pièces, ce qui est contreproductif et moins efficace qu'une identification aléatoire des mauvaises pièces. De manière générale, les experts estiment qu'une des méthodes les plus performantes dans le domaine des semi-conducteurs est le PAT dynamique (voir Section 1.2.2) avec une détection maximale de 5% des CQIs pour un taux de fausses alarmes de moins de 1%. Ces résultats laissent donc une nette marge d'amélioration dans l'identification des anomalies. La première étape consiste à dresser un état de l'art des méthodes existantes sans se restreindre au niveau du domaine d'application.

## 1.3 Approches en dimension standard : $n > p$

De très nombreuses méthodes de détection d'atypiques existent, même en se limitant aux approches non supervisées seulement dédiées aux variables quantitatives. De manière générale, et non plus en se focalisant sur le contrôle de qualité, nous nous concentrons sur les méthodes de détection usuelles applicables lorsque la taille de l'échantillon est supérieure à sa dimension, soit $n > p$.

Dans la première section, nous rappellons les difficultés pour définir une observation atypique ainsi que sur les caractéristiques attendues d'une méthode de détection. Ensuite, en se basant sur l'ouvrage très complet d'Aggarwal (2017), nous synthétisons les différentes approches existantes dans la littérature statistique et informatique. Nous rajoutons égalemment les caractéristiques importantes de chaque méthode ainsi que les packages R qui les mettent en œuvre.

### 1.3.1 Généralités

Tout d'abord, une observation atypique peut être définie comme étant une anormalité, une discordance, une déviation ou une anomalie. Toutes ces appellations ont en commun de renvoyer à un comportement qui semble être incompatible avec le reste de l'ensemble de données, comme défini dans Barnett and Lewis (1994). La subtilité cachée dans l'emploi de ces termes réside dans la caractérisation de l'observation atypique. Par exemple, Aggarwal (2017) explique que la définition d'atypicité est subjective et dépend du degré à partir duquel on considère qu'une déviation est assez importante par exemple. Généralement le terme d'observation « atypique » est utilisé pour définir n'importe quel comportement différent de l'ensemble du reste de la population, alors que le terme « anomalie » fait référence à quelque chose qui intéresse l'analyste, comme un incident de qualité dans le domaine des semi-conducteurs par exemple. Toutefois, dans le cadre des approches non-supervisées, la notion de déviation significativement intéressante n'est pas clairement définie. Les méthodes identifient donc seulement des observations atypiques sans faire la distinction entre celles qui sont de vraies anomalies ou celles produites à cause de tests bruités.

Ensuite, pour caractériser les individus, les méthodes de détection retournent communément un score d'atypicité pour chaque observation et/ou une identification binaire des observations comme normale ou non. Ce score permet de pouvoir ordonner les individus en fonction de leur atypicité. L'identification formelle est généralement basée sur la détermination statistique de la valeur seuil du score à partir duquel les observations sont considérées comme atypiques. D'un point de vue pratique, cette étape est cruciale pour la prise de décision. Toutefois, le seuil statistique utilisé ne permet pas d'assurer que les observations signalées soient significatives d'une réelle anormalité.

Enfin, nous pouvons dresser une liste de caractéristiques souhaitables pour les méthodes de détection. Tout d'abord, nous notons les propositions de Serfling and Mazumder (2013), à savoir : (i) la robustesse de la méthode en présence de valeurs aberrantes, (ii) l'invariance affine faible (i.e. qu'une transformation affine des coordonnées ne devrait pas affecter le classement relatif des scores d'atypicité et donc la comparaison des observations), (iii) l'efficacité computationnelle dans n'importe quelle dimension pratique, et (iv) la non-imposition de l'hypothèse de distributions elliptiques. De plus, Cerioli (2010) fait également remarquer qu'il est pertinent de tester l'absence d'observation atypique. Enfin, Aggarwal (2017) et Markou and Singh (2003) mettent en garde contre le nombre de paramètres à ajuster pour utiliser certaines méthodes. En effet, chaque choix va avoir des conséquences sur la performance de la détection et, dans le cas d'approches non-supervisées, il est souvent très difficile de connaître le paramétrage optimal de la méthode.

En pratique, aucune méthode ne remplit toutes ces caractéristiques et il est donc difficile de pouvoir déterminer celle qui est la meilleure. Les comparaisons des différentes approches sont donc relatives à l'importance donnée à chacun des six critères décrits ci-dessus. Dans notre contexte industriel, nous considérons qu'il est important que la méthode optimale soit affine invariante pour se prémunir des changements d'échelle et idéalement qu'elle soit sans paramétrage et rapide d'exécution pour pouvoir être utilisée par des ingénieurs de tests non statisticiens. Idéalement, la méthode doit être en mesure de tester l'absence d'observation atypique. Par contre, on ne se focalise donc pas sur l'hypothèse de la distribution des données car il est courant que les propriétés théoriques requièrent une distribution elliptique alors qu'en pratique les méthodes tendent à fonctionner correctement même lorsque les hypothèses ne sont pas vérifiées (voir Hastie et al. (2001) Section 4.3 pour l'analyse discriminante par exemple). Enfin, la faible proportion d'anomalies, typique des données de semi-conducteurs, rend non nécessaire l'utilisation de méthodes de détection très robustes, puisqu'il est peu probable que des effets de masque ou de débordement apparaissent.

### 1.3.2 Présentation succincte des méthodes

Dans cette section, nous proposons une synthèse des méthodes de détection d'observations atypiques. En plus des ouvrages mentionnés dans l'introduction (Hawkins, 1980; Barnett and Lewis, 1994; Aggarwal, 2013, 2017), de nombreuses revues de la littérature

existent déjà dont notamment celles de Markou and Singh (2003); Venkatasubramanian et al. (2003); Hodge and Austin (2004); Rousseeuw and Leroy (2005); Agyemang et al. (2006); Chandola et al. (2007); Cateni et al. (2008); Chandola et al. (2009); Kriegel et al. (2010); Singh and Upadhyaya (2012); Zimek et al. (2014a); Pimentel et al. (2014). Toutes ces revues regroupent les différentes méthodes en plusieurs classes. Toutefois, cette classification n'est pas homogène en fonction des chercheurs et de leur communauté (statistique ou informatique). Nous avons choisi ici de suivre celle proposée récemment par Aggarwal (2017), qui distingue les approches en trois groupes : celles basées (i) sur un modèle probabiliste, (ii) sur la détermination d'un sous-espace et (iii) sur la notion de proximité.

L'idée n'est pas de refaire le travail exhaustif réalisé par Aggarwal (2017) dans son ouvrage, mais de s'appuyer sur celui-ci pour proposer un tableau synthétique des méthodes non-supervisées existantes et leurs principales caractéristiques. Des références provenant principalement de la littérature statistique (liste non exhaustive) sont également ajoutées ainsi que les packages R mettant en œuvre ces méthodes. Il faut bien entendu noter que cette synthèse n'est qu'une image à un instant donné, car le développement de nouvelles méthodes et des packages R associés est particulièrement rapide. De plus, les méthodes considérées comme trop complexes sur le plan de calcul ne sont pas présentées plus avant.

#### Méthodes basées sur un modèle probabiliste

Les approches de détection basées sur les modèles probabilistes peuvent être regroupées en trois grandes classes : les valeurs extrêmes univariées, multivariées et la modélisation probabiliste de mélanges.

Si la distribution des données est connue, il est possible d'analyser les valeurs extrêmes par des tests statistiques et d'identifier les observations qui se comportent véritablement de manière anormale. Cette approche, appelée théorie des valeurs extrêmes, est principalement univariée mais adaptable au cas multivarié. Par exemple, la très connue distance de Mahalanobis, présentée en Section 1.2.3 et employée dans tous les chapitres de cette thèse, qui calcule l'éloignement des observations par rapport au centre de la distribution sous l'hypothèse de loi elliptique, peut être considérée comme une méthode linéaire de détection par valeurs extrêmes multivariées. Cette distance peut également être appréhendée comme un cas particulier de mesure de profondeur. En effet, elle ordonne les observations de manière multivariée par rapport à un concept de centralité, ce qui est la définition même de la notion de profondeur. Une autre approche multivariée consiste à considérer une méthode hybride, nommée ABOD (*Angle-based Outlier Detection*), qui se base sur des angles et des distances. La méthode est considérée comme hybride car les angles formés par les observations sont inversement pondérés par la distance entre les points, comme expliquée plus en détail en Section 1.4.2.

Comme le type de distribution est généralement inconnu, on peut réaliser des tests d'adéquation afin de déduire la loi potentielle sous-jacente ainsi que d'estimer ses paramètres. Dans le cas multivarié, une procédure souvent utilisée est la modélisation par un

mélange de gaussiennes à l'aide d'algorithme de type EM (pour *Expectation-Maximisation*), suivi du calcul de la probabilité des points à appartenir à ce mélange. Toutefois, comme le nombre de paramètres augmente avec la complexité de la distribution sous-jacente des données, des problèmes de sur-ajustement peuvent se produire. De plus, les paramètres de ces modèles sont souvent difficiles à interpréter du point de vue pratique de l'analyste. Or les experts du domaine ont besoin de déterminer si les observations détectées comme atypiques sont des anomalies liées à la fiabilité.

Le Tableau 1.2 permet de synthétiser les méthodes appartenant aux trois grandes approches de détection présentées succinctement précédemment ainsi que de préciser leurs principaux avantages et inconvénients par rapport à nos attentes décrites en Section 1.3.1. Les packages R mettant en œuvre les différentes méthodes listées sont également mentionnés.

| Caractéristiques des méthodes | Références |
|---|---|
| VALEURS EXTRÊMES UNIVARIÉES | |
| . Tests statistiques de discordance.<br>. Test de Grubb's.<br>. Tests en fonction de la distribution.<br>. Règles du boxplot.<br>. Cartes de contrôles.<br>R : alphaOutlier, outliers, extremevalues | Barnett and Lewis (1994); Beckman and Cook (1983); Gao and Tan (2006); Grubbs (1950, 1969); Laurikkala et al. (2000); Rocke (1989, 1992); Tatum (1997); Vargas N. (2003) |
| VALEURS EXTRÊMES MULTIVARIÉES | |
| **- Distance de Mahalanobis (MD),** | |
| . Hypothèse de normalité requise.<br>. Affine invariante.<br>. Scores d'atypicité et test formel d'identification.<br>. Pas de paramètres à ajuster.<br>. Calcul peu complexe en $O(p^2)$.<br>R : mvoutlier; rrcovHD | Becker and Gather (1999); Cerioli et al. (2009); Cerioli (2010); Geun Kim (2000); Gnanadesikan and Kettenring (1972); Laurikkala et al. (2000); Mardia et al. (1979); Rocke and Woodruff (1996); Wilks (1962) |
| **$T^2$ de Hotelling,** | |
| . Cartes de contrôles multivariées.<br>. Equivalente à la distance de Mahalanobis. | Hotelling (1931); Hotteling (1947); Jensen et al. (2007); Mnassri et al. (2008); Sullivan and Woodall (1996); Vargas N. (2003) |

| Caractéristiques des méthodes | Références |
|---|---|
| **Distance de Mahalanobis robuste** | |
| . Test d'absence d'atypiques possible.<br>. Cas particulier de profondeur.<br>R : mvoutlier; CerioliOutlierDetection; faoutlier; robustX; rrcovHD | Billor et al. (2000); Campbell (1980); Cerioli et al. (2009); Cerioli (2010); Filzmoser et al. (2005); Hadi et al. (2009); Hardin and Rocke (2005); Jobe and Pokojovy (2015); Mardia et al. (1979); Maronna and Zamar (2002); Rousseeuw and Van Zomeren (1990); Serfling (1980) |
| - **Profondeur** | |
| . Calcul peut être très complexe.<br>R : depth | Aggarwal (2013, 2017); Johnson and Wichern (1998); Kriegel et al. (2010); Ruts and Rousseeuw (1996) |
| - **Angles (ABOD)** | |
| . Approche hybride basée sur des distances et des angles.<br>. Seulement des scores d'atypicité, pas d'identification formelle.<br>. Calcul complexe en $O(n^3)$, mais des optimisations existent.<br>. Adaptée si $n < p$.<br>R : abodOutlier; HighDimOut | Kriegel et al. (2008); Campos et al. (2015); Kriegel et al. (2010); Pham and Pagh (2012); Radovanović et al. (2010); Laurikkala et al. (2000); Shyu et al. (2003); Hadi et al. (2009) |
| Modélisation probabiliste de mélanges | |
| . Nombreux paramètres à ajuster.<br>. Calcul complexe. | Dempster et al. (1977); Gao and Tan (2006); Kriegel et al. (2011) |

Table 1.2: Synthèse des méthodes de détection d'atypiques basées sur les modèles probabilistes et leurs propriétés.

### Méthodes basées sur la détermination d'un sous-espace

La deuxième approche telle que définie par Aggarwal (2017) concerne les méthodes basées sur les modèles linéaires de type régression ou détermination d'un sous-espace. Aggarwal (2017) interprète les modèles linéaires non pas avec une définition statistique classique mais en termes de variables latentes. Toutefois, comme nous nous intéressons exclusivement aux méthodes non-supervisées, nous ne présentons pas les méthodes de régression. Nous nous focalisons sur l'approche qui consiste à déterminer un sous-espace dans lequel le comportement atypique des observations est plus aisément identifiable.

Une des méthodes les plus connues est l'Analyse en Composantes Principales (ACP ou PCA en anglais), présentée en Section 1.2.3, qui cherche à résumer l'information concernant

la structure de variance-covariance des $p$ variables dans les premières $k < p$ composantes principales. Une fois que les observations sont projetées dans ce sous-espace, celles qui se comportent différemment sont considérées comme atypiques. Cette approche est donc particulièrement efficiente dès lors que les données sont très corrélées. Bien que cette méthode ne soit pas affine invariante mais orthogonale invariante, elle est une référence très utilisée en pratique et nous évaluons ses performances sur des exemples en Sections 1.2.5, 2.6 et 3.4.4.

L'ACP peut également être considérée comme un cas particulier de projections révélatrices (ou *Projection Pursuit* en anglais). Elles ont été introduites par Friedman and Tukey (1974) et comme le présentent entres autres Hadi et al. (2009), l'idée est de trouver un sous-espace de projection de dimension inférieure, dans lequel la structure intéressante apparaît. Dans le contexte de la détection d'atypiques, les méthodes optimisent un indice caractéristique de la présence de valeurs aberrantes afin de trouver la projection qui met le plus en évidence ces observations. Dans le cas de l'ACP, cet indice est la maximisation de la variance de chaque composante. D'après Huber (1985), les méthodes de ce type ont l'avantage de ne se focaliser que sur les représentations contenant la structure des données et de faire abstraction du bruit. Toutefois, les algorithmes pour ces méthodes sont très coûteux en temps de calcul et nous ne les utiliserons pas. Une méthode de ce type sera toutefois citée lorsque nous présenterons la méthode ROBPCA en Section 1.4.2.

Aggarwal (2017), dans la Section 3.6 de son ouvrage, appréhende un perceptron, la forme la plus simple d'un réseau de neurones, comme une ACP. Il explique en détail comment définir ce réseau de neurones dans un cadre non-supervisée, nommé auto-encodeur, puisque la phase d'apprentissage est réalisée de manière non supervisée. Finalement, pour une certaine fonction d'activation, il conclut à l'équivalence entre ce réseau de neurones et l'ACP si $p-1$ composantes sont sélectionnées. Aggarwal (2017), dans la Section 3.6 de son ouvrage, appréhende un perceptron, la forme la plus simple d'un réseau de neurones, comme une ACP. Il explique en détail comment définir ce réseau de neurones dans un cadre non-supervisée, nommé auto-encodeur, puisque la phase d'apprentissage est réalisée de manière non supervisée. Finalement, pour une certaine fonction d'activation, il conclut à l'équivalence entre ce réseau de neurones et l'ACP si $p-1$ composantes sont sélectionnées.

D'autres méthodes de factorisation de matrices que l'ACP existent et permettent de déterminer un sous-espace qui peut s'avérer intéressant pour mettre en évidence le comportement atypiques de certaines observations. La méthode ICS (Tyler et al., 2009), qui va faire l'objet des chapitres suivants de cette thèse, est une méthode de factorisation de matrices particulière. Quelques références sont données dans le Tableau 1.3 suivant. Ce tableau synthétise l'ensemble des méthodes, mentionnées ci-dessus, qui se basent sur la détermination d'un sous-espace.

| Caractéristiques des méthodes | Références |
|---|---|
| Analyse en Composantes Principales | |
| - **ACP** | |
| . Hypothèse : données dans un sous-espace de dimension $k < p$.<br>. Seulement orthogonale invariante.<br>. Choix du paramètre $k$.<br>. Interprétation assez difficile.<br>. Calcul peu complexe.<br>R : rrcov | Barnett and Lewis (1994); Campbell (1980); Dutta et al. (2007); Filzmoser and Todorov (2013); Fujimaki et al. (2005); Gnanadesikan and Kettenring (1972); Jolliffe (2002); Mnassri et al. (2008); Parra et al. (1996); Rousseeuw and Leroy (2005) |
| - **ACP robuste** | |
| . Mêmes caractéristiques que l'ACP.<br>R : rrcov; rrcovHD | Candès et al. (2011); Cardot and Godichon (2015); Devlin et al. (1981); Hubert et al. (2002, 2005); Kwitt and Hofmann (2006); Locantore et al. (1999); She et al. (2016); Shyu et al. (2003) |
| - **ACP non linéaire** (ex : ACP à noyau) | |
| . Adaptation possible de l'ACP pour des dépendances non linéaires. | Barnett and Lewis (1994); Rousseeuw and Leroy (2005) |
| Projections révélatrices | |
| . Hypothèse : données dans un sous-espace de dimension $k < p$.<br>. Choix de l'indice à optimiser.<br>. Calcul complexe.<br>R : rrcov; REPPlab | Maronna and Yohai (1995); Peña and Prieto (2001a,b); Croux and Ruiz-Gazen (2005); Ruiz-Gazen et al. (2010) |
| Réseaux de Neurones : perceptrons et auto-encodeurs | |
| . Cas particulier de l'ACP.<br>. Calcul complexe. | An and Cho (2015); Aggarwal (2017) |
| Factorisation de matrices | |
| - **ICS** | |
| . ACP Généralisée.<br>. Affine invariante.<br>. Calcul peu complexe.<br>. Robustification possible.<br>R : ICS; ICSOutlier; ICSShiny | Caussinus and Ruiz-Gazen (1990); Caussinus et al. (2003a); Penny and Jolliffe (1999); Tyler et al. (2009); Bookstein and Mitteroecker (2014); Alashwali and Kent (2016); Archimbaud et al. (2016); Fischer et al. (2017) |
| - **Autres méthodes** | |
| . Généralisation de méthodes comme l'ACP à d'autres fonctions objectifs.<br>. Robustification possible.<br>R : denoiseR | Farcomeni and Greco (2016); Josse and Sardy (2016); Josse et al. (2016a); Xiong et al. (2011) |

Table 1.3: Synthèse des méthodes de détection d'atypiques basées sur la détermination d'un sous-espace et leurs propriétés.

### Méthodes basées sur la notion de proximité

Contrairement à la détermination de sous-espaces contenant de l'information sur la structure des données, l'approche basée sur la proximité cherche des régions de l'espace dans lesquelles certaines observations vont être isolées. Ces individus sont ceux identifiés comme atypiques. Les principales méthodes présentées dans le Tableau 1.4 s'appuient sur la classification non supervisée (*clustering* en anglais), la densité ou les plus proches voisins.

La première méthode de classification non supervisée n'est pas dédiée à la détection d'observations atypiques, puisqu'elle permet de déterminer des groupes d'observations. Toutefois, elle peut s'avérer intéressante si on considère que les groupes contenant peu d'observations peuvent correspondre aux observations potentiellement atypiques. Par contre, seule une identification des observations atypiques est effectuée et aucune mesure d'atypicité n'est disponible. Dans le reste de ce manuscrit, cette méthode n'est donc pas comparée à d'autres approches.

À l'inverse, les méthodes basées sur la densité calculent un indice d'atypicité pour chaque observation mais ne permettent pas de les identifier de manière théorique. Une des méthodes les plus connues est le LOF (*Local Outlier Factor* en anglais) qui calcule le degré d'éloignement d'une observation à ses plus proches voisins en termes de densités locales. Cette méthode a été testée sur des données industrielles par l'entreprise ippon innovation au cours d'un travail préliminaire (Guillouet, 2012), mais les résultats n'ont pas permis d'améliorer la détection par rapport à l'algorithme déjà en place. De plus, comme les calculs peuvent être complexes et longs, cette méthode n'est pas détaillée plus avant et est seulement évaluée sur un exemple dans la Section 3.4.3.

Les méthodes basées sur les $k$ plus proches voisins utilisent la distance de chaque observation à ses $k$ plus proches voisins comme score d'atypicité. Plus la distance est importante, plus l'observation est isolée. Ces approches en termes de distances sont sans doute les plus utilisées pour leur facilité de mise en œuvre et leur interprétabilité en termes de variables initiales. Toutefois, les méthodes nécessitent généralement des temps de calcul assez importants et ne sont donc pas illustrées sur des exemples dans ce manuscrit.

| Caractéristiques des méthodes | Références |
|---|---|
| Classification non supervisée | |
| . Nombreux paramètres : choix du nombre de clusters, du modèle et de la méthode d'initialisation.<br>. Calcul peu complexe.<br>. Pas de scores d'atypicité.<br>R : CrossClustering; kmodR | Duda et al. (2012); Jain and Dubes (1988); Rousseeuw and Kaufman (1990); Smith et al. (2002) |

| Caractéristiques des méthodes | Références |
|---|---|
| DENSITÉ | |
| **- LOF** | |
| . Hypothèse : la densité autour d'un atypique est différente de celle autour de ses voisins.<br>. Détection d'atypiques locaux et globaux.<br>. Choix du nombre $k$ à ajuster.<br>. Seulement une mesure d'atypicité, pas d'identification des atypiques.<br>. Calcul souvent complexe.<br>. Nombreuses variantes.<br>R : DMwR2; Rlof | Breunig et al. (1999, 2000); Hadi et al. (2009); Tang et al. (2002); Zimek et al. (2012) |
| DISTANCES, KNN | |
| . Hypothèse : les atypiques ont un voisinage moins dense (ils sont éloignés de leurs voisins).<br>. Facilement généralisable à différents types de données.<br>. Calcul complexe en $O(n^2)$. | Bay and Schwabacher (2003); Campos et al. (2015); Ghoting et al. (2006); Knorr and Ng (1998, 1999); Pimentel et al. (2014); Ramaswamy et al. (2000); Rohlf (1975); Tao et al. (2006); Wu and Jermaine (2006) |

TABLE 1.4: Synthèse des méthodes de détection d'atypiques basées sur la notion de proximité et leurs propriétés.

Pour conclure, il apparaît qu'aucune méthode ne répond aux six caractéristiques décrites en introduction de la Section 1.3.2. En fait, chacune présente ses propres avantages et faiblesses dont il convient de tirer profit en fonction des applications. À partir de ce constat, un nouveau concept s'est développé : combiner les résultats de plusieurs méthodes afin d'améliorer la détection finale. Un vaste domaine de recherche est consacré à cet objectif, appelé *outlier ensembles* (Aggarwal and Sathe, 2017) ou méthodes d'ensemble en français. L'idée générale est d'appliquer différentes procédures de détection d'atypiques au jeu de données considéré, et de combiner les scores d'atypicité. Avant de pouvoir les agréger d'une quelconque manière, il convient de les normaliser car ceux-ci ne sont généralement pas comparables. En termes de fonction d'agrégation, la moyenne ou la maximisation sont les deux méthodes les plus couramment employées. Entre autres, Aggarwal (2017); Aggarwal and Sathe (2017); Dang et al. (2014); Gao and Tan (2006); Keller et al. (2012); Kriegel et al. (2011); Lazarevic and Kumar (2005); Müller et al. (2010b,a, 2011); Nguyen et al. (2010); Schubert et al. (2012); Zimek et al. (2012, 2013, 2014b) décrivent les principales méthodes utilisées dans la littérature. Ces approches sont relativement récentes car elles peuvent également résoudre les problèmes liés à la grande dimension, comme expliqué dans la section suivante.

## 1.4 Approches en grande dimension - faible taille d'échantillon (HDLSS) : $n < p$

Avec l'émergence des nouvelles technologies, mesurer des caractéristiques et les stocker est devenu de plus en plus facile, à tel point qu'aujourd'hui il est courant de devoir traiter des jeux de données avec plus de variables que d'observations. Ce contexte de « grande dimension - faible taille d'échantillon », plus connu sous son acronyme anglais HDLSS (*High dimension-low sample size*) intéresse fortement les communautés informatique et statistique. En effet, la majorité des méthodes de détection d'observations atypiques présentées dans la section précédente ne peuvent plus être appliquées principalement à cause du « fléau de la dimension ». Plus concrétement, ce phénomène, connu en anglais sous le nom de *curse of dimensionality*, regroupe différents challenges : l'effet de la concentration des distances, l'ajout d'attributs non pertinents pour la détection d'observations atypiques ou simplement les problèmes d'efficacité de certains algorithmes.

La première section caractérise plus précisément les différents challenges que l'on vient d'évoquer. Les sections suivantes se consacrent aux grandes approches utilisées dans ce contexte : l'analyse en dimension globale ou l'analyse en sous-espaces de l'espace originel. Dans ce travail, nous utilisons indistinctement les termes HDLSS et grande dimension pour désigner ce contexte particulier.

### 1.4.1 Le fléau de la dimension

Tout d'abord, Beyer et al. (1999) ont mis en évidence un problème de « concentration des distances » ou de « concentration des mesures d'atypicité » avec l'augmentation de la dimensionnalité. En considérant des distances calculées avec une norme $L_k$ et $k \geq 1$, ils montrent que toutes les paires de points de données deviennent presque équidistantes les unes des autres, et cela pour un grand nombre de distributions. En conséquence, en HDLSS, la notion de proximité ne s'applique plus car il n'est plus possible de discriminer des voisins comme proches ou lointains. De plus, sous l'hypothèse d'uniformité des données, Aggarwal et al. (2001) vont plus loin et démontrent que le problème de la pertinence de notions comme la proximité, la distance ainsi que les plus proches voisins, est en fait sensible à la valeur de $k$ lors de calculs de distances avec la norme $L_k$. Plus spécifiquement, ils stipulent que seules les normes $L_1$ et $L_2$, ne donnent pas des indices d'atypicité égaux pour toutes les observations. À partir de ce constat, ils suggèrent même d'utiliser une norme fractionnelle avec $0 < k < 1$.

Ce problème de concentration des distances n'apparaît que lorsque les variables ajoutées n'apportent pas d'information pertinente concernant l'atypicité des observations. Concrètement, si un individu se comporte de manière anormale sur tous les tests considérés alors on parle plutôt d'une bénédiction (*self-similarity blessing*) car l'information à retrouver est très présente dans les données initiale. Ce phénomène se confirme également si les attributs sont fortement corrélés entre eux. Toutefois, dans le contexte industriel, on

constate généralement la situation inverse avec la présence d'une grande proportion d'attributs non pertinents. En conséquence, les données deviennent de plus en plus éparses (*sparse data*) car le nombre d'observations est faible comparé au nombre de dimensions, les mesures d'atypicités basées sur des distances sont donc faussées puisque pratiquement identiques. Néanmoins, Zimek et al. (2012) notent que le classement est toujours raisonnable même s'il devient presque impossible de choisir un seuil basé sur ces distances pour identifier les observations atypiques. En grande dimension, le principal challenge est donc de trouver le sous-espace d'attributs pertinents qui met en évidence le comportement anormal de certains individus. Or, avec l'accroissement de la dimensionnalité, le nombre de sous-espaces à considérer augmente exponentiellement, ce qui rend cette recherche très complexe.

Enfin, concernant la grande dimension, une croyance assez répandue est que chaque point dans un espace en grande dimension est un atypique. En fait, Zimek et al. (2012) expliquent que ce n'est pas exactement le cas. Ils précisent que pour chaque observation, il est toujours possible de trouver un sous-espace dans lequel celle-ci apparaît comme anormale. Il faut donc être vigilant à ce problème de biais de sollicitation de données (*data-snooping bias*) qui peut être associé à du sur-ajustement (*overfitting*) et doit être correctement traité, principalement dans le cas de procédure d'apprentissage.

À la vue de ces challenges, Aggarwal and Yu (2001) considèrent que pour traiter correctement des jeux de données en grande dimension, les méthodes de détection doivent avoir les caractéristiques suivantes :

- Gestion efficace des problèmes liés à la présence de données clairsemées.
- Interprétabilité de l'anomalie, i.e. la raison pour laquelle on peut dire que cette observation s'est comportée différemment.
- Comparabilité de la mesure d'atypicité. Une distance calculée dans un sous-espace de dimension $k$ n'est pas directement comparable à celle calculée dans un sous-espace de dimension $k+1$, par exemple.
- Calcul peu complexe, et cela même pour des problèmes de très grande dimension. Par exemple, les algorithmes basés sur une exploration combinatoire de l'espace ne sont pas efficaces.
- Prise en compte du comportement local des données pour déterminer si une observation est atypique ou pas.

À ces propriétés proposées par des chercheurs de la communauté informatique, on peut vouloir rajouter le test de l'absence d'observation atypique, comme dans le cas des données en dimension standard. Par contre, la propriété d'affine invariance des scores est difficile à obtenir lorsque l'on analyse des données parcimonieuses. Dans ce contexte, nous pouvons être amenés à relâcher notre exigence et à considérer des méthodes qui sont invariantes par transformation orthogonale. On peut aussi s'intéresser, comme le proposent Serfling and Mazumder (2013), à une affine invariance « faible » qui conserve l'ordre des observations par rapport à une mesure d'atypicité.

Les revues de la littérature de Kriegel et al. (2010) et Zimek et al. (2012) suggèrent de classer les méthodes en deux grandes approches : les analyses basées sur l'ensemble des dimensions et celles basées sur la projection dans des sous-espaces de dimension inférieure. En suivant cette classification, nous présentons succinctement les méthodes qui semblent les plus pertinentes par rapport à notre contexte industriel et peu complexes en calculs.

### 1.4.2 Les analyses en dimension globale

Cette première approche, dite en dimension globale, rassemble des méthodes qui analysent l'espace dans sa dimension globale, i.e. sans avoir à traiter séparément des sous-espaces. Ce concept est le même que celui de la plupart des méthodes usuelles en dimension standard, à savoir celles basées sur la proximité, sur de la classification non supervisée, sur des distances ou sur l'ACP. On s'intéresse particulièrement aux adaptations des méthodes de type distance de Mahalanobis et ACP, qui semblent répondre le mieux à nos attentes dans le contexte industriel. On investigue également une autre démarche très utilisée en pratique qui consiste à prétraiter les données en les projetant dans un sous-espace de dimension égale au rang des données, puis d'appliquer les méthodes usuelles de détection en dimension standard. Les autres méthodes sont quant à elles trop complexes en calculs pour notre application. Toutefois, on n'écarte pas l'approche basée sur les angles qui est capable d'analyser des données HDLSS.

#### Méthode basée sur les angles : ABOD

La méthode hybride ABOD (*Angle-based Outlier Detection*), brièvement introduite dans la Section 1.3.2 est une des seules méthodes qui ne nécessite aucune adaptation pour pouvoir être utilisée sur des données en grande dimension. La mesure d'atypicité est un facteur nommé ABOF (*Angle-Based Outlier Factor*) qui mesure la variabilité du spectre des angles formés à partir du point $\mathbf{x}$ avec l'ensemble des autres observations de l'espace. Ces angles sont inversement pondérés par la distance entre les points, ce qui rend la méthode hybride. Cette approche est souvent considérée comme plus adaptée au cas de données en grande dimension que les méthodes calculant des distances. Toutefois, Radovanović et al. (2010) ont montré que les mesures basées sur les angles ne sont pas immunisées contre le fléau de la dimension en raison des effets de concentration dans la mesure du cosinus, ce qui impacte le spectre des angles. De plus, la distribution de l'indice d'atypicité est inconnue et il n'existe donc pas de règle d'identification des observations anormales. Enfin, la complexité des calculs en $O(n^3)$ est importante mais peut être réduite à l'ordre de $O(n^2)$ (Kriegel et al., 2008) ou $O(n \log n)$ (Pham and Pagh, 2012). L'exemple en section 3.4.3 permet d'illustrer cette méthode.

### Adaptation de la distance de Mahalanobis

La distance de Mahalanobis est une méthode très utilisée en pratique au vu de ses nombreux avantages. Elle ne peut malheureusement pas être calculée dans le cas de données en grande dimension. En effet, l'estimateur de la matrice de variance-covariance est nécessairement singulier à partir du moment où le nombre de dimensions dépasse le nombre d'observations et ne peut plus être inversé. De nombreux chercheurs ont donc proposé des solutions pour adapter cette distance au cas HDLSS.

Une des solutions les plus simples consiste à déterminer l'inverse généralisée de l'estimateur de covariance ou réussir à définir un estimateur de la matrice de variance-covariance qui soit inversible, comme le proposent Ledoit and Wolf (2004, 2012). Sinon, il est également possible de régulariser d'autres estimateurs plus robustes de la variance-covariance de manière à obtenir une estimation de la dispersion qui soit inversible (voir entre autres Ollila and Tyler (2014); Verbanck et al. (2015); Ro et al. (2015)).

### Adaptation de l'ACP

Certaines variantes de l'ACP présentent l'avantage d'être directement applicables sur des données HDLSS. Ces variantes regroupent les ACP basées sur des projections révélatrices (Filzmoser et al., 2008; Croux et al., 2007) et les ACP creuses (Bernard and Saporta, 2013; Croux et al., 2013; Hubert et al., 2016; Reynkens et al., 2015; Shen and Huang, 2008; Zou et al., 2006). Pour adapter la version classique de l'ACP, Ledoit and Wolf (2004, 2012); Ollila and Tyler (2014); Verbanck et al. (2015) proposent de considérer des estimateurs de dispersion régularisés, comme dans le cas de la distance de Mahalanobis.

Une autre approche, nommée ROBPCA, a été mise au point par Hubert et al. (2005) spécialement pour des données de type HDLSS. Cette méthode combine des idées de projections révélatrices et d'estimation robuste de la matrice de variance-covariance. Plus spécifiquement, les données sont projetées dans un sous-espace de dimension inférieure ou égale à $n - 1$ grâce à une décomposition en valeurs singulières. Après cette réduction de dimension, l'idée est d'identifier un sous-ensemble de $h$ observations les moins susceptibles d'être anormales selon l'estimateur de Stahel-Donoho, en se basant sur des projections révélatrices. La décomposition spectrale de la matrice de variance-covariance calculée à partir de ces $h$ observations va permettre de décider du nombre de composantes à retenir dans la suite de l'analyse. Les données sont également projetées sur le sous-espace formé par les premiers vecteurs propres de la décomposition. Enfin, un estimateur robuste de type MCD repondéré peut être utilisé pour calculer les composantes d'une ACP robuste. Les observations sont ensuite identifiées comme anormales à l'aide d'un diagnostic graphique qui prend en compte les distances SD et OD. Cette méthode, qui est devenue une référence dans le cas de données en grande dimension, est mise en œuvre dans le package rrcov et est testée sur des données réelles dans le Chapitre 5. Quelques améliorations de l'algorithme ont été proposées dans Engelen et al. (2005). Les propriétés de robustesse et

le comportement asymptotique de ROBPCA ont été étudiés par Debruyne and Hubert (2009).

**Réduction de la dimension comme prétraitement**

Enfin, il est également possible de ne considérer que la première étape de la méthode ROBPCA, qui consiste à prétraiter les données en réduisant leur dimensionnalité, afin d'être ensuite en mesure d'appliquer les méthodes standards de détection d'observations atypiques.

La décomposition en valeurs singulières, connue sous son acronyme anglais SVD (*Singular Value Decomposition*) est l'une des méthodes les plus populaires pour effectuer cette réduction. Les détails théoriques sont présentés dans la Section 4.3.3. En quelques mots, cette procédure consiste à transformer de manière affine les données afin de les projeter dans un sous-espace de dimension inférieure ou égale au rang $r$ des données. S'assurer de garder les $r$ premières composantes de la décomposition garantit de ne pas perdre d'information. Il est alors possible d'appliquer les méthodes classiques de détection d'atypiques comme la distance de Mahalanobis ou l'ACP par exemple.

Parmi d'autres, Filzmoser et al. (2008) proposent différents algorithmes basés sur des composantes obtenues après une réduction de dimension. Nous présentons ici seulement l'algorithme Sign1 qu'ils proposent et que nous testons dans la Section 3.4.2 de ce manuscrit. Cette méthode se base sur l'ACP sphérique proposée par Locantore et al. (1999) qui consiste à normaliser les données de manière robuste et à les projeter sur une sphère avant de calculer les composantes principales robustes qui en découlent. Elle garde $r = \min(n-1, p-1)$ composantes et calcule l'inverse de la matrice de variance-covariance empirique afin de pouvoir déduire les distances de Mahalanobis de chaque observation dans le cas de la grande dimension.

Néanmoins, Branco and Pires (2015) remarquent que si le rang des données $r$ est égal à $n-1$ et qu'on projette les données sur les $n-1$ premières composantes, alors les distances de Mahalanobis calculées à partir de ces nouvelles observations $\mathbf{x}_i^*$ sont constantes :

$$\mathrm{MD}^2_{\boldsymbol{\mu},\mathrm{COV}}(\mathbf{x}_i^*) = (\mathbf{x}_i^* - \boldsymbol{\mu})'\mathrm{COV}^{-1}(\mathbf{x}_i^* - \boldsymbol{\mu}) = \frac{(n-1)^2}{n} \tag{1.13}$$

Tyler (2010) note également que dans le cas où les données se trouvent en position générale, alors tous les estimateurs de dispersion affine équivariants sont proportionnels à la matrice de variance covariance. Une solution envisageable est de ne projeter les données que dans un sous-espace de dimension inférieur au rang $r$, mais cela à condition d'accepter de perdre éventuellement de l'information sur la structure des données. À ce sujet, She et al. (2016) mettent en garde contre cette pratique dans le cas où les observations atypiques sont présentes dans le sous-espace orthogonal complémentaire car les anomalies peuvent ne pas être détectées. La Section 4.3.3 discute également plus en détail des problèmes pouvant survenir lors de ce pré-traitement des données.

### 1.4.3 Les analyses de sous-espaces de l'espace originel

À l'inverse des méthodes présentées précédemment, une nouvelle approche remet en question l'analyse des données dans l'espace global. En effet, les données de type HDLSS sont souvent contaminées par un nombre non négligeable de variables non pertinentes pour l'identification d'anormalités. On peut donc légitimement supposer que les atypiques sont davantage visibles dans un sous-espace local de dimension inférieure, constitué des attributs pertinents. Une analyse en pleine dimension va masquer ce comportement déviant en dimension inférieure. Les méthodes basées sur la proximité qui prennent en compte toutes les dimensions ne sont donc plus adaptées. L'idée principale est de découvrir le sous-espace qui permet d'identifier les observations anormales. Toutefois, ce n'est pas une tâche aisée.

Tout d'abord, la recherche de sous-espaces dans l'espace total est très complexe en calculs surtout lorsqu'une approche combinatoire est utilisée. Ensuite, il est souvent difficile de réussir à sélectionner le « bon » sous-espace. D'autant plus qu'Aggarwal (2017) constate que l'omission de certaines variables pertinentes a des effets plus graves sur l'efficience de la détection que l'inclusion de variables non pertinentes dans l'analyse. Pourtant, réussir à identifier les sous-espaces appropriés rend plus aisée l'interprétabilité des anormalités, principalement quand ceux-ci sont décrits en fonction des attributs originels.

Aggarwal (2017) consacre le Chapitre 5 de son ouvrage à passer en revue les différentes méthodes permettant de détecter les observations atypiques en se basant sur des sous-espaces de l'espace originel. Nous évoquons ici seulement les idées directrices de ces approches ainsi que celles présentées par Zimek et al. (2012) dans leur revue de la littérature. Les détails techniques peuvent être trouvés dans les articles mentionnés. Les méthodes ne sont pas présentées plus avant car ce n'est pas l'axe de recherche qui a été privilégié dans le cadre de notre contexte au vu de sa complexité.

#### *Feature Bagging* et méthodes d'ensembles

Une des approches les plus simples consiste à constituer des sous-espaces de $r < p$ attributs, d'analyser ces sous-espaces avec les méthodes classiques de détection d'atypiques et de combiner les résultats avec des méthodes d'ensembles. Toutefois, plusieurs difficultés apparaissent. Elles concernent principalement trois aspects : (i) le choix du nombre de sous-espaces à considérer, (ii) si l'on veut partitionner l'espace ou faire des tirages avec remise et (iii) l'agrégation et la normalisation des résultats. Le dernier point est particulièrement important si les scores d'atypicités sont calculés dans des sous-espaces de différentes dimensions. À titre d'exemple, le package HighDimOut met en œuvre la méthode présentée par Lazarevic and Kumar (2005) qui consiste à tirer à chaque itération un échantillon aléatoire de $p/2 < r < p$ variables et de calculer des scores pour chaque observation à l'aide de la méthode LOF. La mesure d'anormalité finale est la somme cumulative de tous les scores obtenus à chaque itération. L'algorithme confidentiel GAT développé dans le

cadre de ce travail de thèse et introduit dans le Chapitre 5, est également basé sur une méthode d'ensembles.

#### Autres algorithmes

De très nombreux algorithmes ont été développés principalement par la communauté informatique. Ils se basent sur les notions classiques présentées à la Section 1.3.2, à savoir la distance entre observations, la densité, la classification non supervisée, les probabilités ou les projections révélatrices. Entre autres, on peut mentionner les méthodes HOS (*High-dimensional Outlying Subspaces*) de Zhang et al. (2004) ; SOD (*Subspace Outlier Degree*) de Kriegel et al. (2009) mise en œuvre dans le package HighDimOut ; HiCS (*High-Contrast Subspaces*) de Keller et al. (2012) ; OutRank (*Projected Clustering Ensembles*) de Muller et al. (2008, 2012) ; OUTRES (*Local Selection of Subspace Projections*) de Müller et al. (2010b, 2011) et COP (*Correlation Outlier Probability*) de Kriegel et al. (2012).

Par contre, comme en dimension standard, aucune de ces méthodes ne remplit toutes les conditions décrites en Section 1.4.1. Ces algorithmes souffrent en effet d'au moins un des trois inconvénients majeurs suivants : (i) la non normalisation des scores d'atypicité avant agrégation, (ii) une recherche pas assez extensive des sous-espaces, ce qui amène à ne pas pouvoir identifier les observations atypiques ou (iii) ils engendrent des calculs trop complexes.

En conclusion, il apparaît clairement que la détection d'observations atypiques dans le cas de la grande dimension est un sujet d'actualité particulièrement investigué par la communauté statistique comme informatique et présentant de nombreux challenges. Toutefois, jusqu'à présent, aucune méthode proposée ne permet de répondre à toutes les propriétés souhaitables mentionnées dans la Section 1.4.1, à savoir l'invariance par transformation affine ou orthogonale, le test d'absence d'observation atypique, l'interprétabilité des mesures d'atypicité ou encore la faible complexité de la méthode. Bien qu'il semble illusoire d'arriver à développer une méthode qui remplisse tous ces objectifs, de nombreuses avancées sont encore possibles dans ce domaine.

## 1.5 Conclusion et perspectives

La première partie de ce chapitre s'est focalisée sur le contrôle de la qualité dans le contexte industriel. Cette partie a mis en évidence que les standards du domaine automobile préconisent uniquement l'utilisation de méthodes univariées. Toutefois, comme ces approches ne sont pas satisfaisantes car elles engendrent trop de rejets pour un niveau de détection acceptable, certaines entreprises ont recours à des méthodes de détection multivariées de type distance de Mahalanobis ou ACP, mais qui ne remplissent pas non plus toutes leurs attentes. Après une étude extensive de la littérature des différentes approches non-supervisées de détection d'atypiques, nous concluons qu'aucune méthode ne remplit tous les critères attendus. La suite du travail de thèse va donc se consacrer à proposer

une méthode de détection dont les propriétés à privilégier vont être déterminées par les caractéristiques du domaine d'application.

Dans le cas de la détection de défauts industriels qui nous intéresse, la principale particularité est le faible pourcentage d'anomalies à identifier (généralement $< 2\%$) dont l'atypicité est contenue dans un sous-espace de petite dimension en comparaison du nombre total de variables. Comme les risques d'effets de masque ou de débordement sont peu probables, la robustesse de la méthode n'est donc pas une priorité. Par contre, l'affine invariance de la méthode est très importante car les données réelles analysées peuvent être dans des unités très différentes et l'échelle des valeurs ne doit pas affecter les résultats. De plus, comme le nombre de tests sur les composants électroniques est en constante augmentation, il est nécessaire que la méthode soit peu complexe en termes de calculs et qu'elle ne dépende pas d'un grand nombre de paramètres. Enfin, il faut également qu'elle soit en mesure d'analyser des jeux de données contenant des variables colinéaires ou en plus grand nombre que les observations.

Pour répondre au mieux à ces spécificités, dans la suite de la thèse nous nous concentrons principalement sur la méthode ICS. Tout d'abord, dans le Chapitre 2 nous développons une méthodologie adaptée à la détection d'observations atypiques basée sur cette méthode affine invariante que nous comparons à certaines des approches présentées dans ce chapitre. Nous proposons également une mise en œuvre en R à travers deux packages présentés dans le Chapitre 3. Ensuite, dans le Chapitre 4, nous adaptons la méthode ICS dans le cas où des variables sont colinéaires ou en plus grand nombre que les observations. Enfin, dans le dernier chapitre, nous présentons brièvement l'algorithme confidentiel développé pour l'entreprise ippon innovation, qui inclue une version adaptée de la méthode ICS et que nous appliquons à des données réelles de l'industrie spatiale.

# Chapter 2

# Multivariate Outlier Detection with ICS

This chapter is a reprint of Archimbaud, A., Nordhausen, K., and Ruiz-Gazen, A. (2016). ICS for Multivariate Outlier Detection with Application to Quality Control. *submitted*.

**Abstract**

*In high reliability standards fields such as automotive, avionics or aerospace, the detection of anomalies is crucial. An efficient methodology for automatically detecting multivariate outliers is introduced. It takes advantage of the remarkable properties of the Invariant Coordinate Selection (ICS) method. Based on the simultaneous spectral decomposition of two scatter matrices, ICS leads to an affine invariant coordinate system in which the Euclidian distance corresponds to a Mahalanobis Distance (MD) in the original coordinates. The limitations of MD are highlighted using theoretical arguments in a context where the dimension of the data is large. Unlike MD, ICS makes it possible to select relevant components which removes the limitations. Owing to the resulting dimension reduction, the method is expected to improve the power of outlier detection rules such as MD-based criteria. It also greatly simplifies outliers interpretation. The paper includes practical guidelines for using ICS in the context of a small proportion of outliers which is relevant in high reliability standards fields. The choice of scatter matrices together with the selection of relevant invariant components through parallel analysis and normality tests are addressed. The use of the regular covariance matrix and the so called matrix of fourth moments as the scatter pair is recommended. This choice combines the simplicity of implementation together with the possibility to derive theoretical results. A simulation study confirms the good properties of the proposal and compares with other scatter pairs. This study also provides a comparison with Principal Component Analysis and MD. The performance of our proposal is also evaluated on several real data sets using a user-friendly R package accompanying the paper.*

**Keywords:** Affine Invariance, Mahalanobis Distance, Principal Component Analysis, Scatter Estimators, Unsupervised Outlier Identification.

## Sommaire


**2.1 Introduction . . . . . 47**
**2.2 Behavior of the Mahalanobis distance in large dimension . . . 49**
**2.3 Invariant Coordinate Selection . . . . . 51**
2.3.1 Scatter matrices . . . . . 51
2.3.2 ICS principle . . . . . 53
**2.4 ICS implementation for outlier detection . . . . . 54**
2.4.1 The choice of the scatter pair . . . . . 54
2.4.2 The invariant components selection . . . . . 55
2.4.3 Outlier identification . . . . . 56
**2.5 Simulations . . . . . 56**
2.5.1 Simulation framework . . . . . 56
2.5.2 Selecting the invariant components . . . . . 58
2.5.3 Detecting outliers with ICS . . . . . 59
2.5.4 Comparing ICS with the Mahalanobis distance and PCA . . . . 60
**2.6 Data Analysis . . . . . 62**
2.6.1 Glass recycling . . . . . 63
2.6.2 Reliability Data . . . . . 64
2.6.3 High-tech parts . . . . . 65
**2.7 Conclusion and perspectives . . . . . 66**
**2.8 Appendix . . . . . 68**
2.8.1 Proof of Proposition 1 . . . . . 68
2.8.2 Derivation of the eigenvalues and eigenvectors of the simultaneous diagonalization of COV and $\text{COV}_4$ for particular mixtures . . . . 71

## 2.1 Introduction

Detecting outliers in multivariate data sets is of particular interest in many physical (Beckman and Cook, 1983), industrial, medical and financial applications (Aggarwal, 2017). Some classical statistical detection methods are based on the Mahalanobis distance and its robust counterparts (see e.g. Rousseeuw and Van Zomeren (1990), Cerioli et al. (2009), Cerioli (2010)) or on robust principal component analysis (see e.g Hubert et al. (2005)). One advantage of the Mahalanobis distance is its affine invariance while Principal Component Analysis (PCA) is only invariant under orthogonal transformations. For its part, PCA allows some components selection and facilitates the interpretation of the detected outliers. All these methods are adapted to the context of casewise contamination while other methods are adapted to the case of cellwise contamination (see e.g. Agostinelli et al. (2015) and Rousseeuw and Van den Bossche (2017)). Furthermore, several other recent references tackle the problem of outlier detection in high dimension where the number of observations may be smaller than the number of variables (see e.g. Croux et al. (2013) and Hubert et al. (2016)).

In the present paper, we propose an alternative to the Mahalanobis distance and to PCA, in a casewise contamination context and when the number of observations is larger than the number of variables. As stated in Tarr et al. (2016) on page 405: "the cellwise contamination is prevalent in large, automatically generated data sets, found in data mining and bioinformatics, where there is often little quality control over the inputs". In the present paper, the focus is on applications with high level of quality control, such as in the automotive, avionics or aerospace fields, where only a small proportion of outliers, up to 2%, is plausible. From our experience in such application fields, a small proportion of parts potentially defective are to be detected with very limited false detection. Moreover, even if in such fields the trend is to increase the number of measurements, there are still many applications where the number of observations is larger than the number of variables and, in such a context, an improved affine invariant method with an easy characterization of the outliers is still of interest.

The method we consider is the Invariant Coordinate Selection (ICS) as proposed by Tyler et al. (2009). The principle of ICS is quite similar to Principal Component Analysis (PCA) with coordinates or components derived from an eigendecomposition followed by a projection of the data on selected eigenvectors. However, ICS differs in many respects from PCA. It relies on the simultaneous spectral decomposition of two scatter matrices instead of one for PCA. While principal components are orthogonally invariant but scale dependent, the invariant components are affine invariant for affine equivariant scatter matrices. Moreover, under some elliptical mixture models, the Fisher's linear discriminant subspace coincides with a subset of invariant components in the case where group identifications are unknown (see Theorem 4 in Tyler et al. (2009)). This remarkable property is of interest for outlier detection since outliers can be viewed as data observations that differ from the remaining data and form separate clusters.

Despite its attractive properties, ICS has not been extensively studied in the literature on outlier detection. An early version of ICS was proposed in Caussinus and Ruiz-Gazen (1990) for multivariate outlier detection and studied further in e.g. Penny and Jolliffe (1999) and Caussinus et al. (2003b) for two specific scatter matrices. Recent articles by Nordhausen et al. (2008) and Tyler et al. (2009) argue that ICS is useful for outlier detection. However, a thorough evaluation of ICS in this context is still missing and the present paper is a first step aimed at filling the gap.

Our first objective is to explain the link between ICS and the Mahalanobis distance. First, we prove that Euclidian distances calculated using all invariant components are equivalent to Mahalanobis distances calculated using the original variables. Then, in the case where the number of variables is large (but still with a larger number of observations) and outliers are contained in a small dimensional subspace, we recommend selecting a small number of invariant components. Such a selection is motivated by looking at the approximate probability in large dimension of the difference between the Mahalanobis distance of an outlying observation and the Mahalanobis distance of an observation from the majority group. We prove that this probability decreases toward zero when the dimension increases which is undesirable. This shortcoming can be avoided by a proper selection of invariant components.

Then, we focus on the case where the majority of the data behaves in a regular way and only a small fraction of the data might be considered outliers. Examples include, for instance, financial fraud detection or production error identification in industrial processes where there is a high level of quality control. Our goal is to provide practical guidelines for using ICS in this context of unsupervised detection of a small proportion of outliers. More precisely, we implement and compare different pairs of scatter matrices estimators and different methods for selecting relevant invariant components through an extensive simulation study. We consider several contamination models with a percentage of contamination equal to 2%, which is relevant in the context of high reliability standards fields. Results are given in terms of true positive and false negative discoveries for several mixture models. We advocate a simple choice for the scatter matrices pair, namely the covariance and the fourth moment matrices. Such estimators are simple to implement and some theoretical results can be derived for some particular mixtures as detailed in 2.8.2. Regarding components selection, we recommend two methods: the so-called parallel analysis (Peres-Neto et al., 2005) and a skewness-based normality test. We also show that our proposal improves over the Mahalanobis distance criterion and over different versions of PCA through simulations and the use of three real data sets. One of the key benefits of our approach compared to competitors is its ability not to detect outliers when there is no outlier present in the data set, at least in the Gaussian case. When outliers are absent, the proposed procedure is likely to select none of the invariant components. Another practical benefit, as illustrated on one of the three real examples, is the ease of interpretation of the detected outliers using the selected invariant coordinates. Mimicking PCA, the user can draw some scatter plots of the invariant components or look at the correlations between

the invariant components and the original variables. More complex procedures (advocated for instance in Willems et al. (2009)) when using the Mahalanobis distance can thereby be avoided.

This article is organized as follows. In Section 2.2 we observe the behavior of the usual and the robust Mahalanobis distances for large dimensions when outliers lie in a small dimensional subspace. This result motivates the use of selected invariant components for outlier detection. ICS is described in a general framework in Section 2.3 and in the context of a small proportion of outliers in Section 2.4. Section 2.5 provides results from a simulation study and derives practical guidelines for the choice of the scatter matrices pair and the components selection method. Comparisons with the Mahalanobis distance and PCA are also provided. Three real data sets are analyzed in Section 2.6. Finally, conclusions and perspectives are drawn in Section 2.7. The proof of Proposition 1 is given in 2.8.1 while some additional propositions are given in 2.8.2. Supplementary material is also provided. It contains some scatterplot matrices to visualize the six simulated data sets and the R code to generate these data sets. It also includes the R code to reproduce the results of Table 4 for the Reliability data and the HTP data sets.

## 2.2 Behavior of the Mahalanobis distance in large dimension

Let $\mathbf{X} = (X_1, \ldots, X_p)'$ be a $p$-multivariate real random vector and assume the distribution of $\mathbf{X}$ is a mixture of $(q+1)$ Gaussian distributions with $q+1 < p$, different location parameters $\boldsymbol{\mu}_h$, for $h = 0, \ldots, q$, and the same definite positive covariance matrix $\boldsymbol{\Sigma}_W$:

$$\mathbf{X} \sim (1-\epsilon)\,\mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_W) + \sum_{h=1}^{q} \epsilon_h\,\mathcal{N}(\boldsymbol{\mu}_h, \boldsymbol{\Sigma}_W) \tag{2.1}$$

where $\epsilon = \sum_{h=1}^{q} \epsilon_h < 1/2$.

Such a distribution can be interpreted as a model for outliers where the majority of the data follows a given Gaussian distribution and outliers are clustered in $q$ clusters with Gaussian distributions with different locations than the majority group. This model is a generalization of the well-known mean-shift outlier model to more than two groups.

For such a model, the mean is $\boldsymbol{\mu}_{\mathbf{X}} = (1-\epsilon)\,\boldsymbol{\mu}_0 + \sum_{h=1}^{q} \epsilon_h \boldsymbol{\mu}_h$, the within covariance matrix is $\boldsymbol{\Sigma}_W$, the between covariance is $\boldsymbol{\Sigma}_B = (1-\epsilon)(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_{\mathbf{X}})(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_{\mathbf{X}})' + \sum_{h=1}^{q} \epsilon_h (\boldsymbol{\mu}_h - \boldsymbol{\mu}_{\mathbf{X}})(\boldsymbol{\mu}_h - \boldsymbol{\mu}_{\mathbf{X}})'$, where the prime symbol denotes the transpose vector or matrix, and the total covariance matrix is $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_B + \boldsymbol{\Sigma}_W$. Let us consider the following squared Mahalanobis distances:

$$d^2(\mathbf{X}) = (\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})' \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}}) \tag{2.2}$$

$$d_R^2(\mathbf{X}) = (\mathbf{X} - \boldsymbol{\mu}_0)' {\boldsymbol{\Sigma}_W}^{-1} (\mathbf{X} - \boldsymbol{\mu}_0). \tag{2.3}$$

These distances are affine invariant in the sense that $d^2(\mathbf{AX}+\mathbf{b}) = d^2(\mathbf{X})$ and $d_R^2(\mathbf{AX}+\mathbf{b}) = d_R^2(\mathbf{X})$, for any full rank $p \times p$ matrix $\mathbf{A}$ and any $p$-vector $\mathbf{b}$. The distance $d$ (resp. $d_R$) can be interpreted as a non-robust (resp. robust) Mahalanobis distance. Of course in practice, the different parameters are unknown and should be estimated, but the results we derive below give some intuition for the finite sample case. Let us now introduce distinct $p$-random vectors that would correspond to the different mixture components of $\mathbf{X}$. Let $\mathbf{X}_{no}$, where *no* stands for "non-outlier", follows a normal distribution $\mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_W)$ and $\mathbf{X}_{o,h}$, where $o$ stands for "outlier", follows a normal distribution $\mathcal{N}(\boldsymbol{\mu}_h, \boldsymbol{\Sigma}_W)$, with $h = 1, \ldots, q$. We assume that $\mathbf{X}_{no}$ and $\mathbf{X}_{o,h}$, for $h = 1, \ldots, q$, are independent, and we are interested in the behavior of the difference between the squared distance of $\mathbf{X}_o$ and of $\mathbf{X}_{no}$ for both Mahalanobis distances, when dimension $p$ increases. The distribution of these differences is not easy to handle especially for non-robust distance, but we can look at the asymptotic distribution for large $p$. When using the Mahalanobis distance or robust distance for outlier identification, we expect the probability of these differences to be large.

Under the mixture distribution defined previously, we have the following proposition. Its proof makes use of the Lindeberg-Feller central limit theorem for $p$ going to infinity and is given in 2.8.1.

**Proposition 1.** *Assume that $q$ is fixed, then under model (2.1):*

$$\frac{1}{2\sqrt{p}}\left(d^2(\mathbf{X}_{o,h}) - d^2(\mathbf{X}_{no}) - \mathbb{E}\left(d^2(\mathbf{X}_{o,h}) - d^2(\mathbf{X}_{no})\right)\right)$$

*and*

$$\frac{1}{2\sqrt{p}}\left(d_R^2(\mathbf{X}_{o,h}) - d_R^2(\mathbf{X}_{no}) - \mathbb{E}\left(d_R^2(\mathbf{X}_{o,h}) - d_R^2(\mathbf{X}_{no})\right)\right)$$

*converge in distribution to a standard Gaussian distribution when $p$ goes to infinity and the expectations $\mathbb{E}\left(d^2(\mathbf{X}_{o,h}) - d^2(\mathbf{X}_{no})\right)$ and $\mathbb{E}\left(d_R^2(\mathbf{X}_{o,h}) - d_R^2(\mathbf{X}_{no})\right)$ do not depend on $p$.*

Note that under model (2.1), the expectations can be made explicit but their expressions are complex and not detailed further.

The conclusion of Proposition 1 is that if outliers belong to a reduced dimension space (equal at most to $q$ in model (2.1)) and $p$ is large, then the probability that the Mahalanobis distance of an outlier exceeds the Mahalanobis distance of a non-outlier is small, because according to the asymptotic result, the variance of the differences increases when $p$ increases. This makes the outlier identification more difficult. If the $q$-subspace is known, it is easy to avoid the problem of the $p-q$ noisy dimensions by projecting the data set on this subspace and calculating a distance based on the $q$ dimensions that does not depend on $p$. This is exactly what ICS is all about, providing the data-analyst with the ability to select a subspace displaying the outliers in an unsupervised way, and project the data on this subspace. Figure 2.1 illustrates in some sense Proposition 1 results and the competitive advantage of ICS compared to the Mahalanobis distance on a simple artificial data set. This set which will be discussed in the simulation framework as "Case 1", contains 1000 observations with one cluster of 20 outliers location shifted and plotted in black. The

dimension of the data set increases from $p = 6$ on the left panels, to 25 on the middle ones and 50 on the right panels. The top panels plot the non-robust Mahalanobis distances using the usual covariance estimator while the middle panels plot robust Mahalanobis distances using the (reweighted) MCD estimator (Rousseeuw, 1986). The bottom panels plot the distances based on an automatic selection of invariant components for ICS with a pair of scatter matrices estimators detailed later in the present paper. When $p$ increases, it becomes more difficult to separate the outlying observations from the rest of the data using the Mahalanobis distances while the separation remains much better using selected invariant components. ICS is now detailed, and the choice of the scatter pair together with the selection of the invariant components is discussed.

## 2.3 Invariant Coordinate Selection

### 2.3.1 Scatter matrices

For a $p$-variate dataset $\mathbf{X}_n = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$, any $p \times p$ matrix symmetric and definite positive $\mathbf{V}(\mathbf{X}_n)$ is a scatter matrix if it is affine equivariant in the sense that

$$\mathbf{V}(\mathbf{X}_n\mathbf{A} + \mathbf{1}_n\mathbf{b}') = \mathbf{A}'\mathbf{V}(\mathbf{X}_n)\mathbf{A},$$

where $\mathbf{A}$ is a full rank $p \times p$ matrix, $\mathbf{b}$ a $p$-vector and $\mathbf{1}_n$ an $n$-vector full of ones.

The literature contains numerous scatter matrices suggestions (see Nordhausen and Tyler (2015) for a recent discussion and many references). Tyler et al. (2009) classify them into three classes depending on their robustness properties in terms of breakdown point and influence function. Class I scatter matrices have a zero or almost zero breakdown value and an unbounded influence function. Relevant scatter matrices from this class are the regular covariance matrix

$$\mathrm{COV}(\mathbf{X}_n) = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})',$$

where $\bar{\mathbf{x}}$ denotes the empirical mean, and the so called scatter matrix of fourth moments

$$\mathrm{COV}_4(\mathbf{X}_n) = \frac{1}{(p+2)n}\sum_{i=1}^{n} r_i^2(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})',$$

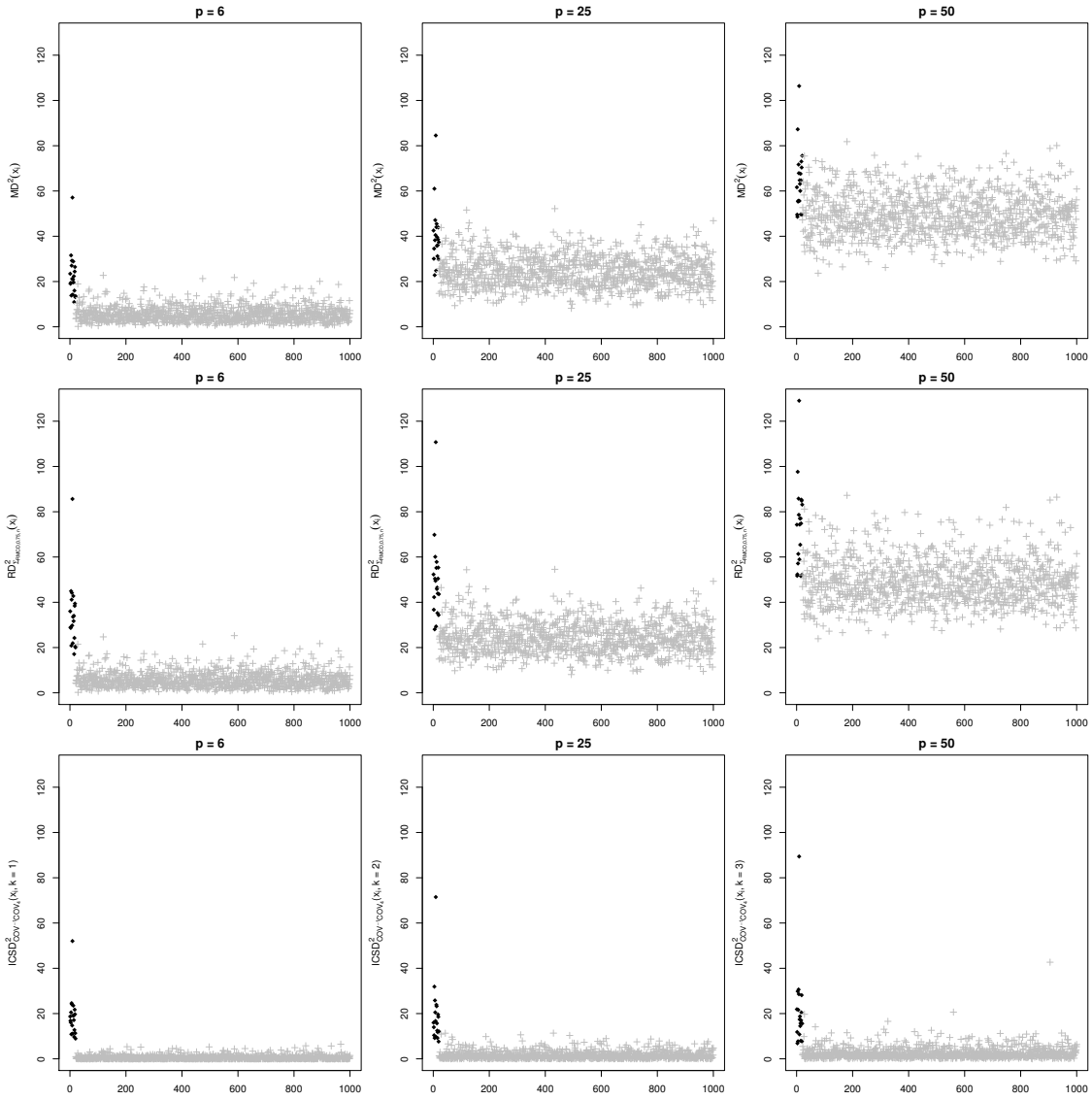
where $r_i^2 = (\mathbf{x}_i - \bar{\mathbf{x}})'\mathrm{COV}(\mathbf{X}_n)^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})$ is the classical squared Mahalanobis distance.

Class II consists of scatter matrices with a bounded influence function but a breakdown point not larger than $(p+1)^{-1}$. From this class, we will later use the following location and scatter matrix estimators defined through the implicit expressions:

$$\begin{aligned}
\mathbf{m}_C(\mathbf{X}_n) &= \sum_{i=1}^{n}(w(r_i^2)\mathbf{x}_i)/\sum_{i=1}^{n} w(r_i^2), \\
\mathrm{MLC}(\mathbf{X}_n) &= \frac{1}{n}\sum_{i=1}^{n} w(r_i^2)(\mathbf{x}_i - \mathbf{m}_C(\mathbf{X}_n))(\mathbf{x}_i - \mathbf{m}_C(\mathbf{X}_n))',
\end{aligned}$$

where $r_i^2 = (\mathbf{x}_i - \mathbf{m}_C(\mathbf{X}_n))'\mathrm{MLC}(\mathbf{X})^{-1}(\mathbf{x}_i - \mathbf{m}_C(\mathbf{X}_n)$ and $w(r_i^2) = (p+1)/(r_i^2+1)$.

Figure 2.1 – Squared distances (top: non-robust Mahalanobis, middle: robust Mahalanobis, bottom: Euclidian using invariant components with an automatic selection) for $p = 6$ (resp. 25 and 50) on the left (resp. middle and right) panels for a sample of 1000 observations drawn from a mixture of two normal distributions with the 20 location shifted observations in black.



These location and scatter matrix estimators are the maximum likelihood estimators of an elliptical Cauchy distribution and belong to the well-known class of M-estimators.

Class III scatter matrices are high-breakdown scatter matrices, and the reweighted Minimum Covariance Determinant (MCD) estimator is perhaps the most popular example from this class. For a given $h \in [0.5; 1]$, the $\text{MCD}_h$ searches for the $hn$ observations $\mathbf{X}_{hn}$ such that $\text{COV}(\mathbf{X}_{hn})$ has the smallest determinant and then is made more efficient by reweighting observations appropriately (see Rousseeuw (1986) and Cator and Lopuhaä (2012) for more details). The associated location estimator is a reweighted version of the average of the $hn$ observations.

While the Mahalanobis distance and PCA are based on one scatter matrix, ICS is based on the simultaneous use of two scatter matrices denoted below by $\mathbf{V}_1(\mathbf{X}_n)$ and $\mathbf{V}_2(\mathbf{X}_n)$. We will choose among the four estimators recalled previously and consider that class III estimators are more robust than class II, which are themselves more robust than class I. For the two class I estimators $\text{COV}(\mathbf{X}_n)$ and $\text{COV}_4(\mathbf{X}_n)$, we will consider $\text{COV}(\mathbf{X}_n)$ more robust than $\text{COV}_4(\mathbf{X}_n)$ because the norm of its influence function is smaller.

### 2.3.2 ICS principle

Formally, the goal of ICS is to find the $p \times p$ matrix $\mathbf{B}(\mathbf{X}_n)$ and diagonal matrix $\mathbf{D}(\mathbf{X}_n)$ such that:

$$\mathbf{B}(\mathbf{X}_n)\mathbf{V}_1(\mathbf{X}_n)\mathbf{B}(\mathbf{X}_n)' = \mathbf{I}_p \quad \text{and}$$

$$\mathbf{B}(\mathbf{X}_n)\mathbf{V}_2(\mathbf{X}_n)\mathbf{B}(\mathbf{X}_n)' = \mathbf{D}(\mathbf{X}_n).$$

$\mathbf{D}(\mathbf{X}_n)$ contains the eigenvalues of $\mathbf{V}_1(\mathbf{X}_n)^{-1}\mathbf{V}_2(\mathbf{X}_n)$ in decreasing order, while the rows of the matrix $\mathbf{B}(\mathbf{X}_n) = (\mathbf{b}_1, \ldots, \mathbf{b}_p)'$ contain the corresponding eigenvectors so that:

$$\mathbf{V}_1(\mathbf{X}_n)^{-1}\mathbf{V}_2(\mathbf{X}_n)\mathbf{B}(\mathbf{X}_n)' = \mathbf{B}(\mathbf{X}_n)'\mathbf{D}(\mathbf{X}_n).$$

Using any affine equivariant location estimator $\mathbf{m}(\mathbf{X}_n)$, the corresponding scores

$$\mathbf{Z}_n = (\mathbf{z}_1, \ldots, \mathbf{z}_n)' = (\mathbf{X}_n - \mathbf{1}_n\mathbf{m}(\mathbf{X}_n)')\mathbf{B}(\mathbf{X}_n)'$$

are the so-called invariant coordinates or components. They are affine invariant in the sense that

$$(\mathbf{X}_n^* - \mathbf{1}_n\mathbf{m}(\mathbf{X}_n^*)')\mathbf{B}(\mathbf{X}_n^*)' = (\mathbf{X}_n - \mathbf{1}_n\mathbf{m}(\mathbf{X}_n)')\mathbf{B}(\mathbf{X}_n)'\mathbf{J}$$

for $\mathbf{X}_n^* = \mathbf{X}_n\mathbf{A} + \mathbf{1}_n\mathbf{b}'$ with any full rank $p \times p$ matrix $\mathbf{A}$ and any $p$-vector $\mathbf{b}$. $\mathbf{J}$ is a $p \times p$ diagonal matrix with diagonal elements $\pm 1$, which means the invariant coordinates change at most their signs. For convenience, the dependence on $\mathbf{X}_n$ is dropped from the different matrices when the context is obvious.

Because $\mathbf{V}_1^{-1} = \mathbf{B}'\mathbf{B}$, the proof of the following proposition is immediate.

**Proposition 2.** *Let us consider an affine equivariant location estimator* $\mathbf{m}$ *and two scatter matrices* $\mathbf{V}_1$ *and* $\mathbf{V}_2$*. The Euclidian norm of an observation using its invariant coordinates corresponds to the Mahalanobis distance of this observation from* $\mathbf{m}$ *in the sense of* $\mathbf{V}_1$*.*

*Formally, it means that for observation* $i = 1, \ldots, n$,

$$\mathbf{z}_i'\mathbf{z}_i = (\mathbf{x}_i - \mathbf{m})'\mathbf{V}_1^{-1}(\mathbf{x}_i - \mathbf{m})$$

Tyler et al. (2009), p. 554, underlines the exchangeability between the roles of $\mathbf{V}_1$ and $\mathbf{V}_2$. However, as can be observed from Proposition 2, exchanging the two scatter matrices has an impact on the scale of the invariant coordinates and not only on the fact that the eigenvalues are the inverse of the others and the eigenvectors are in reverse order. In the following, we propose to use the location estimator associated with the scatter matrix $\mathbf{V}_1$ and take $\mathbf{V}_1$ "more" robust than $\mathbf{V}_2$, as Alashwali and Kent (2016) so that the Euclidian distance using all invariant components leads to more robust Mahalanobis distance.

## 2.4 ICS implementation for outlier detection

Identifying outliers with ICS is a three step procedure. The first step consists in choosing a pair of scatter matrices and calculating the invariant coordinates. The second step is the selection of the relevant invariant components and the calculation of the Euclidian norm of the $n$ observations using only the selected components. The last step is the outlier identification with the choice of a cut-off value $c$ such that observations with a norm larger than $c$ are flagged as outliers.

### 2.4.1 The choice of the scatter pair

When the objective is outlier detection, Caussinus and Ruiz-Gazen (1990, 2003) and Tyler et al. (2009) recommend using class I scatter estimators such as the classical one or some weighted scatter matrix. The main reason for this choice is that these estimators are simple and can be computed rapidly. Moreover, the nice properties of ICS given by Theorems 3 and 4 in Tyler et al. (2009) are true even for non-robust estimators such as the COV and $\text{COV}_4$ scatter matrices. For this particular pair, the formulation of Theorem 3, which applies to a mixture of two Gaussian distributions with different locations and proportional scatter matrices, can be made much more precise. As explained in Tyler et al. (2009), for a proportion of outliers smaller than $(3 - \sqrt{3})/6$ (around 21%), the first invariant component displays the outliers. Similar results can be derived for other particular mixtures as detailed in 2.8.2. More precisely, for a symmetrically contaminated Gaussian distribution with equal covariance matrices (which is similar to the so-called barrow wheel distribution), the first component will display the structure as soon as the contamination level is smaller than 33%. And this is also true for a Gaussian mixture with inflated variance in $q$ directions: as soon as the contamination is smaller than 50%, the invariant components associated with the $q$ largest eigenvalues will span the subspace of interest. For other scatter pairs, this calculus is not analytically tractable anymore and so a comparison through simulations is worthwhile. In the present paper, we propose comparing four pairs of scatter matrix estimators taken from the three different classes

based on simulations. The first pair is based on two class I estimators $\mathbf{V}_1 = \text{COV}$ and $\mathbf{V}_2 = \text{COV}_4$, while the others are based on class II and I with $\mathbf{V}_1 = \text{MLC}$ and $\mathbf{V}_2 = \text{COV}$, class III and I with $\mathbf{V}_1 = \text{MCD}$ and $\mathbf{V}_2 = \text{COV}$ and class III and II scatter estimators with $\mathbf{V}_1 = \text{MCD}$ and $\mathbf{V}_2 = \text{MLC}$.

### 2.4.2 The invariant components selection

In the present subsection, we focus on the case of a small proportion of outliers that could be as high as 20% if we take into account the theoretical properties of ICS for the $\text{COV} - \text{COV}_4$ pair as detailed in 2.8.2. We assume that the outliers belong to a subspace of dimension $q \leq p$, and we aim at providing some procedures to automatically select a number of invariant components close to $q$. Beginning with the first component, we test whether each invariant component is significantly relevant via two different sequential approaches. For both approaches, as soon as one invariant component, - let us say number $(k+1)$, - is not significantly relevant, we stop the procedure and select the $k$ first components. In this particular context of sequential multiple testing, some adjustments on the initial significance level $\alpha$ are necessary. Following Dray (2008), we apply the Bonferroni correction on the significance level and consider a level $\alpha_j = \alpha/j$ for each component $j = 1, \ldots, p$.

The first approach consists in a Parallel Analysis (PA) based on Monte Carlo simulations. For some given dimensions $n$ and $p$, many samples are generated following a standard multivariate Gaussian distribution, and for each sample and a given scatter pair, the eigenvalues of the simultaneous diagonalization of the two scatter matrices are computed. Cut-offs for the eigenvalues are then derived using the empirical quantiles of the eigenvalues from the simulated Gaussian data. This method is common for selecting components in PCA as described in Peres-Neto et al. (2005). It was already used in Caussinus et al. (2003b) for ICS but only for a particular pair of scatters. The second approach makes use of the fact that relevant components for displaying outliers do not follow a Gaussian distribution. It is thus based on univariate normality tests for each component beginning with the first one as previously described. The five tests we compare are the D'Agostino test of skewness (DA), the Anscombe-Glynn (AG) test of kurtosis, the Bonett-Seier (BS) test of Geary's kurtosis, the Jarque-Bera (JB) test based on both skewness and kurtosis and the Shapiro-Wilk (SW) normality test (see Yazici and Yolacan (2007) and Bonett and Seier (2002) for a complete description of these five tests).

Note that automated selection procedures are necessary in a simulation framework but may not be the best alternative when analyzing one data set. This point will be detailed further in the data analysis section of the present paper, where we also explore the possibility of using a scree plot as in PCA.

### 2.4.3 Outlier identification

Once having selected $k$ invariant components, the last procedure step is the identification of outlying observations. For each observation $i = 1, \ldots, n$, we calculate its squared "ICS distance" which corresponds to its squared Euclidian norm in the invariant coordinate system taking into account the first $k$ coordinates:

$$(\text{ICS distance})_i^2 = \sum_{j=1}^{k} \left(z_i^j\right)^2$$

where $z_i^j$ denotes the $j$th coordinate of the score $\mathbf{z}_i$. As the distribution of the ICS distances is unknown, we derive cut-offs based on Monte Carlo simulations from the standard Gaussian distribution. For a given data dimension, a scatter pair and a number $k$ of selected components, we generate many samples and compute the ICS distances. A cut-off is derived for a fixed level $\gamma$ as the $1-\gamma$ percentile of these distances. An observation with a distance higher than this cut-off is flagged as an outlier.

The implementation of ICS for outlier detection in the next two sections is performed in R 3.1.2 (R Core Team, 2017) using the packages ICS (Nordhausen et al., 2008), ICSOutlier (Archimbaud et al., 2016), mvtnorm (Genz and Bretz, 2009), moments (Komsta and Novomestky, 2015), robustX (Stahel et al., 2013) and robustbase (Rousseeuw et al., 2016).

## 2.5 Simulations

### 2.5.1 Simulation framework

ICS performance for outlier detection is evaluated through an extensive simulation study in the particular context of a proportion of outliers fixed at 2%. As already indicated, this small proportion is consistent with some current practice in industrial applications where the data already meet the standard quality controls and only a few observations, clearly identified as multivariate outliers, may be disregarded. The different models we consider are well-known models in the robust statistics literature (Hampel et al., 1986). Using the $\text{COV} - \text{COV}_4$ scatter pair and for the two components mixture models or the scale-shift model, it is possible to derive some theoretical conditions on the contamination level which insure that the first invariant components point in the directions of the outliers (see 2.8.2 for details).

In this framework, we discuss the impact of the scatter pair together with the components selection strategy and the choice of the cut-off for identifying outliers. Some of the conclusions and recommendations drawn from this study are used as guidelines for the data analysis conducted in Section 2.6 in different industrial settings. Concerning the scatter matrices, the four pairs (i) $\text{COV} - \text{COV}_4$, (ii) $\text{MLC} - \text{COV}$, (iii) $\text{MCD} - \text{COV}$ and (iv) $\text{MCD} - \text{MLC}$ are evaluated. In pairs (ii)-(iv) the scatter matrices come from different classes, while in pair (i) both come from class I. For the MCD, given that the proportion

of outliers is small, the value $h = 0.75$ which is often advocated (Croux and Haesbroeck, 1999) is used throughout the simulations, leading to a 25% breakdown point.

For each of the six setups, we generate 1000 samples with sample size $n = 1000$ and dimension $p$ equal to 6, 25 and 50. For all cases, the uncontaminated data follow a Gaussian distribution with mean 0 and covariance matrix $\Sigma_i$, $i = 0, \ldots, 5$, depending on the setup. Except for Case 0 which contains no outlier, we generate exactly 20 outliers in each sample so that the proportion of outliers is 2% in all samples. We use the notation $\boldsymbol{e}_i$ for the $p$-vector with a one in the $i$th coordinate and zero elsewhere. For each setup we give the dimension $q$ of the subspace spanned by the outliers. For dimension $p = 6$, Figure 2.6 gives the scatterplot matrix of the variables for one simulation of each of the six cases in order to visualize easily the structure of the data sets. Some affine transformation could have been performed in order to mask the structure like in Stahel and Mächler (2009) for the barrow wheel distribution. Such transformation has no impact on the MD and ICS results but may change completely the PCA results.

**Case 0** ($q = 0$)**:** $\boldsymbol{\Sigma}_0 = \mathbf{I}_p$ with no outlier.

**Case 1** ($q = 1$)**:** $\boldsymbol{\Sigma}_1 = \text{diag}(1, 4, \ldots, 4)$ with outliers clustered in one direction with distribution $\mathcal{N}(6\boldsymbol{e}_1, \boldsymbol{\Sigma}_1)$.

**Case 2** ($q = 1$)**:** $\boldsymbol{\Sigma}_2 = \text{diag}(0.1, 1, \ldots, 1)$ with outliers following a distribution $H$ such that $\mathbf{h} = (h_1, \mathbf{h}_2')' \sim H$ means that $h_1 \sim \chi_5$ and $\mathbf{h}_2 \sim \mathcal{N}(\mathbf{0}, 0.2\mathbf{I}_{p-1})$. The data follows the so-called barrow wheel distribution as introduced in Hampel et al. (1986) and using a slightly modified setting compared to Stahel and Mächler (2009). No rescaling or rotation has been performed. In any case such transformations have no impact on the ICS results. Hence, outliers are generated along the same direction on both sides of the uncontaminated data cloud.

**Case 3** ($q = 2$)**:** $\boldsymbol{\Sigma}_3 = \text{diag}(1, 1, 4, \ldots, 4)$ with outliers clustered in two directions with 12 (resp. 8) observations following a $\mathcal{N}(6\boldsymbol{e}_1, \boldsymbol{\Sigma}_3)$ (resp. $\mathcal{N}(6.2\boldsymbol{e}_2, \boldsymbol{\Sigma}_3)$) distribution.

**Case 4** ($q = 6$)**:** $\boldsymbol{\Sigma}_4 = \mathbf{I}_p$ with outliers clustered in six directions with Gaussian distribution with mean $\boldsymbol{\mu}_i = (6 + 0.1(i-1))\boldsymbol{e}_i$, $i = 1, \ldots, 6$ and covariance $\mathbf{I}_p$, with 4 (resp. 3) outliers in the first two (resp. last four) clusters.

**Case 5** ($q \leq 6$)**:** $\boldsymbol{\Sigma}_5 = \mathbf{I}_p$ with outliers generated in up to six directions via scale shifts with a covariance matrix $\tilde{\boldsymbol{\Sigma}}_5 = \text{diag}(5, \ldots, 5)$ if $p \leq 6$ and $\text{diag}(5, 5, 5, 5, 5, 5, 1, \ldots, 1)$ if $p > 6$. The 20 outliers are generated by drawing observations from a $N(\mathbf{0}, \tilde{\boldsymbol{\Sigma}}_5)$ distribution and keeping the ones with at least one variable (among the first six) larger than the maximum value or smaller than the minimum value of the non-outlying observations.

Details concerning the implementation of the simulations can be found in the supplementary material. To compare the performance of the methods, we provide the percentage of outliers correctly identified (denoted by TP for "True Positive") and the percentage of non-outlying observations erroneously identified as outliers (FN for "False Negative").

### 2.5.2 Selecting the invariant components

Before examining the performance of ICS in terms of TP and FP, we observe the selected dimensions using the D'Agostino (DA) and the Parallel Analysis PA methods for a level $\alpha = 5\%$. Table 2.1 below gives the average of these dimensions over the 1000 simulations for the different cases. Note that the results for the other normality tests proposed in Subsection 2.4.2 have not been reported because they do not improve the performance compared with the DA and PA methods.

| Scatters | $p$ | Case 0 ($q=0$) | | Case 1 ($q=1$) | | Case 2 ($q=1$) | | Case 3 ($q=2$) | | Case 4 ($q=6$) | | Case 5 ($q\leq 6$) | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | DA | PA | DA | PA | DA | PA | DA | PA | DA | PA | DA | PA |
| COV - $COV_4$ | 6 | 0.14 | 0.08 | 1.06 | 1.58 | 1.00 | 1.00 | 1.96 | 2.90 | 2.67 | 6.00 | 1.34 | 5.96 |
| | 25 | 0.42 | 0.09 | 1.25 | 1.98 | 1.27 | 1.09 | 2.09 | 4.33 | 2.95 | 10.48 | 1.62 | 8.41 |
| | 50 | 0.80 | 0.06 | 1.59 | 1.82 | 1.53 | 2.02 | 2.37 | 4.35 | 2.93 | 11.48 | 1.99 | 7.41 |
| MLC - COV | 6 | 0.12 | 0.08 | 1.05 | 1.45 | 0.99 | 1.08 | 1.98 | 2.77 | 2.13 | 5.97 | 1.09 | 5.36 |
| | 25 | 0.23 | 0.08 | 1.15 | 1.75 | 1.10 | 1.04 | 2.03 | 3.59 | 2.08 | 9.25 | 1.18 | 6.27 |
| | 50 | 0.46 | 0.06 | 1.34 | 1.76 | 0.48 | 20.31 | 2.16 | 3.87 | 2.10 | 8.86 | 1.32 | 5.23 |
| MCD - COV | 6 | 0.15 | 0.05 | 1.06 | 1.05 | 1.00 | 1.00 | 2.01 | 2.21 | 2.06 | 6.00 | 1.08 | 5.62 |
| | 25 | 0.38 | 0.07 | 1.28 | 1.29 | 1.21 | 1.02 | 2.15 | 2.84 | 2.08 | 9.24 | 1.13 | 6.56 |
| | 50 | 0.65 | 0.05 | 1.46 | 1.51 | 1.43 | 1.06 | 2.33 | 3.33 | 1.79 | 6.94 | 1.13 | 3.45 |
| MCD - MLC | 6 | 0.08 | 0.07 | 1.04 | 0.99 | 1.00 | 1.00 | 2.05 | 0.52 | 1.75 | 0.03 | 0.66 | 0.05 |
| | 25 | 0.25 | 0.05 | 1.17 | 1.14 | 1.11 | 1.04 | 2.09 | 2.40 | 1.28 | 1.96 | 0.68 | 1.54 |
| | 50 | 0.56 | 0.05 | 1.42 | 1.49 | 1.40 | 1.00 | 2.27 | 2.83 | 1.28 | 1.96 | 0.78 | 1.13 |

Table 2.1 – Averaged numbers of selected invariant components for the DA and PA methods.

Under setups 0, 1, 2 and 3, the results from Table 2.1 are overall quite good and comparable for the different scatter pairs. Only certain specific results have to be pointed out for the pairs MLC − COV (Case 2 with $p = 50$ for DA and PA) and MCD − MLC (Case 3, $p = 6$ for PA), and these points require further investigation. Moreover, for the four setups, the differences between procedures DA and PA are small, with some overestimation of the dimension for PA in Case 3 when $p = 25$ or 50.

The results are not as good for Cases 4 and 5 which correspond to larger $q$ values than the other setups, in particular for the DA procedure that leads to an important underestimation of the dimension for all scatter pairs. The PA procedure gives better results in this context except for the MCD − MLC pair, which leads to an important underestimation in all cases. The results for COV − COV4 and MCD − COV are quite similar despite a larger overestimation of the dimension for COV − COV4 in Cases 4 and 5, for $p = 25$ and $p = 50$, when using the PA procedure.

These first results are in favor of the pairs COV − COV4 and MCD − COV but need to be confirmed by studying the performance of the methods in terms of TP and FP.

### 2.5.3 Detecting outliers with ICS

Table 2.2 gives the TP (except for Case 0) and FP averaged over the 1000 simulations and averaged also over Cases 1 to 5 to save space. The $\gamma$ level for the identification cut-off is fixed at 2%.

| Averaged Measures in % | TP | | | FP | | | FP Case 0 | | |
|---|---|---|---|---|---|---|---|---|---|
| $p$ | 6 | 25 | 50 | 6 | 25 | 50 | 6 | 25 | 50 |
| True subspace | 95.10 | 96.68 | 93.92 | 0.10 | 0.07 | 0.12 | | | |
| ICS true $q$ COV - $COV_4$ | 96.92 | 92.52 | 80.13 | 0.57 | 0.33 | 0.49 | | | |
| ICS true $q$ MLC - COV | 97.53 | 92.24 | 64.97 | 1.27 | 0.63 | 1.01 | | | |
| ICS true $q$ MCD - COV | 97.53 | 93.59 | 81.99 | 1.48 | 0.92 | 0.83 | | | |
| ICS true $q$ MCD - MLC | 97.29 | 92.75 | 80.09 | 1.82 | 1.49 | 1.07 | | | |
| ICS DA COV - $COV_4$ | 77.01 | 78.14 | 70.26 | 0.43 | 0.38 | 0.57 | 0.30 | 0.70 | 1.17 |
| ICS DA MLC - COV | 76.34 | 75.84 | 54.18 | 1.03 | 0.56 | 0.65 | 0.25 | 0.42 | 0.76 |
| ICS DA MCD - COV | 76.77 | 77.61 | 69.31 | 1.22 | 0.94 | 0.86 | 0.30 | 0.68 | 0.95 |
| ICS DA MCD - MLC | 71.51 | 71.41 | 65.18 | 1.53 | 1.23 | 0.95 | 0.18 | 0.48 | 0.87 |
| ICS PA COV - $COV_4$ | 96.73 | 91.39 | 76.29 | 0.70 | 0.64 | 0.79 | 0.10 | 0.11 | 0.09 |
| ICS PA MLC - COV | 96.89 | 91.68 | 62.77 | 1.39 | 0.85 | 1.10 | 0.15 | 0.12 | 0.09 |
| ICS PA MCD - COV | 97.36 | 93.36 | 76.66 | 1.50 | 1.03 | 0.89 | 0.11 | 0.14 | 0.09 |
| ICS PA MCD - MLC | 44.95 | 76.92 | 65.44 | 0.95 | 1.32 | 0.90 | 0.08 | 0.11 | 0.09 |

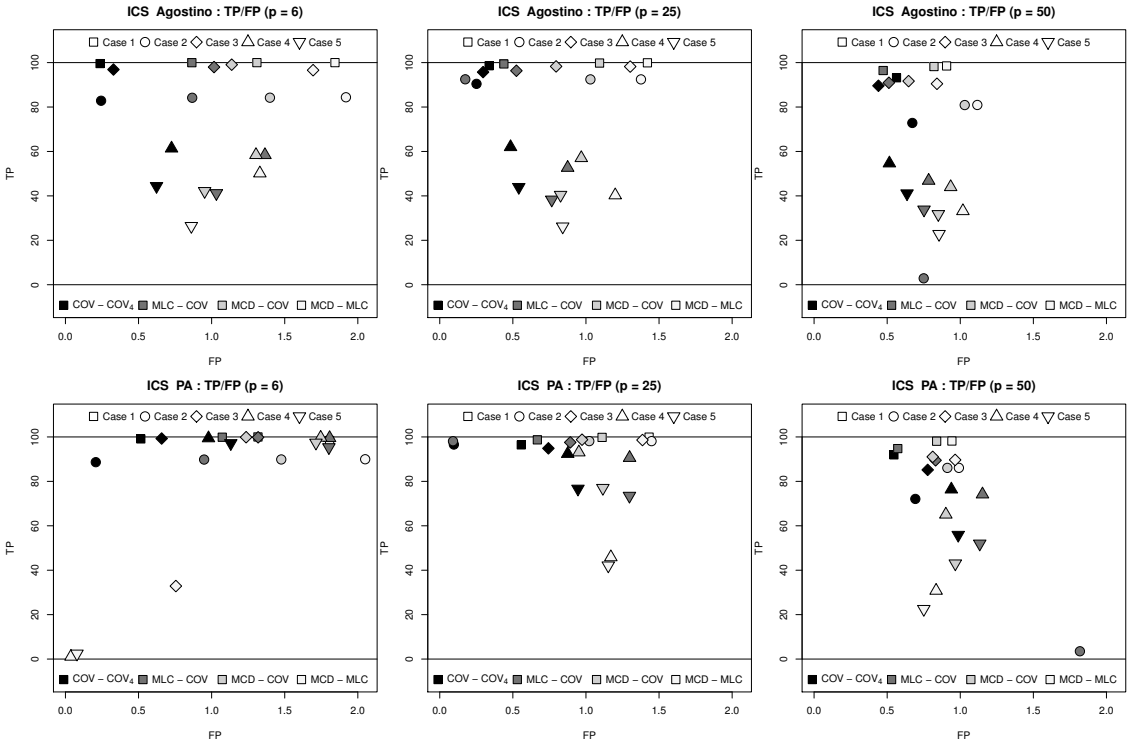
Table 2.2 – TP and FP results for ICS (averaged results for Cases 1 to 5).

The first row of Table 2.2 gives some kind of oracle performance measure obtained by calculating TP and FP values using the Euclidian norm of the projected data on the true subspace containing the outliers (known for each Case). As anticipated, the results are very good regardless of the dimension. The next results are obtained when the true number of invariant components is selected but the invariant components are estimated using different scatter pairs. They give another oracle performance measure. Compared with the first row of Table 2.2, these results are globally good in terms of TP but perform less well in terms of FP. They give an idea of the impact of the scatter matrices estimation when the number of invariant components is the true one. In this context, the COV - COV4 scatter pair clearly outperforms the others with similar TP values but smaller FP values.

Then, the results are given for the two automated selection DA and PA. Compared with the previous results, they give some insight into the impact of the dimension selection procedures. When looking at Cases 1 to 5, there is no method for dimension selection that outperforms the other. PA is the best in terms of TP, but D'Agostino is the best in terms of FP. However, for Case 0, any dimension $p$ and any scatter pair, the PA selection leads to less than 0.15% FP on average, while the values are larger for DA. The choice $\mathrm{COV}-\mathrm{COV}_4$ is the best in most situations from the FP point of view.

Figure 2.2 gives more details concerning the TP and the FP values for the Cases 1 to 5. It contains scatter plots of the TP against the FP for D'Agostino (top) and PA (bottom) and for the different values of $p$. Note that Case 2 for $p = 50$ and $\mathrm{MLC} - \mathrm{COV}$ is very specific with often no component selected and will not be considered further in our comments. For DA, the results are clearly ordered in terms of TP according to the different Cases, from the largest TP values for Case 1 to the smallest ones for Case 5. There are only tiny differences between the scatter pairs. With respect to FP values, the

Figure 2.2 – Averaged TP and FP results for ICS detailed for Cases 1 to 5.




results are now ordered according to the different scatter pairs from the smallest values for $\mathrm{COV} - \mathrm{COV}_4$ to the largest values for MCD − MLC. These differences are more limited for Cases 4 and 5 than for Cases 1 to 3 and decrease for all cases when $p$ increases.

For PA, the results differ. If we except MCD−MLC, all scatter pairs lead to very similar and good TP values when $p = 6$ while $\mathrm{COV} - \mathrm{COV}_4$ is clearly the best when comparing FP values. For the particular pair MCD − MLC and Cases 4 and 5, as observed from Table 2.1, no dimension is selected, and so no outlier can be detected. When $p$ increases, in general the results become worse for TP, in particular for Cases 4 and 5, while they become close together for FP.

From this simulation results, we recommend using the pair COV − COV4. For this scatter pair, the results for DA and PA, - compared to the ones obtained when the true dimension $q$ is known, - do not make it possible to conclude in favor of one of the two selection methods. While the TP values are better and closer to the oracle for PA, the FP values are better and closer to the oracle for DA.

### 2.5.4 Comparing ICS with the Mahalanobis distance and PCA

Table 2.3 recalls the TP and the FP values for ICS focusing on COV − COV4 but also gives the values when using non-robust (MD) and robust (RD) Mahalanobis distances and PCA (unstandardized and standardized). RD is obtained using a 25% breakdown point reweighted MCD estimator. For the Mahalanobis distances, we only report the

results when the cut-off values are the usual ones, based on a chi-squared distribution quantile (of order 2%) or are adjusted to take into account some asymptotic corrections for RD and the method is denoted GM (see CerioliOutlierDetection (Green and Martin, 2017a) and Green and Martin (2017b) for the implementation). Other criteria obtained through simulations have been implemented but do not bring any improvement and are not reported. Concerning PCA and robust PCA, the outlier detection procedure is quite complex since the method is not aimed at detecting outliers. Atypical observations may thus be displayed on any of the $p$ principal components (Jolliffe, 2002). Basically, the procedure consists in selecting some components and calculating, on the one hand, a distance in the space spanned by the selected components (after some standardization), and, on the other hand, a distance in the space orthogonal to the previous space (see Hubert et al. (2005) for details). In our comparison, following Hubert et al. (2005), observations associated with at least one large distance are flagged as outliers using some cut-off values based on quantiles of order 99% for each distance. We tried different methods for principal components selection but report only the results obtained when the dimension is chosen as the best possible among all possible dimensions (from 1 to $p$). More precisely, it means that the results give the smallest FP value among all the results that were found to maximize TP value. Automated methods were also tested but the results were never better than the ones reported. Some robust PCA methods where the usual covariance or correlation matrix is replaced by some robust estimators were also implemented but did not lead to better results and are not reported neither. Results are averaged for Cases 1 to 5 in order to save space.

| Averaged Measures | | TP | | | FP | | | FP Case 0 | |
|---|---|---|---|---|---|---|---|---|---|
| $p$ | 6 | 25 | 50 | 6 | 25 | 50 | 6 | 25 | 50 |
| ICS DA COV - $COV_4$ | 77.01 | 78.14 | 70.26 | 0.43 | 0.38 | 0.57 | 0.30 | 0.70 | 1.17 |
| ICS PA COV - $COV_4$ | 96.73 | 91.39 | 76.29 | 0.70 | 0.64 | 0.79 | 0.10 | 0.11 | 0.09 |
| MD | 94.14 | 72.80 | 52.03 | 1.24 | 1.91 | 2.40 | 2.09 | 2.35 | 2.69 |
| RD GM | 94.03 | 78.80 | 53.57 | 0.20 | 0.26 | 0.26 | 0.21 | 0.29 | 0.23 |
| RD | 97.37 | 91.34 | 75.26 | 1.78 | 1.88 | 1.94 | 2.09 | 2.12 | 2.10 |
| PCA | 98.55 | 91.58 | 84.43 | 1.14 | 1.22 | 1.17 | 2.01 | 1.91 | 1.84 |
| PCA std | 80.98 | 80.88 | 47.85 | 0.58 | 0.84 | 1.52 | 1.99 | 1.80 | 1.54 |

Table 2.3 – Comparison of ICS with MD, RD and PCA (averaged results in % for Cases 1 to 5).

The performance of MD, RD and PCA compared to the other methods is particularly low when focusing on the FP measure. For standardized PCA, results are better when the dimension is equal to 6 but the method cannot compete when the dimension increases. ICS with DA and PA together with RD when using the GM correction lead to better performance. When $p = 6$, RD GM gives the best results with very low FP on average for Cases 1 to 5. When $p = 50$, the method still leads to low false detection but at the cost of a low true positive detection compared to ICS. For Case 0, RD GM exhibits good results but ICS PA outperforms it. In conclusion, we advocate the use of ICS with the scatter pair $\mathrm{COV} - \mathrm{COV4}$, which is very easy to compute and exhibits good performance. In this framework where the majority of the data follows a Gaussian distribution, we recommend

the PA components selection method, but the DA method is an interesting alternative with a very low computational cost.

Note that in case the majority of the data does not follow a Gaussian distribution, the different cutoffs are not valid anymore and should be adapted.

## 2.6 Data Analysis

We analyze three real data sets using ICS and compare ICS with several competitors that are Mahalanobis distance or PCA variants, including ROBPCA as introduced by Hubert et al. (2005) and implemented in the package rrcov (Todorov and Filzmoser, 2009). All data are from industrial processes and contain potentially a small proportion of outliers. The last two data sets in particular come from industrial processes where there is a high level of quality control and only a small proportion of observations can be diagnosed as outliers. It implies that the False Positive rate is crucial and should be as small as possible.

For each of the three data sets, we give details concerning the observations considered as true outliers in the following subsections. Table 2.4 provides the number of True Positive (NTP) and the number of False Positive (NFP) for the three data sets. Be careful that for this particular table, the values are not given as proportions. For ICS, we only report results for the scatter pair $\text{COV} - \text{COV}_4$ because, in most cases, this pair leads to the best results, which is consistent with our simulation conclusions and confirms our recommendations. For ICS and PCA methods, the results depend on the number of selected components, and we show results for three different types of selection. The "best selection" results are obtained by trying all possible dimensions between 1 and $p$ and taking, for each method, the dimension $k$, which leads to the smallest NFP among those that maximize the NTP. This procedure leads to some kind of oracle measure of the maximum performance of the methods. The second type of results are obtained through automated components selection methods as detailed in the previous section for ICS and using the rule proposed in the package rrcov for ROBPCA. Moreover, for ICS, only the DA and PA automated components selection methods are reported, because they give the best results in general. As can be observed from the last two data sets, and also from our experience on other data sets, the automated procedures for ICS tend to select too many components. One possible reason is that these procedures rely on the Gaussian distribution of the main bulk of the data, and such an assumption may not be fulfilled in practice. Therefore, we propose to use the scree plot as an alternative visual selection method that leads to a third type of results for ICS. The scree plot is very well-known for PCA (Jolliffe, 2002) and can be applied in the same way for ICS except that for the scatter pair COV − COV4, the eigenvalues are to be interpreted in terms of kurtosis (see Tyler et al. (2009)), instead of variance for PCA. The scree plots for the three examples are given on Figure 2.3. For the three scree plots, some invariant components (two for

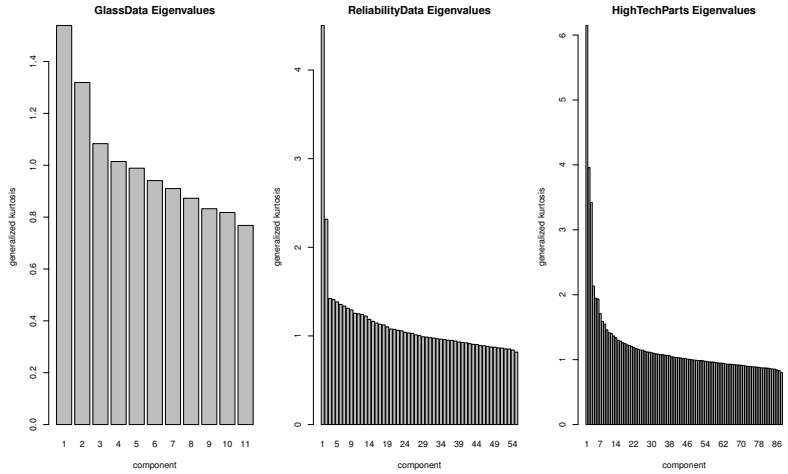



Figure 2.3 – Scree plots for ICS with COV − COV4 for the three data sets.

the Glass and the Reliability data sets and three for the HTP data) clearly differ from the other components due to their high eigenvalues. The results for these components selection are reported in the last row of Table 2.4.

| | Glass | | | Reliability | | | HighTech | | |
|---|---|---|---|---|---|---|---|---|---|
| | NTP (/3) | NFP (/109) | $k$ (/11) | NTP (/2) | NFP (/518) | $k$ (/55) | NTP (/2) | NFP (/900) | $k$ (/88) |
| MD | 3 | 4 | | 2 | 52 | | 2 | 119 | |
| RD | 3 | 15 | | | | | 2 | 243 | |
| RD GM | 3 | 7 | | | | | 2 | 223 | |
| *Best selection* | | | | | | | | | |
| ICS COV − $COV_4$ | 3 | 3 | 2 | 2 | 1 | 1 | 2 | 0 | 1 |
| PCA | 3 | 9 | 5 | 2 | 41 | 52 | 2 | 21 | 1 |
| PCA std | 3 | 4 | 2 | 2 | 22 | 40 | 2 | 25 | 6 |
| ROBPCA | 3 | 13 | 5 | | | | 2 | 50 | 1 |
| *Automated selection* | | | | | | | | | |
| ICS COV − $COV_4$ DA | 3 | 3 | 2 | 2 | 23 | 12 | 2 | 39 | 14 |
| ICS COV − $COV_4$ PA | 3 | 3 | 2 | 2 | 42 | 28 | 2 | 87 | 50 |
| PCA | 1 | 5 | 1 | 0 | 6 | 12 | 2 | 24 | 3 |
| PCA std | 1 | 4 | 1 | 2 | 31 | 20 | 2 | 28 | 4 |
| ROBPCA | 3 | 17 | 1 | | | | 2 | 80 | 2 |
| *Scree plot selection* | | | | | | | | | |
| ICS COV − $COV_4$ | 3 | 3 | 2 | 2 | 1 | 2 | 2 | 5 | 3 |

Table 2.4 – NTP, NFP and number $k$ of selected components for the three real data examples.

### 2.6.1 Glass recycling

The so-called glass data set is analyzed by Cerioli and Farcomeni (2011) and consists of 112 glass fragments collected for recycling, of which 109 are true glass fragments and 3 are contaminated ceramic glass fragments. The 11 variables are the log of spectral measures recorded for each fragment. For all methods, the outliers are flagged by using cut-offs defined through simulations at the 5% level so that results are comparable with Cerioli and Farcomeni (2011). For this example, ICS detects the three outliers and has only three false detections, and the results are the same for the three types of components selection

(best, automated or scree plot). ICS has the highest performance in comparison with the competitors considered here but also in comparison with the results reported in Table 6 of Cerioli and Farcomeni (2011). The non robust Mahalanobis distance, which is equivalent to ICS with $\text{COV} - \text{COV}_4$ when all components are selected, also performs quite well on this example. All three outliers are detected, and there are only four false detections compared to three when two invariant components are selected among the eleven.

For the next two examples, to obtain an acceptable quality control performance, true outliers should be detected with up to 2% of observations flagged as outliers, taking into account the true outliers and the false detections. Moreover, the results for these two examples are readily reproducible using the R code provided in the supplementary material of the present paper.

### 2.6.2 Reliability Data

The Reliability data are available in the R package REPPlab (Fischer et al., 2016b) and contain 55 variables measured on 520 units during a production process. The quality standards for this process are respected for each variable, and the objective is to detect some potential multivariate faulty units representing less than 2% of the 520 observations. In Fischer et al. (2016a), two observations (414 and 512) are detected as the most severe outliers. For simplicity, we consider these two observations as the only true outliers. However, there may be other outliers, and the NFP numbers should be viewed with caution for this example in comparison with the other two data sets, where some auxiliary information concerning the true outliers is known. For this example and the next one, the outliers are flagged by using cut-offs defined through simulations at the 2% level.

In Table 2.4, the results for the MCD are not reported. As mentioned in Fischer et al. (2016a), computing the MCD (at least with a breakdown point equal or larger than 25%) is not possible on this data set because 497 observations among the 520 take exactly the same value on the 24th variable. Note that from our experience, this problem occurs quite recurrently on real data sets in some industrial context, and, as illustrated below, removing such variables may lead to a loss of relevant information. This is, however, not a problem for ICS when using the scatter pair $\text{COV} - \text{COV}_4$, and the method shows very good performance for the Reliability data when selecting only two components. The only observation declared as a false positive is observation 57, which is also flagged as an outlier in Fischer et al. (2016a) (although not as extreme as the other two). The selection of two components is suggested by the scree plot analysis. The automated selection procedures or the use of all invariant components (Mahalanobis distance) show poor performance with a number of false positives higher than the 2% rate that is acceptable. PCA is even less successful with many false positive in the best selection case and sometimes no detection at all when the components selection is automated.

Moreover, when the number of selected invariant components is small, ICS makes the detected outliers easy to interpret by drawing scatter plots of the selected components

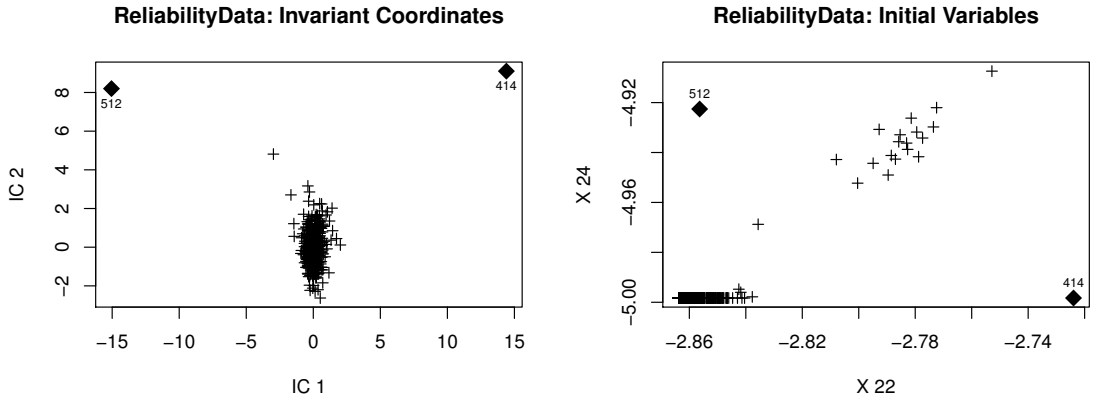



Figure 2.4 – Scatter plot of the first two invariant components (left panel) and scatter plot of the variables numbered 22 and 24 (right panel) for the Reliability data set.

and by observing the correlations between the components and the original variables. Figure 2.4 illustrates this point. The two selected invariant components are plotted on the left panel and clearly lead to the identification of observations 414 and 512 as outliers. When calculating the correlations between these invariant components and the 55 original variables, it appears that they are essentially correlated with variables 22 and 24. These two variables are thus plotted on the right panel of Figure 2.4 and reveal that observation 414 (resp. 512) combines in an unusual way a high (resp. small) value on variable 22 with a small (resp. large) value on variable 24. Note that removing variable 24 in order to compute the MCD estimate precludes the ability to detect the two outliers.

### 2.6.3 High-tech parts

The third real data set contains 902 high-tech parts designed for consumer products and characterized by 88 electronic measures; it is available in the R package ICSOutlier (Archimbaud et al., 2016). To anonymize the data collected, the measures have been mean-centered. We do not have access to the original data, but we know that they were cleaned from univariate outliers using some preliminary standard quality control rules. No multivariate outlier detection method was applied and the parts were sold. However, two parts (denoted by R1 and R2 in what follows) among the 902 were found to be defective and returned to the manufacturer. Our objective is to check whether these two observations could have been detected before being sold, using some multivariate outlier detection method in an unsupervised way, with less than 2% of observations flagged as outliers.

From Table 2.4, the result based on only one component (best selection) for ICS is perfect, with two outliers detected and no false detection. The results are much worse for all other methods, with too many false detections. This is especially true when considering the Mahalanobis distance with no selection of components. The results for ICS are rather mediocre when using the DA or (even worse) the PA automated selection methods which

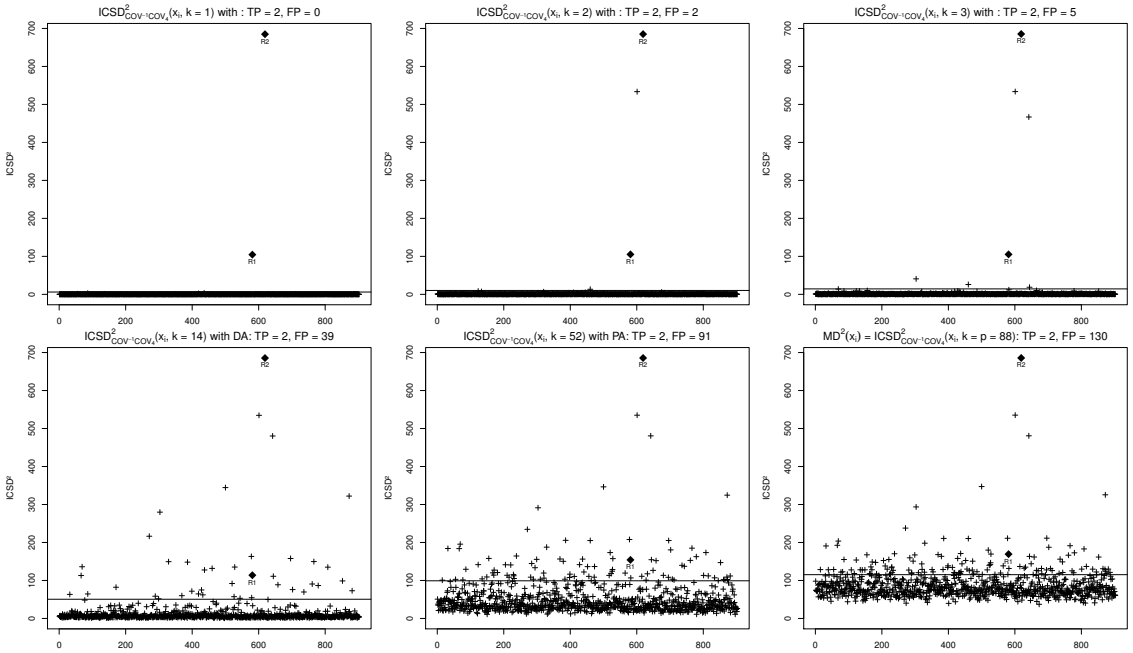



Figure 2.5 – Plots of the squared ICS distances for different numbers of invariant components selected for the HTP data set.

tend to select too many components. Using the scree plot, however, leads unambiguously to a more drastic selection, with three eigenvalues larger than the others. Using three components leads to good performance, with five NFP and all together seven detected outliers, which is less than 2% and thus acceptable. Figure 2.5 gives more insight on the influence of the number of selected invariant components on the detection performance and echoes Figure 2.1. The six scatter plots give the squared ICS distances when the number of components increases. The top-left plot corresponds to one component, which is the best possible selection. Then, the NFP increases when more components are selected. The bottom-left plot corresponds to DA selection, while the bottom-middle plots correspond to PA selection. On the bottom-right plot, all 88 components are taken into account, which corresponds to the squared Mahalanobis distance, and the result is the worst. Note that for this data set, PCA performs better than the Mahalanobis distance even if the number of NFP is still unacceptable.

Finally, ICS is shown to be appropriate for the three data sets when using the scree plot selection method, while the performance of its competitors depends on the data set.

## 2.7 Conclusion and perspectives

The remarkable theoretical properties of ICS are confirmed in the context of multivariate outlier detection with a small proportion of outliers. In particular, the ability of ICS to recover the Fisher's linear discriminant subspace in the case where group identifications are unknown has been verified on simulations, with a majority of the data following a Gaussian distribution, but also on some real data set where the Gaussian assumption is

not true anymore. So, as stated for Linear Discriminant Analysis by Hastie et al. (2001), it seems that the applicability of ICS extends beyond the realm of Gaussian data. However, this remark does not apply to the components selection procedures we propose. From our simulation study, we advocate the use of some selection methods such as DA or PA. But such methods are not convincing when analyzing real data sets as they tend to select too many components. The reason is certainly the fact that the majority of the data does not follow a Gaussian distribution while the cut-offs we propose depend heavily on this assumption. The data analysis of real data sets highlights the advantage of using the scree plot for selecting the number of components.

Contrary to PCA, the method is not only orthogonal invariant but also scale invariant and is aimed at detecting outliers. More precisely, the present paper demonstrates the good performance of ICS, when using the scatter pair $\text{COV} - \text{COV}_4$ and selecting the first components in a context of a small proportion of outliers. The simulation study together with the data analysis illustrates that ICS consistently detects outliers, when they are present, with a small proportion of false detections, while the success of its competitors depends more on the data set under study. This is particularly true for PCA whose results depend a lot on the way the data are scaled. If the outliers are not concentrated in a small dimension subspace, ICS is equivalent to the Mahalanobis distance. For large dimensions and when outliers are contained in a small dimensional subspace, using ICS may improve greatly with respect to the Mahalanobis distance as illustrated by certain theoretical properties and applications. Moreover, selecting a small number of invariant components makes outlier interpretation much easier.

A perspective of the work is to consider multiple testing procedures for the choice of the cut-off for the distances as proposed by Cerioli (2010) and Cerioli and Farcomeni (2011). Moreover, instead of defining cut-offs independently for the components selection and the outlier detection steps, it would be of interest to propose some alternative which would control the overall false positive rate of the global procedure. Another perspective is to consider the case of a large proportion of outliers. In such a context, the scatter pair choice has to be revisited together with the components choice. If outliers are contained in a small dimensional subspace, the $\text{COV} - \text{COV}_4$ pair, even if it is not robust, may still be a good alternative given the ICS theoretical properties. However, small kurtosis values are now also of interest, and thus invariant components associated with small eigenvalues should be examined. In such a context, the recent papers Nordhausen et al. (2016) and Nordhausen et al. (2017) are of particular interest. The problem of high dimension and small sample size is also relevant in our industrial context for some particular applications. The adaptation of ICS to such data sets is a work in progress.

## Acknowledgements

The work of Klaus Nordhausen was partly supported by the Academy of Finland (grant 268703). The article is based upon work from CRoNoS COST Action IC1408, supported by COST (European Cooperation in Science and Technology).

## Supplementary material

The four supplemental files are contained in a single archive.

**readme.txt:** brief description of the contents of the supplemental files.

**Scatterplot_simulations.pdf:** figure with six scatterplot matrices to visualize the distribution of one data set from each of the Cases 0 to 5 and with $p = 6$.

**CodeR_simulations_functions.r:** R code to generate the simulated data (Cases 1 to 5).

**CodeR_examples.r:** R code to derive the results of Table 4 for the Reliability data and the HTP data sets.

## 2.8 Appendix

### 2.8.1 Proof of Proposition 1

Let us denote by $\mathbf{M}$ the $p \times q$ matrix whose columns contain the vectors $\boldsymbol{\mu}_h$, $h = 1, \ldots, q$. Given the affine invariance property of ICS, we assume w.l.o.g. that $\boldsymbol{\mu}_0 = \mathbf{0}$, that $\boldsymbol{\Sigma}_W = \mathbf{I}_p$ where $\mathbf{I}_p$ denotes the $p \times p$ identity matrix and that the last $p - q$ rows of $\mathbf{M}$ contain zeros so that: $\mathbf{M} = [\mathbf{M}_q, \mathbf{0}]'$ where $\mathbf{M}_q$ is a $q \times q$ matrix. In the following, we also assume for convenience that the dimension of the vector space spanned by the columns of $\mathbf{M}$ is $q$. Otherwise, we would have to reparametrize the mixture distribution with a number of clusters smaller than $q + 1$ and equal to one plus the dimension of the subspace spanned by the columns of $\mathbf{M}$. Under these assumptions, we determine that the total covariance matrix can be written as

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_q & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{p-q} \end{bmatrix}$$

where $\boldsymbol{\Sigma}_q$ denotes a non-singular $q \times q$ matrix. We also denote by $\mathbf{X}_q$ (resp. $\boldsymbol{\mu}_{\mathbf{X}_q}$) the first $q$ rows of $\mathbf{X}$ (resp. of $\boldsymbol{\mu}_{\mathbf{X}}$).

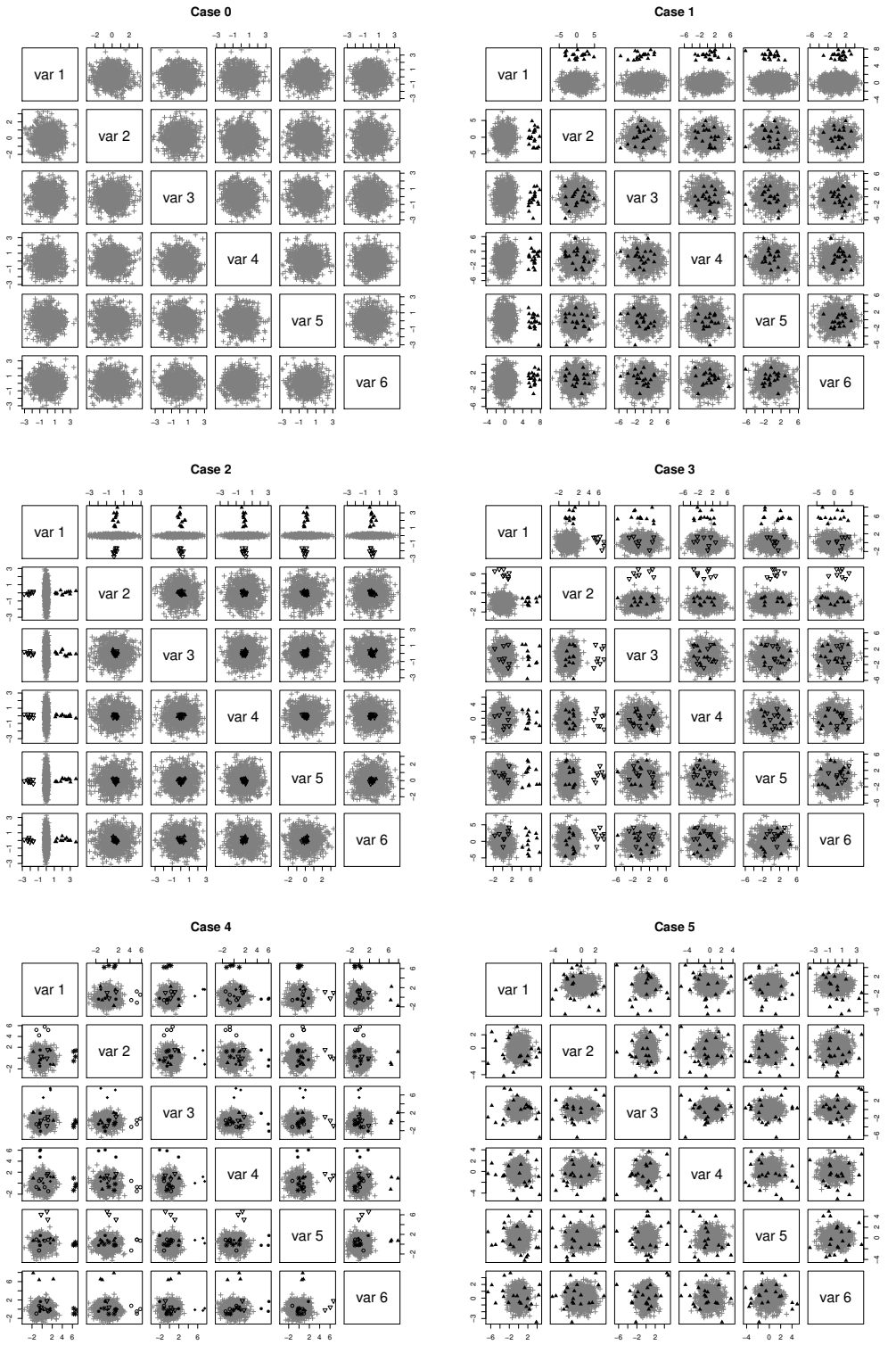



Figure 2.6 – Contents of the file Scatterplot_simulations.pdf: scatterplot matrices for one data set generated according to each of the six simulated cases and with $p = 6$.

Under the mixture distribution (2.1), we have:

$$\begin{aligned} d^2(\mathbf{X}) &= (\mathbf{X}_q - \boldsymbol{\mu}_{\mathbf{X}_q})' \boldsymbol{\Sigma}_q^{-1} (\mathbf{X}_q - \boldsymbol{\mu}_{\mathbf{X}_q}) + \sum_{i=q+1}^{p} X_i^2, \\ d_R^2(\mathbf{X}) &= \sum_{i=1}^{p} X_i^2. \end{aligned}$$

We make use of the Lindeberg-Feller central limit theorem as recalled for instance in Greene (2012), p.1119, which states that:

Let $Y_i$, $i = 1, \ldots, n$, be a sequence of independent random variables with finite means $m_i$ and finite positive variance $\sigma_i^2$. Let

$$\bar{m}_n = \frac{1}{n} \sum_{i=1}^{n} m_i \quad \text{and} \quad \bar{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^{n} \sigma_i^2.$$

If $\lim_{n \to +\infty} \max(\sigma_i)/(n\bar{\sigma}_n) = 0$ and $\lim_{n \to +\infty} \bar{\sigma}_n^2 = \bar{\sigma}^2 < \infty$ then

$$\sqrt{n} \left( \bar{Y}_n - \bar{m}_n \right) \xrightarrow[n \to +\infty]{d} \mathcal{N}(0, \bar{\sigma}^2)$$

We recall that $\mathbf{X}_{no}$ follows a normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I}_p)$ and $\mathbf{X}_{o,h}$ follows a normal distribution $\mathcal{N}(\boldsymbol{\mu}_h, \mathbf{I}_p)$, with the last $p - q$ coordinates of $\boldsymbol{\mu}_h$ equal to 0 and $h = 1, \ldots, q$. We assume that $\mathbf{X}_{no}$ and $\mathbf{X}_{o,h}$, for $h = 1, \ldots, q$, are independent, and we are interested in the behavior of the difference between the squared distance of $\mathbf{X}_o$ and of $\mathbf{X}_{no}$ for both Mahalanobis distances, when dimension $p$ increases and $q$ is fixed. We first look at the convergence in distribution of the difference of the robust Mahalanobis distances when $p$ grows to infinity and we have that

$$d_R^2(\mathbf{X}_{o,h}) - d_R^2(\mathbf{X}_{no}) = \sum_{i=1}^{p} \left( X_{o,h,i}^2 - X_{no,i}^2 \right).$$

Let denote $Y_i = X_{o,h,i}^2 - X_{no,i}^2$, $i = 1, \ldots, p$, and check the Lindeberg-Feller theorem assumptions for this sequence when $p$ goes to infinity. Note that we apply the theorem to $p$ and not to $n$. Given that the vectors $\mathbf{X}_{no}$ and $\mathbf{X}_{o,h}$ are Gaussian vectors with uncorrelated components and are independent between them, the random variables $Y_i$ are independent. For $i = 1, \ldots, p$, $X_{no,i}^2$ follows a chi-squared distribution with one degree of freedom and $X_{o,h,i}^2$ follows the same distribution for $i = q+1, \ldots, p$, while it follows a non central chi-squared distribution with one degree of freedom and noncentrality parameter $\mu_{h,i}^2$, for $i = 1, \ldots, q$. So the expectation of $Y_i$, $m_i = 0$ for $i = q+1, \ldots, p$ and $m_i = \mu_{h,i}^2$ for $i = 1, \ldots, q$. The variance of $Y_i$ is finite, positive and equal $\sigma_i^2 = 4$ for $i = q+1, \ldots, p$ and $4(\mu_{h,i}^2 + 1)$ for $i = 1, \ldots, q$. So $\bar{\sigma}_p^2 = (4/p)[p + \sum_{i=1}^{q} \mu_{h,i}^2]$ and tends to 4 when $p$ goes to infinity. Let denote $\sigma_{\max}^2 = \max(\sigma_i^2) = 4(\max\{\mu_{h,i}^2, i = 1, \ldots, q\} + 1)$, which does not depend on $p$, then we have

$$\lim_{p \to +\infty} \sigma_{\max}/(p\bar{\sigma}_p) = 0.$$

We conclude that:

$$\frac{1}{\sqrt{p}} \left( d_R^2(\mathbf{X}_{o,h}) - d_R^2(\mathbf{X}_{no}) - \sum_{i=1}^{q} \mu_{h,i}^2 \right) \xrightarrow[p \to +\infty]{d} \mathcal{N}(0, 4).$$

For the non-robust Mahalanobis distance, we have:

$$\begin{aligned} d^2(\mathbf{X}_{o,h}) - d^2(\mathbf{X}_{no}) &= (\mathbf{X}_{o,h,q} - \boldsymbol{\mu}_{\mathbf{X}_q})'\boldsymbol{\Sigma}_q^{-1}(\mathbf{X}_{o,h,q} - \boldsymbol{\mu}_{\mathbf{X}_q}) - (\mathbf{X}_{no,q} - \boldsymbol{\mu}_{\mathbf{X}_q})'\boldsymbol{\Sigma}_q^{-1}(\mathbf{X}_{no,q} - \boldsymbol{\mu}_{\mathbf{X}_q}) \\ &\quad + \sum_{i=q+1}^{p} \left( X_{o,h,i}^2 - X_{no,i}^2 \right). \end{aligned}$$

where we denote by $\mathbf{X}_{o,h,q}$ (resp. $\mathbf{X}_{no,q}$) the first $q$ rows of $\mathbf{X}_{o,h}$ (resp. of $\mathbf{X}_{no}$).

$$\text{Let } Y_1 = (\mathbf{X}_{o,h,q} - \boldsymbol{\mu}_{\mathbf{X}_q})'\boldsymbol{\Sigma}_q^{-1}(\mathbf{X}_{o,h,q} - \boldsymbol{\mu}_{\mathbf{X}_q}) - (\mathbf{X}_{no,q} - \boldsymbol{\mu}_{\mathbf{X}_q})'\boldsymbol{\Sigma}_q^{-1}(\mathbf{X}_{no,q} - \boldsymbol{\mu}_{\mathbf{X}_q})$$

and $Y_i = X_{o,h,q+i-1}^2 - X_{no,q+i-1}^2$, for $i = 2, \ldots, p-q+1$. As previously, the $Y_i$s are independent. As the difference of two non degenerate quadratic forms for $q$-dimensional Gaussian vectors, the expectation $m_1$ of $Y_1$ is finite and its variance $\sigma_1^2$ is finite and positive and does not depend on $p$. For $i = 2, \ldots, p-q+1$, $X_{no,i}^2$ and $X_{o,h,i}^2$ follow a chi-squared distribution with one degree of freedom and so the expectation $m_i$ of $Y_i$ equals 0 and the variance $\sigma_i^2 = 4$. So $\bar{m}_p = m_1/(p-q+1)$, $\bar{\sigma}_p^2 = [4(p-q) + \sigma_1^2]/(p-q+1)$ and tends to 4 when $p$ goes to infinity. Let $\sigma_{\max}^2 = \max(\sigma_i^2) = \max(4, \sigma_1^2)$, which does not depend on $p$, then we have

$$\lim_{p \to +\infty} \sigma_{\max}/(p\bar{\sigma}_p) = 0.$$

We conclude that:

$$\frac{\sqrt{p}}{p-q+1} \left( d^2(\mathbf{X}_{o,h}) - d^2(\mathbf{X}_{no}) - m_1 \right) \xrightarrow[p \to +\infty]{d} \mathcal{N}(0, 4)$$

which gives the final result.

### 2.8.2 Derivation of the eigenvalues and eigenvectors of the simultaneous diagonalization of COV and $\text{COV}_4$ for particular mixtures

Let $\mathbf{X} = (X_1, \ldots, X_p)'$ be a $p$-multivariate real random vector and denote by $\mathbf{F}_\mathbf{X}$ the distribution of $\mathbf{X}$. We assume that $p > 2$ and that $\mathbf{F}_\mathbf{X}$ admits fourth moments. The functional versions of $\text{COV}(\mathbf{F}_\mathbf{X})$ and $\text{COV}_4(\mathbf{F}_\mathbf{X})$ which are consistent at the Gaussian distribution are given by:

$$\text{COV}(\mathbf{F}_\mathbf{X}) = \mathbb{E}\left[ (\mathbf{X} - \mathbb{E}(\mathbf{X}))(\mathbf{X} - \mathbb{E}(\mathbf{X}))' \right],$$

$$\text{COV}_4(\mathbf{F}_\mathbf{X}) = \frac{1}{p+2} \mathbb{E}\left[ (\mathbf{X} - \mathbb{E}(\mathbf{X}))'\text{COV}^{-1}(\mathbf{F}_\mathbf{X})(\mathbf{X} - \mathbb{E}(\mathbf{X}))(\mathbf{X} - \mathbb{E}(\mathbf{X}))(\mathbf{X} - \mathbb{E}(\mathbf{X}))' \right].$$

We denote by $\rho_1(\mathbf{F}_\mathbf{X}) \geq \rho_2(\mathbf{F}_\mathbf{X}) \ldots \geq \rho_p(\mathbf{F}_\mathbf{X})$ the eigenvalues of $\text{COV}^{-1}(\mathbf{F}_\mathbf{X})\text{COV}_4(\mathbf{F}_\mathbf{X})$ in decreasing order. The cases we consider below correspond or are very similar to Cases 1, 2 and 5 from the simulations section. For such mixtures and the scatter pair COV-COV4, it is possible to derive conditions under which the ICS method recovers the direction of outlying observations.

**Case 1: mean-shift outlier model**

Let $\mathbf{F_X}$ be a mixture of two Gaussian distributions with different location parameters and the same definite positive covariance matrix $\mathbf{\Sigma}_1$:

$$\mathbf{X} \sim (1-\epsilon)\ \mathcal{N}(\mathbf{0}_p, \mathbf{\Sigma}_1) + \epsilon\ \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma}_1) \tag{2.4}$$

with $\epsilon < 0.5$ and $\boldsymbol{\mu} \neq \mathbf{0}_p$ a $p$-vector.
In this case, the behavior of ICS has already been established. This result is explicitly presented in Tyler et al. (2009) as a particular case of the Theorem 3. Caussinus and Ruiz-Gazen (1994) and Caussinus and Ruiz-Gazen (1995) also derived this condition as a particular case of the symmetrized version of the one-step $W$-estimate used as one of the scatter matrix while the other was the usual covariance matrix. Finally, Alashwali and Kent (2016) also recovered the same result by using arguments from Peña and Prieto (2001b) focusing on projection pursuit based on the kurtosis. As a reminder, the result is the following.

**Proposition 3.**
*Let* $\mathbf{X}$ *follow the distribution (2.4), the eigenvalues of* $\mathrm{COV}^{-1}(\mathbf{F_X})\mathrm{COV}_4(\mathbf{F_X})$ *are such that either:*

$$\begin{array}{llllllll}
(a) & \rho_1(\mathbf{F_X}) & > & \rho_2(\mathbf{F_X}) & = \quad \ldots \quad = & \rho_p(\mathbf{F_X}) & \text{if } \epsilon < (3-\sqrt{3})/6 \ (\approx 21\%), \\
(b) & \rho_1(\mathbf{F_X}) & = & \ldots & = \ \rho_{p-1}(\mathbf{F_X}) > & \rho_p(\mathbf{F_X}) & \text{if } \epsilon > (3-\sqrt{3})/6, \\
(c) & \rho_1(\mathbf{F_X}) & = & \rho_2(\mathbf{F_X}) & = \quad \ldots \quad = & \rho_p(\mathbf{F_X}) & \text{if } \epsilon = (3-\sqrt{3})/6.
\end{array}$$

*Moreover, if (a) (resp. (b)) holds then the eigenvector of* $\mathrm{COV}^{-1}(\mathbf{F_X})\mathrm{COV}_4(\mathbf{F_X})$ *associated with* $\rho_1(\mathbf{F_X})$ *(resp.* $\rho_p(\mathbf{F_X})$*) is proportional to* $\mathbf{\Sigma}_1^{-1}\boldsymbol{\mu}$.

*Remark* 1. In the simulation framework, Case 1 corresponds to model (2.4) with a percentage of contamination equal to 2% and outliers are highlighted on the first component of ICS.

**Case 2a: the barrow wheel distribution**

This distribution was suggested by Stahel and Mächler in the discussion in Tyler et al. (2009) as a "benchmark distribution for multivariate tools". This so-called barrow wheel distribution was first introduced in Hampel et al. (1986). Here, we simplify the model without loss of generality by considering no rescaling nor rotation because of the affine invariance property of the ICS method.
Let the distribution of $\mathbf{X}$ be:

$$\mathbf{X} \sim (1-\epsilon)\ \mathcal{N}(\mathbf{0}_p, \mathbf{\Sigma}_{21}) + \epsilon\ H \tag{2.5}$$

where $\mathbf{\Sigma}_{21} = \mathrm{diag}(\sigma_{11}^2, 1, \ldots, 1)$ and let $\mathbf{Y} = (Y_1, \ldots, Y_p)'$ distributed according to $H$. $H$ is such that $Y_1$ has a symmetric distribution with $Y_1^2 \sim \chi_k^2$ and is independent of $Y_2, \ldots, Y_p \sim \mathcal{N}(\mathbf{0}_p, \mathbf{\Sigma}_{22})$ with $\mathbf{\Sigma}_{22} = \sigma_{22}^2\mathbf{I}_{p-1}$. With such a model, the outliers are generated along the first direction on both sides of the main data.

Tyler et al. (2009) prove in the discussion that their Theorem 4 is still valid under the barrow wheel distribution. Restricting the analysis to COV and $\text{COV}_4$ enables us to derive a more precise result.

**Proposition 4.**
*Let* $\mathbf{X}$ *follow the distribution (2.5), the eigenvalues of* $\text{COV}^{-1}(\mathbf{F_X})\text{COV}_4(\mathbf{F_X})$ *are such that either:*

$$\begin{array}{lllllll} (a) & \rho_1(\mathbf{F_X}) & > & \rho_2(\mathbf{F_X}) & = & \dots & = \rho_p(\mathbf{F_X}) \ , \\ (b) & \rho_1(\mathbf{F_X}) & = & \dots & = & \rho_{p-1}(\mathbf{F_X}) > & \rho_p(\mathbf{F_X}), \\ (c) & \rho_1(\mathbf{F_X}) & = & \rho_2(\mathbf{F_X}) & = & \dots & = \rho_p(\mathbf{F_X}). \end{array}$$

*with* $\rho_1(\mathbf{F_X}) = \dfrac{1}{p+2}\left(\dfrac{3(1-\epsilon)\sigma_{11}^4 + \epsilon(2+k)k}{((1-\epsilon)\sigma_{11}^2 + \epsilon k)^2} + p - 1\right)$

*and* $\rho_2(\mathbf{F_X}) = \dfrac{1}{p+2}\left(\dfrac{3((1-\epsilon) + \epsilon\sigma_{22}^4)}{((1-\epsilon) + \epsilon\sigma_{22}^2)^2} + p - 1\right)$.

*Moreover, if (a) (resp. (b)) holds then the eigenvector of* $\text{COV}^{-1}(\mathbf{F_X})\text{COV}_4(\mathbf{F_X})$ *associated with* $\rho_1(\mathbf{F_X})$ *(resp. with* $\rho_p(\mathbf{F_X})$*) is proportional to* $\mathbf{e}_1 = (1, 0, \dots, 0)'$.

*Remark* 2. In the simulation framework, Case 2 corresponds to model (2.5) with $k = 5$, $\sigma_{11}^2 = 0.1$, $\sigma_{11}^2 = 0.2$ and $\epsilon = 2\%$. In this situation, $\rho_1(\mathbf{F_X}) > \rho_2(\mathbf{F_X})$ and the outliers are highlighted on the first component of ICS.

*Proof.* Let us compute the eigenvalues of $\text{COV}(\mathbf{F_X})^{-1}\text{COV}_4(\mathbf{F_X})$.

*Moments of* $Y_1$*:*
We can decompose $Y_1$ as $Y_1 = S\,C$, with $S = \begin{cases} \text{-1} & \text{with probability } 1/2 \\ 1 & \text{with probability } 1/2 \end{cases}$ and $C \sim \chi_k$.
So, $\mathbb{E}(Y_1) = 0$, $\text{var}(Y_1) = \mathbb{E}(Y_1^2) = \mathbb{E}(C^2) = k$ and $\text{var}(Y_1) = \mathbb{E}(Y_1^2) = 2k$.

*Computation of* $\text{COV}(\mathbf{F_X})^{-1}$*:*
For model (2.5), the expectation is $\mathbb{E}(\mathbf{X}) = \mathbf{0}_p$, the between covariance is $\mathbf{\Sigma}_B = \mathbf{0}$ and the within covariance matrix is $\mathbf{\Sigma}_W = \text{diag}((1-\epsilon)\sigma_{11}^2 + \epsilon k, ((1-\epsilon) + \epsilon\sigma_{22}^2), \dots)$. So,

$$\text{COV}(\mathbf{F_X}) = \begin{pmatrix} \gamma_1 & \mathbf{0} \\ \mathbf{0} & \gamma_2\mathbf{I}_{p-1} \end{pmatrix} \quad \text{and} \quad \text{COV}(\mathbf{F_X})^{-1} = \begin{pmatrix} 1/\gamma_1 & \mathbf{0} \\ \mathbf{0} & 1/\gamma_2\mathbf{I}_{p-1} \end{pmatrix},$$

with $\gamma_1 = (1-\epsilon)\sigma_{11}^2 + \epsilon k$ and $\gamma_2 = (1-\epsilon) + \epsilon\sigma_{22}^2$.

*Computation of* $\text{COV}_4(\mathbf{F_X})$*:*
The scatter matrix based on the fourth moments $\text{COV}_4$ is defined by:

$$\text{COV}_4(\mathbf{F_X}) = \frac{1}{(p+2)}\mathbb{E}(d^2(\mathbf{X} - \mathbb{E}(\mathbf{X}))(\mathbf{X} - \mathbb{E}(\mathbf{X}))')$$

where $d^2 = d(\mathbf{X})^2 = ||\text{COV}(\mathbf{F_X})^{-1/2}(\mathbf{X} - \mathbb{E}(\mathbf{X}))||^2$ is the classical squared Mahalanobis distance. Here, $\mathbb{E}(\mathbf{X}) = \mathbf{0}_p$ so,

$$\text{COV}_4(\mathbf{F_X}) = \frac{1}{(p+2)}\text{diag}(\mathbb{E}(d^2X_1^2), \dots, \mathbb{E}(d^2X_p^2))$$

as all the $X_i$ are independent and $d^2 = \frac{1}{\gamma_1}X_1^2 + \frac{1}{\gamma_2}\sum_{l=2}^p X_l^2$.

The first diagonal term is $\mathbb{E}(d^2X_1^2) = \frac{1}{\gamma_1}\mathbb{E}(X_1^4) + (p-1)\gamma_1$.
$\mathbb{E}(X_1^4)$ can be easily expressed since $X_1 \sim (1-\epsilon)Z_1 + \epsilon Y_1$ with $Z_1 \sim \mathcal{N}(0, \sigma_{11}^2)$ and $\mathbb{E}(Y_1^4) = \text{var}(Y_1^2) + \mathbb{E}(Y_1^2)^2 = (2+k)k$. Then, we apply the following properties to have an expression for $\mathbb{E}(X_1^4)$:

— Additive property of the moments:
$\mathbb{E}(X_1^4) = (1-\epsilon)\mathbb{E}(Z_1^4) + \epsilon\mathbb{E}(Y_1^4)$

— Decomposition of a fourth order moment:
$\mathbb{E}(Z_i^4) = \mathbb{E}((Z_i - \mathbb{E}(Z_i))^4) + 4\mathbb{E}((Z_i - \mathbb{E}(Z_i))^3)\mathbb{E}(Z_i) + 6\mathbb{E}((Z_i - \mathbb{E}(Z_i))^2)\mathbb{E}(Z_i)^2 + 4\mathbb{E}(Z_i - \mathbb{E}(Z_i))\mathbb{E}(Z_i)^3 + \mathbb{E}(Z_i)^4$.

— Computation of moments from a Gaussian distribution:
If $Z \sim \mathcal{N}(\mu, \sigma^2)$, then $\mathbb{E}((Z-\mu)^{2k}) = \left((2k)!\sigma^{2k}\right)/(2^k k!)$ and $\mathbb{E}((Z-\mu)^{2k+1}) = 0$.

So, $\mathbb{E}(X_1^4) = 3(1-\epsilon)\sigma_{11}^4 + \epsilon(2+k)k$.
Finally, $\mathbb{E}(d^2X_1^2) = \frac{1}{\gamma_1}(3(1-\epsilon)\sigma_{11}^4 + \epsilon(2+k)k) + (p-1)\gamma_1$.

All the other diagonal terms are equal to $\mathbb{E}(d^2X_j^2) = \frac{1}{\gamma_2}\mathbb{E}(X_j^4) + (p-1)\gamma_2$ for $j = 2, \ldots, p$. Since $X_j \sim (1-\epsilon)Z_1 + \epsilon Z_2$ with $Z_1 \sim \mathcal{N}(0,1)$ and $Z_2 \sim \mathcal{N}(0, \sigma_{22}^2)$, we can apply the same procedure as previously and we obtain: $\mathbb{E}(X_j^4) = 3((1-\epsilon) + \epsilon\sigma_{22}^4)$ and so, for $j = 2, \ldots, p$, $\mathbb{E}(d^2X_j^2) = \frac{1}{\gamma_2}(3((1-\epsilon) + \epsilon\sigma_{22}^4)) + (p-1)\gamma_2$.

*Computation of* $\text{COV}(\mathbf{F_X})^{-1}\text{COV}_4(\mathbf{F_X})$*:*
Now we can express $\text{COV}(\mathbf{F_X})^{-1}\text{COV}_4(\mathbf{F_X})$ as:

$$\text{COV}(\mathbf{F_X})^{-1}\text{COV}_4(\mathbf{F_X}) = \frac{1}{(p+2)}\begin{pmatrix} \frac{1}{\gamma_1}\mathbb{E}(d^2X_1^2) & \mathbf{0} \\ \mathbf{0} & \frac{1}{\gamma_2}\mathbb{E}(d^2X_j^2)\mathbf{I}_{p-1} \end{pmatrix}$$

So, the eigenvalues of $\text{COV}(\mathbf{F_X})^{-1}\text{COV}_4(\mathbf{F_X})$ are simply its diagonal terms and the eigenvector associated with $\mathbb{E}(d^2X_1^2)/\gamma_1$ is $\mathbf{e}_1$.
If $\frac{1}{\gamma_1}\mathbb{E}(d^2X_1^2) > \frac{1}{\gamma_2}\mathbb{E}(d^2X_j^2)$ then $\rho_1(\mathbf{F_X}) > \rho_2(\mathbf{F_X})$,

$$\text{with: } \rho_1(\mathbf{F_X}) = \frac{1}{p+2}\left(\frac{3(1-\epsilon)\sigma_{11}^4 + \epsilon(2+k)k}{((1-\epsilon)\sigma_{11}^2 + \epsilon k)^2} + p - 1\right)$$

$$\text{and } \rho_2(\mathbf{F_X}) = \frac{1}{p+2}\left(\frac{3((1-\epsilon) + \epsilon\sigma_{22}^4)}{((1-\epsilon) + \epsilon\sigma_{22}^2)^2} + p - 1\right).$$

and the eigenvector associated with $\mathbb{E}(d^2X_1^2)/\gamma_1$ is $\mathbf{e}_1$. And so Proposition 4 is proven. □

The eigenvalues expression can be easily simplified in the case when $\mathbf{\Sigma}_{21} = \mathbf{\Sigma}_{22}$ and so we derive the following corollary.
*Corollary* 1. If $\mathbf{X}$ follows the distribution (2.5) with $\sigma_{11} = \sigma_{22} = 1$, the eigenvalues of $\text{COV}^{-1}(\mathbf{F_X})\text{COV}_4(\mathbf{F_X})$ are such that either:

$$\begin{array}{llllllll}
\text{(a)} & \rho_1(\mathbf{F_X}) & > & \rho_2(\mathbf{F_X}) & = \ \ldots \ = & \rho_p(\mathbf{F_X}) & \text{if } \epsilon < (k-3)/(3(k-1)), \\
\text{(b)} & \rho_1(\mathbf{F_X}) & = \ \ldots \ = & \rho_{p-1}(\mathbf{F_X}) & > & \rho_p(\mathbf{F_X}) & \text{if } \epsilon > (k-3)/(3(k-1)), \\
\text{(c)} & \rho_1(\mathbf{F_X}) & = & \rho_2(\mathbf{F_X}) & = \ \ldots \ = & \rho_p(\mathbf{F_X}) & \text{if } \epsilon = (k-3)/(3(k-1)).
\end{array}$$

Moreover, if (a) (resp. (b)) holds then the eigenvector of $\text{COV}^{-1}(\mathbf{F_X})\text{COV}_4(\mathbf{F_X})$ associated with $\rho_1(\mathbf{F_X})$ (resp. $\rho_p(\mathbf{F_X})$) is proportional to $\mathbf{e}_1$.

The bound $(k-3)/(3(k-1))$ on the contamination is minimum for $k=4$ and equals $1/9$. It increases with $k$ and its limit equals to $1/3 \simeq 33\%$ of contamination when $k$ grows to infinity.

**Case 2b: symmetric contamination of a Gaussian distribution**

We can also mimic the barrow wheel distribution by the following mixture of three Gaussian distributions:

$$\mathbf{X} \sim (1-\epsilon)\,\mathcal{N}(\mathbf{0}_p, \mathbf{\Sigma}_{21}) + \tfrac{\epsilon}{2}\,\mathcal{N}(\delta\mathbf{e}_1, \mathbf{\Sigma}_{22}) + \tfrac{\epsilon}{2}\,\mathcal{N}(-\delta\mathbf{e}_1, \mathbf{\Sigma}_{22}) \tag{2.6}$$

with $\mathbf{\Sigma}_{21} = \mathrm{diag}(\sigma_{11}^2, \sigma_{12}^2, \ldots, \sigma_{12}^2)$, $\mathbf{\Sigma}_{22} = \mathrm{diag}(\sigma_{21}^2, \sigma_{22}^2, \ldots, \sigma_{22}^2)$ and $\delta \neq 0$.
In this case, we can derive the following proposition.

**Proposition 5.**
*Let* $\mathbf{X}$ *follow the distribution (2.6), the eigenvalues of* $\mathrm{COV}^{-1}(\mathbf{F_X})\mathrm{COV}_4(\mathbf{F_X})$ *are such that either:*

$$\begin{array}{lccccccc}
(a) & \rho_1(\mathbf{F_X}) & > & \rho_2(\mathbf{F_X}) & = & \ldots & = & \rho_p(\mathbf{F_X})\ , \\
(b) & \rho_1(\mathbf{F_X}) & = & \ldots & = & \rho_{p-1}(\mathbf{F_X}) & > & \rho_p(\mathbf{F_X}), \\
(c) & \rho_1(\mathbf{F_X}) & = & \rho_2(\mathbf{F_X}) & = & \ldots & = & \rho_p(\mathbf{F_X}).
\end{array}$$

*with* $\rho_1(\mathbf{F_X}) = \dfrac{1}{p+2}\left(\dfrac{3(1-\epsilon)\sigma_{11}^4 + \epsilon(3\sigma_{21}^4 + 6\sigma_{21}^2\delta^2 + \delta^4)}{((1-\epsilon)\sigma_{11}^2 + \epsilon(\sigma_{21}^2 + \delta^2))^2} + p - 1\right)$

*and* $\rho_2(\mathbf{F_X}) = \dfrac{1}{p+2}\left(\dfrac{3((1-\epsilon)\sigma_{12}^4 + \epsilon\sigma_{22}^4)}{((1-\epsilon)\sigma_{12}^2 + \epsilon\sigma_{22}^2)^2} + p - 1\right)$.

*Moreover, if (a) (resp. (b)) holds then the eigenvector of* $\mathrm{COV}^{-1}(\mathbf{F_X})\mathrm{COV}_4(\mathbf{F_X})$ *associated with* $\rho_1(\mathbf{F_X})$ *(resp.* $\rho_p(\mathbf{F_X})$*) is proportional to* $\mathbf{e}_1$.

*Proof.* Let us compute the eigenvalues of $\mathrm{COV}(\mathbf{F_X})^{-1}\mathrm{COV}_4(\mathbf{F_X})$.

*Computation of* $\mathrm{COV}(\mathbf{F_X})^{-1}$*:*
For the model (2.6), the expectation is $\mathbb{E}(\mathbf{X}) = \mathbf{0}_p$, the within covariance matrix is $\mathbf{\Sigma}_W = (1-\epsilon)\mathbf{\Sigma}_{21} + \epsilon\mathbf{\Sigma}_{22} = \mathrm{diag}((1-\epsilon)\sigma_{11}^2 + \epsilon\sigma_{21}^2, ((1-\epsilon)\sigma_{12}^2 + \epsilon\sigma_{22}^2), \ldots)$ and the between covariance is $\mathbf{\Sigma}_B = \epsilon\delta^2\mathbf{e}_1\mathbf{e}_1'$. So,

$$\mathrm{COV}(\mathbf{F_X}) = \begin{pmatrix} \gamma_1 & \mathbf{0} \\ \mathbf{0} & \gamma_2\mathbf{I}_{p-1} \end{pmatrix} \quad \text{and} \quad \mathrm{COV}(\mathbf{F_X})^{-1} = \begin{pmatrix} 1/\gamma_1 & \mathbf{0} \\ \mathbf{0} & 1/\gamma_2\mathbf{I}_{p-1} \end{pmatrix},$$

with $\gamma_1 = (1-\epsilon)\sigma_{11}^2 + \epsilon\sigma_{21}^2 + \epsilon\delta^2$ and $\gamma_2 = (1-\epsilon)\sigma_{12}^2 + \epsilon\sigma_{22}^2$.

*Computation of* $\mathrm{COV}_4(\mathbf{F_X})$*:*
As already defined in Proof 2.8.2,

$$\mathrm{COV}_4(\mathbf{F_X}) = \frac{1}{(p+2)}\mathrm{diag}(\mathbb{E}(d^2X_1^2), \ldots, \mathbb{E}(d^2X_p^2))$$

where $d^2 = \frac{1}{\gamma_1}X_1^2 + \frac{1}{\gamma_2}\sum_{l=2}^p X_l^2$.

The first diagonal term is $\mathbb{E}(d^2X_1^2) = \frac{1}{\gamma_1}\mathbb{E}(X_1^4) + (p-1)\gamma_1$.
$\mathbb{E}(X_1^4)$ can be easily expressed since $X_1 \sim (1-\epsilon)Z_1 + \frac{\epsilon}{2}Z_2 + \frac{\epsilon}{2}Z_3$ with $Z_1 \sim \mathcal{N}(0, \sigma_{11}^2)$, $Z_2 \sim \mathcal{N}(\delta, \sigma_{21}^2)$ and $Z_3 \sim \mathcal{N}(-\delta, \sigma_{21}^2)$. Then, we apply the same properties as in Proof 2.8.2 and so we obtain $\mathbb{E}(X_1^4) = 3(1-\epsilon)\sigma_{11}^4 + \epsilon(3\sigma_{21}^4 + 6\sigma_{21}^2\delta^2 + \delta^4)$.
Finally, $\mathbb{E}(d^2X_1^2) = \frac{1}{\gamma_1}(3(1-\epsilon)\sigma_{11}^4 + \epsilon(3\sigma_{21}^4 + 6\sigma_{21}^2\delta^2 + \delta^4)) + (p-1)\gamma_1$.

All the other diagonal terms are equal to $\mathbb{E}(d^2X_j^2) = \frac{1}{\gamma_2}\mathbb{E}(X_j^4) + (p-1)\gamma_2$ for $j = 2, \ldots, p$. Since $X_j \sim (1-\epsilon)Z_1 + \epsilon Z_2$ with $Z_1 \sim \mathcal{N}(0, \sigma_{12}^2)$ and $Z_2 \sim \mathcal{N}(0, \sigma_{22}^2)$, we can apply the same procedure as previously and we obtain: $\mathbb{E}(X_j^4) = 3((1-\epsilon)\sigma_{12}^4 + \epsilon\sigma_{22}^4)$ and so, for $j = 2, \ldots, p$, $\mathbb{E}(d^2X_j^2) = \frac{1}{\gamma_2}(3((1-\epsilon)\sigma_{12}^4 + \epsilon\sigma_{22}^4)) + (p-1)\gamma_2$.

*Computation of* $\mathrm{COV}(\mathbf{F_X})^{-1}\mathrm{COV}_4(\mathbf{F_X})$*:*
Now we can express $\mathrm{COV}(\mathbf{F_X})^{-1}\mathrm{COV}_4(\mathbf{F_X})$ as:

$$\mathrm{COV}(\mathbf{F_X})^{-1}\mathrm{COV}_4(\mathbf{F_X}) = \frac{1}{(p+2)}\begin{pmatrix} \frac{1}{\gamma_1}\mathbb{E}(d^2X_1^2) & \mathbf{0} \\ \mathbf{0} & \frac{1}{\gamma_2}\mathbb{E}(d^2X_j^2)\mathbf{I}_{p-1} \end{pmatrix}$$

So, the eigenvalues of $\mathrm{COV}(\mathbf{F_X})^{-1}\mathrm{COV}_4(\mathbf{F_X})$ are simply its diagonal terms and the eigenvector associated with $\mathbb{E}(d^2X_1^2)/\gamma_1$ is $\mathbf{e}_1$.
If $\frac{1}{\gamma_1}\mathbb{E}(d^2X_1^2) > \frac{1}{\gamma_2}\mathbb{E}(d^2X_j^2)$ then $\rho_1(\mathbf{F_X}) > \rho_2(\mathbf{F_X})$,

$$\text{with: } \rho_1(\mathbf{F_X}) = \frac{1}{p+2}\left(\frac{3(1-\epsilon)\sigma_{11}^4 + \epsilon(3\sigma_{21}^4 + 6\sigma_{21}^2\delta^2 + \delta^4)}{((1-\epsilon)\sigma_{11}^2 + \epsilon(\sigma_{21}^2 + \delta^2))^2} + p - 1\right)$$

$$\text{and } \rho_2(\mathbf{F_X}) = \frac{1}{p+2}\left(\frac{3((1-\epsilon)\sigma_{12}^4 + \epsilon\sigma_{22}^4)}{((1-\epsilon)\sigma_{12}^2 + \epsilon\sigma_{22}^2)^2} + p - 1\right).$$

And so Proposition 5 is proven. □

The eigenvalues expression above can be easily simplified in the case of equal covariance matrices $\mathbf{\Sigma}_{21} = \mathbf{\Sigma}_{22}$. Moreover, in this case, given that the ICS method is affine invariant, we do not need to assume diagonal matrices. We have thus the following corollary where the condition for the three cases depends only on the percentage of contamination $\epsilon$ and not on the location parameter $\delta$.

*Corollary* 2. For $\mathbf{X}$ simulated as in (2.6) with $\mathbf{\Sigma}_{21} = \mathbf{\Sigma}_{22} = \mathbf{I}_p$, the eigenvalues of $\mathrm{COV}^{-1}\mathrm{COV}_4$ are such that either:

$$\begin{array}{llllllll}
\text{(a)} & \rho_1(\mathbf{F_X}) & > & \rho_2(\mathbf{F_X}) & = & \ldots & = & \rho_p(\mathbf{F_X}) \quad \text{if } \epsilon < 1/3, \\
\text{(b)} & \rho_1(\mathbf{F_X}) & = & \ldots & = & \rho_{p-1}(\mathbf{F_X}) & > & \rho_p(\mathbf{F_X}) \quad \text{if } \epsilon > 1/3, \\
\text{(c)} & \rho_1(\mathbf{F_X}) & = & \rho_2(\mathbf{F_X}) & = & \ldots & = & \rho_p(\mathbf{F_X}) \quad \text{if } \epsilon = 1/3.
\end{array}$$

Moreover, if (a) (resp. (b)) holds then the eigenvector of $\mathrm{COV}^{-1}(\mathbf{F_X})\mathrm{COV}_4(\mathbf{F_X})$ associated with $\rho_1(\mathbf{F_X})$ (resp. $\rho_p(\mathbf{F_X})$) is proportional to $\mathbf{e}_1$.

**Case 5: scale-shift outlier model**

Let $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_p)'$ be a $p$-multivariate real random vector and assume the distribution of $\mathbf{X}$ is a mixture of two Gaussian distributions with the same location parameters

but with a scale change:

$$\mathbf{X} \sim (1-\epsilon)\mathcal{N}(\mathbf{0}_p, \mathbf{I}_p) + \epsilon\mathcal{N}(\mathbf{0}_p, \mathbf{\Sigma}_5) \tag{2.7}$$

with $\epsilon < 0.5$, $\mathbf{\Sigma}_5 = \text{diag}(\alpha\mathbf{I}_q, \mathbf{I}_{p-q})$, $q < p$ and $\alpha > 1$.
This model generates outliers in up to $q$ directions via a scale-shift. In this context, the following proposition arises.

**Proposition 6.**
*Let* $\mathbf{X}$ *follow the distribution (2.7), then the eigenvalues of* $\text{COV}^{-1}(\mathbf{F_X})\text{COV}_4(\mathbf{F_X})$ *are such that:*

$$\rho_1(\mathbf{F_X}) = \cdots = \rho_q(\mathbf{F_X}) > \rho_{q+1}(\mathbf{F_X}) = \ldots \quad = \rho_p(\mathbf{F_X})$$

*Moreover, the eigenvectors of* $\text{COV}^{-1}(\mathbf{F_X})\text{COV}_4(\mathbf{F_X})$ *associated with the* $q$ *largest eigenvalues span the subspace spanned by* $\{\mathbf{e}_1, \ldots, \mathbf{e}_q\}$.

Note that if $q = p$ then all the eigenvalues are equal and ICS is not informative and leads to the Mahalanobis distance.

*Remark* 3. In the simulation framework, Case 5 is similar to model (2.7) with $\alpha = 5$ and thus if $p > 6$, $\rho_1(\mathbf{F_X}) = \cdots = \rho_6(\mathbf{F_X}) > \rho_7(\mathbf{F_X}) = \ldots \quad = \rho_p(\mathbf{F_X})$ and the outliers are highlighted on the first six components, independently of the percentage of contamination $\epsilon$.

*Proof.* Let us compute the eigenvalues of $\text{COV}(\mathbf{F_X})^{-1}\text{COV}_4(\mathbf{F_X})$.

*Computation of* $\text{COV}(\mathbf{F_X})^{-1}$*:*
For the model (2.7), the expectation is $\mathbb{E}(\mathbf{X}) = \mathbf{0}_p$, the within covariance matrix is $\mathbf{\Sigma}_W = \text{diag}(\gamma_1\mathbf{I}_q, \mathbf{I}_{p-q})$, with $\gamma_1 = (1-\epsilon) + \alpha\epsilon$, and the between covariance is $\mathbf{\Sigma}_B = \mathbf{0}_p$. So,

$$\text{COV}(\mathbf{F_X}) = \mathbf{\Sigma}_W \begin{pmatrix} \gamma_1\mathbf{I}_q & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{p-q} \end{pmatrix} \text{ and } \text{COV}(\mathbf{F_X})^{-1} = \begin{pmatrix} 1/\gamma_1\mathbf{I}_q & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{p-q} \end{pmatrix},$$

with $\gamma_1 = (1-\epsilon) + \alpha\epsilon$.

*Computation of* $\text{COV}_4(\mathbf{F_X})$*:*
As already defined in Proof 2.8.2,

$$\text{COV}_4(\mathbf{F_X}) = \frac{1}{(p+2)}\text{diag}(\mathbb{E}(d^2X_1^2), \ldots, \mathbb{E}(d^2X_p^2))$$

where $d^2 = \frac{1}{\gamma_1}\sum_{l=1}^{q} X_l^2 + \sum_{j=q+1}^{p} X_j^2$.

The first $q$ diagonal terms are equal to $\mathbb{E}(d^2X_l^2) = \mathbb{E}(d^2X_1^2)$ for $l = 1, \ldots, q$ and $\mathbb{E}(d^2X_1^2) = \frac{1}{\gamma_1}\mathbb{E}(X_1^4) + (p-1)\gamma_1$. $\mathbb{E}(X_1^4)$ can be easily expressed since $X_1 \sim (1-\epsilon)Z_1 + \epsilon Z_2$ with $Z_1 \sim \mathcal{N}(0,1)$ and $Z_2 \sim \mathcal{N}(0,\alpha)$. Then, we apply the same properties as in Proof 2.8.2 and so we obtain $\mathbb{E}(X_1^4) = 3(1+\epsilon(\alpha^2-1))$. Finally, $\mathbb{E}(d^2X_1^2) = \frac{3}{\gamma_1}(1+\epsilon(\alpha^2-1)) + (p-1)\gamma_1$.

All the other $p-q$ diagonal terms are equal to $\mathbb{E}(d^2X_j^2) = \mathbb{E}(d^2X_{q+1}^2) = \mathbb{E}(X_j^4) + (p-1)$ for $j = q+1, \ldots, p$. Since $X_{q+1} \sim \mathcal{N}(0,1)$, $\mathbb{E}(X_{q+1}^4) = 3$ and so, $\mathbb{E}(d^2X_{q+1}^2) = 3 + (p-1) = p+2$.

*Computation of* $\mathrm{COV}(\mathbf{F_X})^{-1}\mathrm{COV}_4(\mathbf{F_X})$*:*
Now we can express $\mathrm{COV}(\mathbf{F_X})^{-1}\mathrm{COV}_4(\mathbf{F_X})$ as:

$$\mathrm{COV}(\mathbf{F_X})^{-1}\mathrm{COV}_4(\mathbf{F_X}) = \frac{1}{(p+2)}\begin{pmatrix} \frac{1}{\gamma_1}\mathbb{E}(d^2X_1^2)\mathbf{I}_q & \mathbf{0} \\ \mathbf{0} & \frac{1}{\gamma_2}\mathbb{E}(d^2X_{q+1}^2)\mathbf{I}_{p-q} \end{pmatrix}$$

So, the eigenvalues of $\mathrm{COV}(\mathbf{F_X})^{-1}\mathrm{COV}_4(\mathbf{F_X})$ are simply its diagonal terms and the vector space spanned by the eigenvectors associated with $\mathbb{E}(d^2X_1^2)/\gamma_1$ is the one spanned by $\{\mathbf{e}_1, \ldots, \mathbf{e}_q\}$.
If $\frac{1}{\gamma_1}\mathbb{E}(d^2X_1^2) > \mathbb{E}(d^2X_{q+1}^2)$ then $\rho_1(\mathbf{F_X}) = \cdots = \rho_q(\mathbf{F_X}) > \rho_{q+1}(\mathbf{F_X}) = \cdots = \rho_p(\mathbf{F_X})$. This condition is equivalent to: $\frac{3}{\gamma_1^2}(1+\epsilon(\alpha^2-1)) + (p-1) > p+2 \Leftrightarrow \frac{1}{\gamma_1^2}(1+\epsilon(\alpha^2-1)) > 1$ with $\gamma_1 = (1-\epsilon) + \alpha\epsilon$. It leads to the following inequality: $(1-\alpha)^2(1-\epsilon) > 0$ which is true for $\alpha > 1$ and so the outliers are always revealed on the first $q$ components, as long as $\alpha > 1$. If $\alpha = 1$ then all the eigenvalues are equal and ICS fails to detect the structure of outlierness. □

*Remark* 4. Given the affine equivariance of ICS, the result can be generalized to the following mixture model where $\mathbf{\Sigma}_{51}$ (resp. $\mathbf{\Sigma}_{52}$) is a $q \times q$ (resp. $(p-q) \times (p-q)$) matrix not necessarily diagonal:

$$\mathbf{X} \sim (1-\epsilon)\,\mathcal{N}\left(\boldsymbol{\mu}, \begin{pmatrix} \mathbf{\Sigma}_{51} & 0 \\ 0 & \mathbf{\Sigma}_{52} \end{pmatrix}\right) + \epsilon\,\mathcal{N}\left(\boldsymbol{\mu}, \begin{pmatrix} \alpha\mathbf{\Sigma}_{51} & 0 \\ 0 & \mathbf{\Sigma}_{52} \end{pmatrix}\right)$$

with $\epsilon < 0.5$, $\mu$ is a $p$ vector, $q < p$ and $\alpha > 1$.

# Chapter 3

# Unsupervised outlier detection with the R package ICSOutlier

This chapter is a reprint of Archimbaud, A., Nordhausen, K., and Ruiz-Gazen, A. (2017). Unsupervised outlier detection with ICSOutlier. *accepted under revision*. It is followed by a partial reprint of the manual file of the new package we recently developed: ICSShiny.

**Abstract**

*Detecting outliers in a multivariate and Unsupervised context is an important and ongoing problem notably for quality control. Many statistical methods are already implemented in R and are briefly surveyed in the present paper. But only a few lead to the clear identification of potential outliers. In the case of a small level of contamination, the Invariant Coordinate Selection (ICS) method shows remarkable properties for identifying outliers that lie on a low-dimensional subspace in its first invariant components. It is implemented in the ICSOutlier package. The main* `ics.outlier` *function of the package offers the possibility of labelling potential outliers in an completely automated way. The detailed functionalities of this function are explored on four examples, including two real examples in quality control. Comparing with several other approaches, ICS is efficient in properly identifying some defective products while not detecting too many false positives.*

## Sommaire


**3.1 Introduction . . . . . 81**
**3.2 Invariant Coordinate Selection (ICS) for outlier detection . . . 82**
3.2.1 Principle . . . . . 82
3.2.2 Invariant Coordinates Selection . . . . . 84
3.2.3 Measure of outlierness . . . . . 85
3.2.4 Outlier identification . . . . . 85
**3.3 Using package ICSOutlier . . . . . 85**
**3.4 Examples . . . . . 87**
3.4.1 Example with no outlier . . . . . 87
3.4.2 HTP data set . . . . . 88
3.4.3 Reliability data set . . . . . 92
3.4.4 HBK data set . . . . . 95
**3.5 Conclusion and future developments . . . . . 97**
**3.6 Complements on Chapter 3: the ICSShiny package . . . . . 99**

## 3.1 Introduction

The unsupervised detection of multivariate outliers is a timeless subject of interest in statistics, see Aggarwal (2017), Hodge and Austin (2004), Hadi et al. (2009) for a complete overview. Indeed, this can be the goal of the analysis, like in fraud detection, in medical applications or manufacturing-defect detection. It can also be used for preprocessing in almost any statistical analysis that is sensitive to the presence of outliers, e.g., Principal Component Analysis (PCA). Many statistical methods exist and have already been implemented in R through tens of packages. Only some of these packages are dedicated to the unsupervised context and are mentioned below. One of the most common methods for multivariate outlier detection is the Mahalanobis Distance (MD), and many packages are based on this distance: mvoutlier (Filzmoser and Gschwandtner, 2015), CerioliOutlierDetection (Green and Martin, 2017a), rrcovHD (Todorov, 2016), faoutlier (Chalmers and Flora, 2015). Additionally, traditional methods such as the angle-based methods are implemented in the packages abodOutlier (Jimenez, 2015), HighDimOut (Fan, 2015), the distribution-based methods in alphaOutlier (Rehage and Kuhnt, 2016), extremevalues (van der Loo, 2010), HDoutliers (Wilkinson, 2016), outliers (Komsta, 2011), the density-based methods in DMwR (Torgo, 2010) or DMwR2 (Torgo, 2016), HighDimOut, ldbod (Williams, 2016), Rlof (Hu et al., 2015) and the depth-based methods in depth (Genest et al., 2012). Finally, other approaches have been implemented, such as the one based on Projection Pursuit (PP) in REPPlab (Fischer et al., 2016b), and PCA in OutlierDC (Eo and Cho, 2014), pcadapt (Luu et al., 2016), rrcov (Todorov and Filzmoser, 2009) and rrcovHD. However, it is important to note that all the implemented functions do not necessary return outlierness measures and the identities of the outlying observations. Limiting only to some packages fulfilling these conditions and to multivariate methods for numerical variables leads to focusing on mvoutlier, CerioliOutlierDetection and rrcov. These packages implement modified versions of the robust Mahalanobis distance, different algorithms for the outlier identification in high dimensions (Filzmoser and Todorov, 2013), classical and robust PCA. In PCA, the labelling of the outliers is based on the comparison of two outlierness measures: a score distance (SD) based on the first principal components and a distance in the orthogonal space (OD). All these methods rely on distances. To extend the analysis, we include comparisons with Local Outlier Factor (LOF) (Rlof package) and Angle-Based Outlier Factor (abodOutlier package) approaches even if they only return outlierness measures and not the identities of the outlying observations.

These methods dedicated to outlier detection in an unsupervised context have different properties. Contrary to PCA, the Mahalanobis distance and its modified versions are invariant under any affine transformation as long as the location and scatter estimators are affine equivariant. However, several drawbacks have been noticed in the use of these distances. First, Filzmoser et al. (2005) highlight the fact that an outlier is not necessarily an extreme value. So, using a fixed threshold to label outliers for every data set is not the wisest solution and the threshold should be adjusted to the sample size. Moroever,

as Cerioli (2010), they note that the threshold used usually is not adapted to the case in which there is no outlier in the data. Finally, for Cerioli (2010), the comparison of the distances of each observation to the threshold should not be performed independently without taking into account any simultaneous adjustments. Otherwise, the method may lead to many false positives, i.e. non outlying observations that are identified as outliers. Note that this is a serious concern that applies to almost all the outlier-detection methods. To the best of our knowledge, the mvoutlier and CerioliOutlierDetection packages are the first ones that tend to correct these drawbacks, but only for the robust Mahalanobis distance with the Minimum Covariance Determinant (MCD) estimators.

The ICSOutlier (Archimbaud et al., 2016) package is another approach to the detection of outliers. Its current version is able to handle a small proportion of outliers that lie on a subspace, using the affine equivariant Invariant Coordinate Selection (ICS) method, as presented by Archimbaud et al. (2016). The method relies on a generalized diagonalization and on the selection of the invariant components associated with the largest eigenvalues. Note that when more than say 10% of the observations are suspected to be outlying, invariant components associated with the smallest eigenvalues may also be of interest and this alternative will be considered in a future version of the package. ICS fulfills all the properties mentioned previously: (i) detect the absence of outliers, (ii) not be sensitive to the standardization of the data, (iii) be a multivariate method which controls the number of observations identified as outliers. Moreover, (iv) it is easy to use even for non-statisticians. The restriction to a small proportion of outliers that belong to a subspace is of interest in some areas of manufacturing products with a high level of quality control (e.g. automotive, avionics or aerospace). In such areas, engineers are usually not experts in statistics and they need to have an automatic way to perform efficient outlier detection.

In the following sections, the principle of the Invariant Coordinate Selection (ICS) method is recalled as well as the three steps of the outlier detection procedure: (i) invariant components selection, (ii) outlierness measure definition and (iii) outlier identification. Then, we explain how to use the ICSOutlier package and finally we analyze the efficiency of the ICS method compared to other methods on four examples, including a new real data set included in the package.

## 3.2 Invariant Coordinate Selection (ICS) for outlier detection

### 3.2.1 Principle

The ICS method is a powerful method designed for exploring multivariate data by revealing their structure (Nordhausen et al., 2008). As explained in Tyler et al. (2009), it is based on a simultaneous spectral decomposition of two scatter matrices and leads to an affine invariant coordinate system. Readers not so familiar with the concept of

scatter matrices are refered for example to Rousseeuw and Hubert (2013), Nordhausen and Tyler (2015) and references therein for definitions, examples and properties. The ICS (Nordhausen et al., 2008) package has been available for a few years, and contains a function called `ics` for implementing the ICS method. However, the `ics` function cannot be used for directly outlier detection, and a new function, `ics2`, had to be added to the ICS package (version 1.3-0). The arguments of `ics2` are mostly the same as those of the function `ics`, except now it is necessary to add the location vectors associated with the scatter estimators. The main function `ics.outlier` of the package ICSOutlier uses the output of the `ics2` function as an input. The function `ics.outlier` implements the three steps of the procedure of outlier detection and returns in particular the observations labelled as outlier.

Following Archimbaud et al. (2016), for a $p$-variate dataset $\mathbf{X}_n = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$, let $\mathbf{V}_1(\mathbf{X}_n)$ and $\mathbf{V}_2(\mathbf{X}_n)$ be two affine equivariant scatter matrices (symmetric and definite positive) and $\mathbf{m}_1(\mathbf{X}_n)$ and $\mathbf{m}_2(\mathbf{X}_n)$ their associated location estimators. ICS is looking for the diagonal $p \times p$ matrix $\mathbf{D}(\mathbf{X}_n)$ containing the eigenvalues of $\mathbf{V}_1(\mathbf{X}_n)^{-1}\mathbf{V}_2(\mathbf{X}_n)$ in decreasing order and the $p \times p$ matrix $\mathbf{B}(\mathbf{X}_n) = (\mathbf{b}_1, \ldots, \mathbf{b}_p)'$ containing the corresponding eigenvectors as its rows and such that

$$\mathbf{B}(\mathbf{X}_n)\mathbf{V}_1(\mathbf{X}_n)\mathbf{B}'(\mathbf{X}_n) = \mathbf{I}_p.$$

We have:

$$\mathbf{V}_1(\mathbf{X}_n)^{-1}\mathbf{V}_2(\mathbf{X}_n)\mathbf{B}(\mathbf{X}_n)' = \mathbf{B}(\mathbf{X}_n)'\mathbf{D}(\mathbf{X}_n).$$

Letting $\mathbf{V}_1(\mathbf{X}_n)$ be "more robust" than $\mathbf{V}_2(\mathbf{X}_n)$, the interpretation of these eigenvalues is the main point of the ICS method: they correspond to some kurtosis measure that depends on the choice of $\mathbf{V}_1(\mathbf{X}_n)$ and $\mathbf{V}_2(\mathbf{X}_n)$ . For example if $\mathbf{V}_1$ is the usual empirical variance-covariance matrix and $\mathbf{V}_2$ is the scatter matrix based on the fourth moments, the eigenvalues are ordered according to their classical Pearson kurtosis values in decreasing order. In this case and for a small proportion of outliers, it is advisable to analyze the projections that maximize the kurtosis and are associated with the largest eigenvalues (Archimbaud et al., 2016).

Then, using the location estimator $\mathbf{m}_1(\mathbf{X}_n)$ associated with the scatter matrix $\mathbf{V}_1(\mathbf{X}_n)$ , the corresponding centered scores are obtained as $\mathbf{z}_i = \mathbf{B}(\mathbf{X}_n)(\mathbf{x}_i - \mathbf{m}_1(\mathbf{X}_n))$ for $i = 1, \ldots, n$, so that:

$$\mathbf{Z}_n = (\mathbf{z}_1, \ldots, \mathbf{z}_n)' = (\mathbf{X}_n - \mathbf{1}_n\mathbf{m}_1'(\mathbf{X}_n))\mathbf{B}'(\mathbf{X}_n).$$

They are called the invariant components (IC) because of their affine invariance property in the sense that if $\mathbf{X}_n^* = \mathbf{X}_n\mathbf{A} + \mathbf{1}_n\mathbf{b}'$ for any $p \times p$ regular matrix $\mathbf{A}$ and any $p$-vector $\mathbf{b}$,

$$(\mathbf{X}_n^* - \mathbf{1}_n\mathbf{m}_1(\mathbf{X}_n^*)')\mathbf{B}(\mathbf{X}_n^*)' = (\mathbf{X}_n - \mathbf{1}_n\mathbf{m}_1(\mathbf{X}_n)')\mathbf{B}(\mathbf{X}_n)'\mathbf{J},$$

where $\mathbf{1}_n$ is a $n$-vector of ones and $\mathbf{J}$ is a $p \times p$ diagonal matrix with diagonal elements $\pm 1$, which means the invariant coordinates change at most their signs.

These scores are related to the Mahalanobis Distance (MD). For any observation $\mathbf{x}_i$ with $i = 1, \ldots, n$, the squared Euclidian norm of its centered invariant coordinates is exactly the squared Mahalanobis distance from $\mathbf{m}_1(\mathbf{X}_n)$ in the sense of $\mathbf{V}_1(\mathbf{X}_n)$:

$$\mathbf{z}_i'\mathbf{z}_i = (\mathbf{x}_i - \mathbf{m}_1(\mathbf{X}_n))'\mathbf{V}_1(\mathbf{X}_n)^{-1}(\mathbf{x}_i - \mathbf{m}_1(\mathbf{X}_n)).$$

The added value of using ICS over MD is realized when the structure of the data is on a subspace of dimension $q < p$. In this context and with a small percentage of outliers, it is of interest to focus only on the $q$ projections that maximize a kurtosis measure. Indeed, if the number $k$ of selected Invariant Coordinates corresponds to the dimension $q$ of the subspace on which the outliers lie, then the ICS method is expected to recover the subspace of interest for outlier detection. As detailed in Tyler et al. (2009), this subspace corresponds to Fisher's discriminant subspace in case of mixtures of Gaussian distributions. Nevertheless, the main difficulty is to correctly estimate this dimension $k$.

### 3.2.2 Invariant Coordinates Selection

Depends on the combination of the scatters $\mathbf{V}_1(\mathbf{X}_n)$ and $\mathbf{V}_2(\mathbf{X}_n)$ chosen, the ICSOutlier package incorporates automated ways to select these $k$ invariant coordinates. The two approaches proposed in Archimbaud et al. (2016) are implemented here: a test based on a quasi-inferential procedure and some normality tests. Considering $\mathbf{V}_1(\mathbf{X}_n)$ be "more robust" than $\mathbf{V}_2(\mathbf{X}_n)$ and a few percentage of outliers (less than 10% is recommended), the structure of the outlierness should be contained in the first $k$ non-normal components (in the last ones if the percentage of outliers is large). Since the invariant coordinates are already ordered decreasingly according to a kurtosis measure, it is enough to test whether each component is gaussian begining by the one associated with the largest eigenvalue and stop the test procedure as soon as we find a gaussian component. So, if $k + 1$ denotes the rank of the first gaussian component, the components are sequentially tested at the adapted level $\alpha_j = \alpha/j$, for $j = 1, \ldots, k$, as in Dray (2008). Because of the sequentiality, the Bonferroni correction adjusts the significance of each test and ensures a nominal level $\alpha$.

The first approach is a Parallel Analysis (PA) based on Monte Carlo simulations of eigenvalues for Gaussian populations of the same dimension as the initial dataset. This method is common for selecting components in PCA as described in Peres-Neto et al. (2005). It is based on $mEig$ simulations of a Gaussian population and, for each $j = 1, \ldots, p$, the computation of the $1 - \alpha_j$ percentile of the $j^{th}$ eigenvalue of ICS which is considered as a cut-off for the $j^{th}$ component. Then, sequentially from $j = 1$, if the observed $j^{th}$ eigenvalue exceeds the cut-off then the $j^{th}$ invariant component is declared non-gaussian and is selected. As soon as one eigenvalue is smaller than its associated cut-off, the invariant component is considered as gaussian and the test procedure is stopped.

The second approach is directly based on usual normality tests and is applied to the invariant coordinates. In the package the user can choose out of the following five tests:

the D'Agostino test of skewness (DA), the Anscombe-Glynn (AG) test of kurtosis, the Bonett-Seier (BS) test of Geary's kurtosis, the Jarque-Bera (JB) test for normality which is based on both skewness and kurtosis measures and the Shapiro-Wilk (SW) normality test (see Yazici and Yolacan (2007) and Bonett and Seier (2002) for a complete description of each). The process of testing is still sequential with an adaptation of the level of each test. Once the first normal coordinate is found, the test procedure stops and only the non-gaussian coordinates are selected.

### 3.2.3 Measure of outlierness

Let $k$ denote the number of invariant components selected by one of the two test procedures detailed previously or by looking at the screeplot of the eigenvalues. Then, for each observation $\mathbf{x}_i$, $i = 1, \ldots, n$, its squared ICS distance is computed based on the $k$ selected invariant coordinates by:

$$ICSD^2_{\mathbf{V}_1(\mathbf{X}_n)^{-1}\mathbf{V}_2(\mathbf{X}_n)}(\mathbf{x}_i, k) = \mathbf{z}'_{i,k}\mathbf{z}_{i,k}$$

where $\mathbf{z}_{i,k} = \mathbf{B}(\mathbf{X}_n, k)(\mathbf{x}_i - \mathbf{m}_1(\mathbf{X}_n))$, for $i = 1, \ldots, n$, and the $k \times p$ matrix $\mathbf{B}(\mathbf{X}_n, k)$ contains the $k$ first rows of $\mathbf{B}(\mathbf{X}_n)$.

### 3.2.4 Outlier identification

Finally, the outliers are identified based on the comparison of their ICS distance with the expectation under the Gaussian distribution as in Archimbaud et al. (2016). The cut-off is derived from $mDist$ Monte Carlo simulations of a Gaussian population of the same dimension as the initial dataset. For each observation, the ICS distance is computed using the $k$ selected components. For a given level $\beta$, the cut-off corresponds to the average of the $1 - \beta$ percentiles of the distances over all the simulations. An observation is labeled as outlier if its ICS distance is higher than this cut-off. By default, $\beta = 5\%$.

## 3.3 Using package ICSOutlier

The main function available in ICSOutlier is `ics.outlier` which directly calls all the other functions included in the package. First it is necessary to apply the `ics2` function from ICS in order to create an object of class `ics2`. Then the `ics.outlier` function performs the three steps necessary for outlier detection using ICS in an automated way:

(i) Select the invariant coordinates which recover at best the subspace where the outliers are lying using some test procedure.

(ii) Compute the ICS distance based on the selected invariant coordinates as an outlierness measure for each observation.

(iii) Label the outliers using the cut-off value obtained through simulations.

The links between the `ics2` function from the ICS package and the functions available in the ICSOutlier package are illustrated on Figure 3.1.

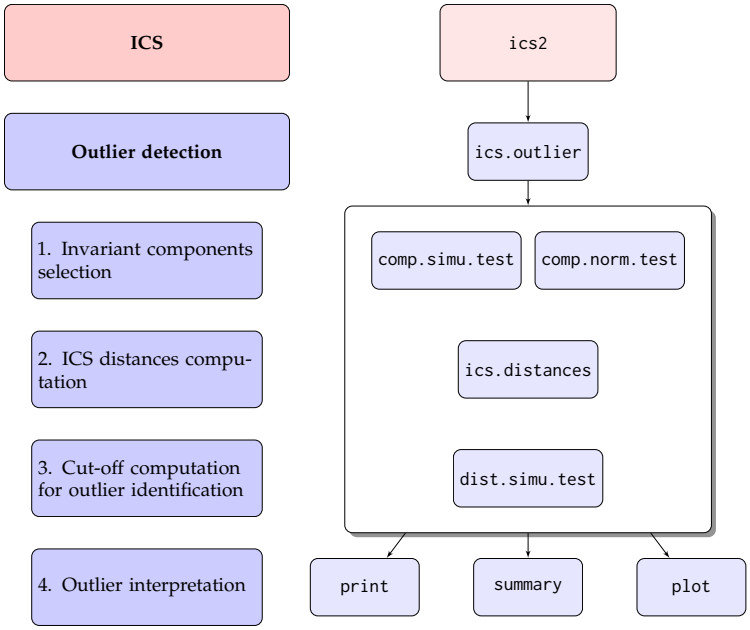



Figure 3.1 – Organization chart of the functions called by `ics.outlier`

The `ics.outlier` function only takes as parameter an object of class `ics2` and not an object of class `ics` because this later returns only uncentered invariant coordinates contrary to the `ics2` class. From a practical point a view, the `S1` and `S2` arguments of the `ics2` function should be the name of the function which returns a list with the first (resp. second) location vector $\mathbf{m}_1$ (resp. $\mathbf{m}_2$) and the first scatter matrix $\mathbf{V}_1$ (resp. $\mathbf{V}_2$).

In addition to this `ics2` object, there are other parameters of `ics.outlier` that can be tuned:

— `method`: name of the method used to select the ICS components. Options are `"simulation"` for the parallel analysis approach and `"norm.test"` for the normality tests approach. Depending on the method, either `comp.norm.test` or `comp.simu.test` is used.

— `test`: name of the marginal normality test to use if `method = "norm.test"`. Possibilities are `"jarque.test"`, `"anscombe.test"`, `"bonett.test"`, `"agostino.test"`, `"shapiro.test"`. Default is `"agostino.test"`.

— `mEig`: number of simulations performed to derive the cut-off values for selecting the ICS components. Only if `method = "simulation"`. See `comp.simu.test` for details.

— `level.test`: `level` for the `comp.norm.test` or `comp.simu.test` functions. The inital level for selecting the invariant coordinates.

— `adjust`: logical. For selecting the invariant coordinates, the level of the test can be adjusted for each component to deal with multiple testing. See `comp.norm.test` and `comp.simu.test` for details. Default is TRUE.
— `level.dist`: `level` for the `dist.simu.test` function. The (1-`level`(s))th quantile(s) used to determine the cut-off value(s) for the ICS distances.
— `mDist`: number of simulations performed to derive the cut-off value for the ICS distances. See `dist.simu.test` for details.

By default the `ics2` function performs ICS with the usual mean vector and the usual empirical variance-covariance matrix returned by `MeanCov` and the location vector based on the third moments associated to the scatter matrix based on the fourth moments returned by `Mean3Cov4`.

## 3.4 Examples

In the following examples, we illustrate first, using an artifical data set, how `ics.outlier` behaves in the case of no outlier and then apply the function to three data sets available in R. For reproducibility all examples have a fixed seed.

### 3.4.1 Example with no outlier

One of the advantages of using ICS for outlier detection is its ability to detect the absence of outliers. If the first invariant coordinate is considered normal that means there is no outlier in the data set. So, in presence of a normal multivariate data set the function should select in most cases no component and hence will not detect outliers.

For the first example we simulate a bivariate normally distributed data set with 500 observations and we determine the cut-offs for identifying the outliers at the level 2.5% for all methods.

```
library("ICSOutlier")
```

```
R> Loading required package: ICS
R> Loading required package: mvtnorm
R> Loading required package: moments
```

```
# Data simulation
set.seed(123)
X <- matrix(rnorm(1000, 0, 0.1), 500, 2)
```

Using the default ICS setting and choosing the Jarque-Bera test of kurtosis to select the components at the default level 5% can easily be done as follows.

```
icsX <- ics2(X)
icsOutlierJB <- ics.outlier(icsX, test = "jarque", level.test = 0.05,
                            level.dist = 0.025)
print(icsOutlierJB)
```

```
R> [1] "0 components were selected and no outliers were detected."
```

Hence in this outlier-free normal data no component was considered non-normal and therefore as desired no outliers were detected.

As a comparison we compute the robust Mahalanobis distances using the MCD with a breakdown point of 25% and compute the commonly used cut-off value coming from the $\chi^2_2$ distribution and the adjusted cut-off value from (Green and Martin, 2017b) as implemented in the package CerioliOutlierDetection.

```
# Robust Mahalanobis distance with MCD estimates with a breakdown point of 25%
library("robustbase")
MCD <-  covMcd(X, alpha = 0.75)
RD <- mahalanobis(X,  MCD$center, MCD$cov)

# Cut-off based on the chi-square distribution
cutoff.chi.sq <- qchisq(0.975, df = ncol(X))
cutoff.chi.sq
```

```
R> [1] 7.377759
```

```
# Cut-off based Green and Martin (2014)
library("CerioliOutlierDetection")
cutoff.GM <- hr05CutoffMvnormal(n.obs = nrow(X), p.dim = ncol(X),
                mcd.alpha = 0.75, signif.alpha = 0.025,
                method = "GM14", use.consistency.correction = TRUE)$cutoff.asy
cutoff.GM
```

```
R> [1] 14.22071
```

To visualize these results, the robust squared Mahalanobis distances are plotted with two different cut-off values in Figure 3.2.

```
colPoints <- ifelse(RD >= min(c(cutoff.chi.sq, cutoff.GM)), 1, grey(0.5))
pchPoints <- ifelse(RD >= min(c(cutoff.chi.sq, cutoff.GM)), 16, 4)
plot(seq_along(RD), RD,  pch = pchPoints, col = colPoints,
     ylim=c(0, max(RD, cutoff.chi.sq, cutoff.GM) + 2), cex.axis = 0.7,
     cex.lab = 0.7, ylab = expression(RD**2), xlab = "Observation Number")
abline(h = c(cutoff.chi.sq, cutoff.GM), lty = c("dashed", "dotted"))
legend("topleft", lty = c("dashed", "dotted"), cex = 0.7, ncol = 2, bty = "n",
       legend = c(expression(paste(chi[p]**2, " cut-off")), "GM cut-off"))
```

Hence in this example using the $\chi^2_2$ cut-off yields to 12 outliers while the more sophisticated cut-off does not classify any observation as outlying.

Another situation where the `ics.outlier` function can conclude to an absence of outliers is when the squared ICS distances are below the corresponding cut-off for all observations, as illustrated in the help file of the package: `?ics.outlier`.

### 3.4.2 HTP data set

The real data set `HTP`, introduced in Archimbaud et al. (2016), is included in this package. The data set provides the results of 88 numerical tests for 902 high-tech parts.

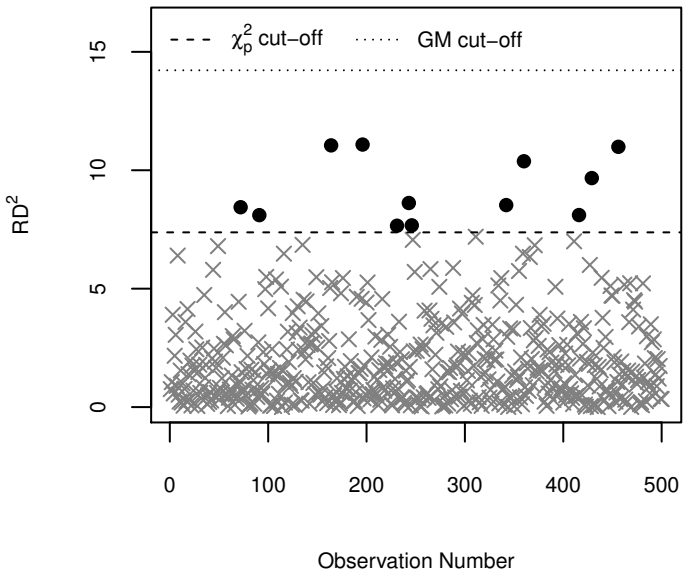



Figure 3.2 – Squared robust Mahalanobis distances and two different cut-off values.

Based on these results the producer considered all parts functional and all of them were sold. However two parts, 581 and 619, showed defects in use and were returned to the manufacturer. These two observations can thus be considered as outliers and the objective is to detect them. We use the `ics.outlier` function with its default settings, as most users initially would do.

```
# HTP dataset
library("ICSOutlier")
set.seed(123)
data(HTP)
outliers <- c(581, 619)

# default ICS
icsHTP <- ics2(HTP)

# Outlier detection with selection of components based on normality tests
# by default it can take quite long as mDist = 10000
icsOutlierDA <- ics.outlier(icsHTP)
summary(icsOutlierDA)

R>
R> ICS based on two scatter matrices and two location estimates
R> S1:  MeanCov
R> S2:  Mean3Cov4
R>
R> Searching for a small proportion of outliers
R>
R> Components selected at nominal level 0.05: 14
```

```
R> Selection method: norm.test (agostino.test)
R> Number of outliers at nominal level 0.025: 43

plot(icsOutlierDA, cex.lab = 0.7, cex.axis = 0.7)
points(outliers, icsOutlierDA@ics.distances[outliers], pch = 5)
text(outliers, icsOutlierDA@ics.distances[outliers], outliers, pos = 2,
     cex = 0.7)
```

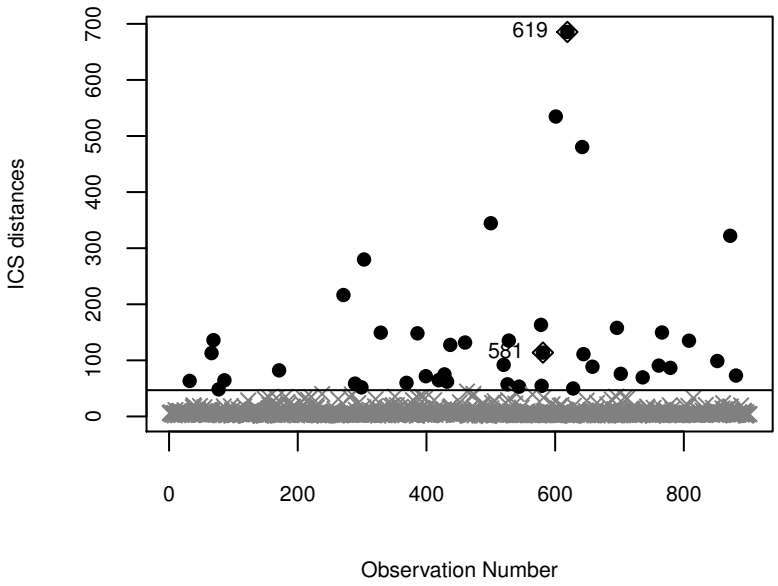



Figure 3.3 – Squared ICS distances for HTP data with default parameters.

Based on this result we are able to identify the two outliers among the 5% of the observations declared as outliers, taking into account the 14 first components selected by the D'Agostino normality test.

However, following Archimbaud et al. (2016), a simple alternative for selecting components is to use the screeplot of the `icsHTP` object.

```
screeplot(icsHTP, cex.lab = 0.7, cex.axis = 0.7, cex.names = 0.7,
          cex.main = 0.7)
```

Based on this screeplot, three components might be a reasonable choice and then the functions `ics.distances` and `dist.simu.test` can be used to obtain the desired distances and cut-off value.

```
ics.dist.scree <- ics.distances(icsHTP, index = 1:3)
ics.cutOff <- dist.simu.test(icsHTP, 1:3)
```

Before visualizing these results we also apply two competing outlier detection methods. First the Finite-Sample Reweighted MCD (FSRMCD) outlier detection test of Cerioli (Cerioli, 2010), implemented in the package CerioliOutlierDetection and secondly the SIGN1

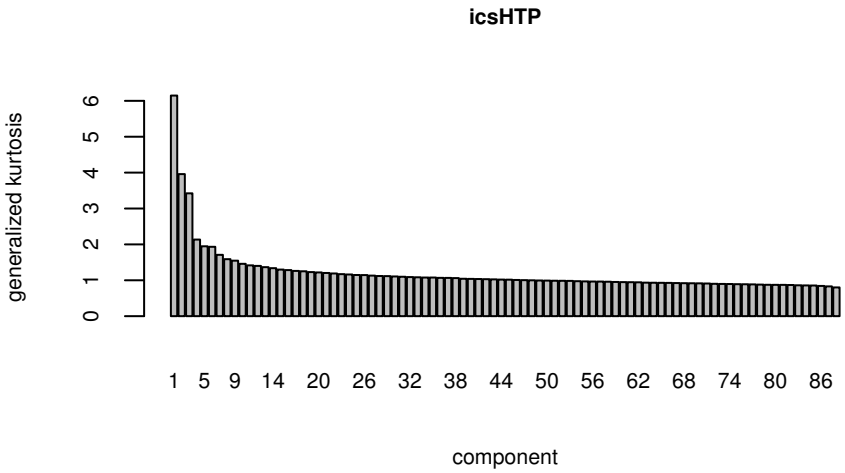



Figure 3.4 – Screeplot of ICS eigenvalues for HTP data and default parameters.

algorithm (Filzmoser et al., 2008) implemented in the package mvoutlier. For the FSRMCD algorithm, we choose as nominal level $\alpha = 1 - \gamma^{1/n}$ for the individual outlier tests, with $\gamma$ the nominal size of the intersection test and $n$ the number of observations. Then the ICS distances based on the three selected components are plotted against the distances from the two competing methods together with the corresponding cut-off values.

```
# FSRMCD
library("CerioliOutlierDetection")
FSRMCD <- cerioli2010.fsrmcd.test(HTP, mcd.alpha = 0.75,
                                  signif.alpha = 1 - 0.975**(1/nrow(HTP)))
FSRMCD.cutoff <- min(FSRMCD$mahdist.rw[which(FSRMCD$outliers == TRUE)])

# SIGN1
library("mvoutlier")
SIGN1 <- sign1(HTP, qcrit = 0.975)

par(mfrow = c(1, 2))
par(mar = c(4, 4, 2, 0.2))

# Comparison ICS vs FSRMCD
colPoints <- ifelse(ics.dist.scree >= ics.cutOff , 1, grey(0.5))
pchPoints <- ifelse(ics.dist.scree >= ics.cutOff, 16, 4)
plot(FSRMCD$mahdist.rw, ics.dist.scree, col = colPoints, pch = pchPoints,
     cex.lab = 0.7, cex.axis = 0.7, cex.main = 0.7,
     main = "ICS vs FSRMCD", ylab = "ICS Distances", xlab = "FSRMCD")
points(FSRMCD$mahdist.rw[outliers], ics.dist.scree[outliers], pch = 5)
text(FSRMCD$mahdist.rw[outliers], ics.dist.scree[outliers],
     labels = outliers, pos = 2, cex = 0.7)
abline(h = ics.cutOff, v = FSRMCD.cutoff, lty = "dashed")

# Comparison ICS vs SIGN1
plot(SIGN1$x.dist, ics.dist.scree,  col = colPoints, pch = pchPoints,
     cex.lab = 0.7, cex.axis = 0.7, cex.main = 0.7,
     main = "ICS vs SIGN1", ylab = "ICS Distances", xlab = "SIGN1")
points(SIGN1$x.dist[outliers], ics.dist.scree[outliers], pch = 5)
```

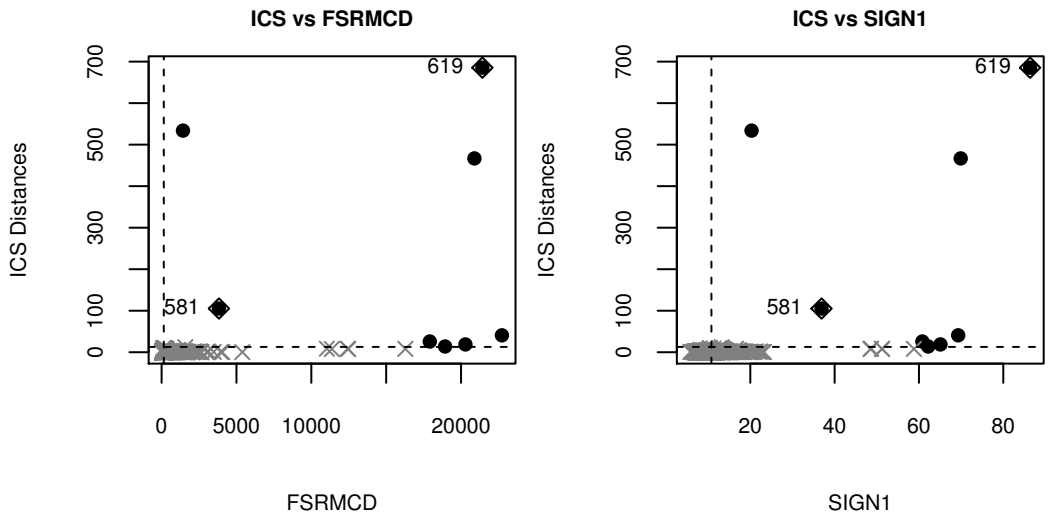



Figure 3.5 – Squared ICS distances with 3 components against FSRMCD (left panel) and SIGN1 (right panel) measures of outlierness for HTP data.

```
text(SIGN1$x.dist[outliers], ics.dist.scree[outliers],
     labels = outliers, pos = 2, cex = 0.7)
abline(h = ics.cutOff, v = SIGN1$const, lty = "dashed")
```

As illustrated in Figure 3.5, the cut-offs used for the FSRMCD and SIGN1 algorithms plotted on the x-axis lead to too many false positives. Even when focusing only on the rank of the outlying observations it is clear that the ICS method outperforms the other methods. The outlier labelled 619 is obviously outlying with all methods, but the outlier labelled 581 has the 4th highest squared ICS distance compared to 15th and 12th for the distances computed with FSRMCD and SIGN1 algorithms.

Note that Archimbaud et al. (2016) contains a more detailed analysis of this dataset comparing ICS with other methods and also gives the impact of the choice of different scatter matrices. For the present data set the default scatter combination used here seems the best.

### 3.4.3 Reliability data set

The Reliability data set comes also from an industrial context and is part the of REPPlab package. It contains 55 variables measured during the production process of 520 units. The challenge for this data is to identify the produced items with a fault not detected by the marginal tests. According to Fischer et al. (2016a), the observations numbered 414 and 512 are most likely outliers.

A special feature of this data set is that for example variable 24 is nearly constant which makes this data set hard to use with high-breakdown methods. For example the MCD cannot be computed for this data set when all variables are used. The use of ICS and other distance-based outlier detection methods is also detailed in Archimbaud et al. (2016).

Let us now illustrate the use of other scatter matrices than the default one together with the way to pass arguments to some of the functions. This time the invariant components are selected using simulations (Parallel analysis) with an initial decision level of 5%. These results are then compared to two methods not included in Archimbaud et al. (2016), namely the Local Outlier Factor (LOF) and the Angle-Based Outlier Factor (ABOD) outlier detection methods. For LOF we use the `lof` function from the Rlof package which parallelizes the computation of the local outlier factor using $l$ neighbours for each observation. The number of neighbours is determined based on the guidelines from Breunig et al. (2000): we compute the outlierness measures for $l$ between 5 and 50 and we aggregate the results by taking the maximum of the local outlier factor for each observation. Then we calculate the angle-based outlier factor for each observation through the `abod` function from the abodOutlier package on a random sample of 10% (the default) of the data (not on the entire data set because it is too time-consuming).

The location and scatter combinations we use here are the regular mean vector and covariance matrix and the joint maximum likelihood estimation of location and scatter of a t-distribution with one degree of freedom, also known as Cauchy MLE estimate. The function `MeanCov` returns the mean vector and regular covariance matrix as required for `ics2` and for the Cauchy MLE we use the function `tM` where the degree of freedom is specified using the argument `df`. As the Cauchy MLE is considered the more robust estimator it should be specified in `ics2` as `S1`.

```
# ReliabilityData example:
# the observations 414 and 512 are suspected to be outliers
library("ICSOutlier")
library("REPPlab")
```

```
R> Loading required package: rJava
R> Loading required package: lattice
R> Loading required package: LDRTools
```

```
set.seed(123)
data(ReliabilityData)
outliers <- c(414, 512)

# ICS with MLE Cauchy and the Mean-Cov
icsReliabilityData <- ics2(ReliabilityData, S1 = tM, S2 = MeanCov,
                           S1args = list(df = 1))

# Outlier detection with selection of components based on simulations
icsOutlierPA <- ics.outlier(icsReliabilityData, method = "simulation",
                            level.test = 0.05, mEig = 5000,
                            level.dist = 0.01, mDist = 5000)
icsOutlierPA
```

```
R> [1] "39 components were selected and 86 outliers were detected."
```

```
# LOF: Local Outlier Factor
library("Rlof")
```

```
R> Loading required package: doParallel
R> Loading required package: foreach
```

```
R> Loading required package: iterators
R> Loading required package: parallel

X.lof <- lof(ReliabilityData, 5:50, cores = 2)
X.lof.max <- apply(X.lof, 1, max)

# ABOD: Angle-Based Outlier Factor
library("abodOutlier")

R> Loading required package: cluster

X.abod <- abod(ReliabilityData, method = "randomized")

par(mfrow = c(1, 2))
par(mar = c(4, 4, 2, 0.2))

# Comparison ICS vs LOF
colPoints <- ifelse(icsOutlierPA@ics.distances >= icsOutlierPA@ics.dist.cutoff,
                    1, grey(0.5))
pchPoints <- ifelse(icsOutlierPA@ics.distances >= icsOutlierPA@ics.dist.cutoff,
                    16, 4)
plot(X.lof.max, icsOutlierPA@ics.distances, col = colPoints, pch = pchPoints,
     cex.lab = 0.7, cex.axis = 0.7, cex.main = 0.7,
     main = "ICS vs LOF", ylab = "ICS Distances", xlab = "LOF")
points(X.lof.max[outliers], icsOutlierPA@ics.distances[outliers], pch = 5)
text(X.lof.max[outliers], icsOutlierPA@ics.distances[outliers],
     labels = outliers, pos = 4, cex = 0.7)

# Comparison ICS vs ABOD
plot(X.abod, icsOutlierPA@ics.distances, col = colPoints, pch = pchPoints,
     cex.lab = 0.7, cex.axis = 0.7, cex.main = 0.7,
     main = "ICS vs ABOD", ylab = "ICS Distances", xlab = "ABOD")
points(X.abod[outliers], icsOutlierPA@ics.distances[outliers], pch = 5)
text(X.abod[outliers], icsOutlierPA@ics.distances[outliers],
     labels = outliers, pos = 4, cex = 0.7)
```

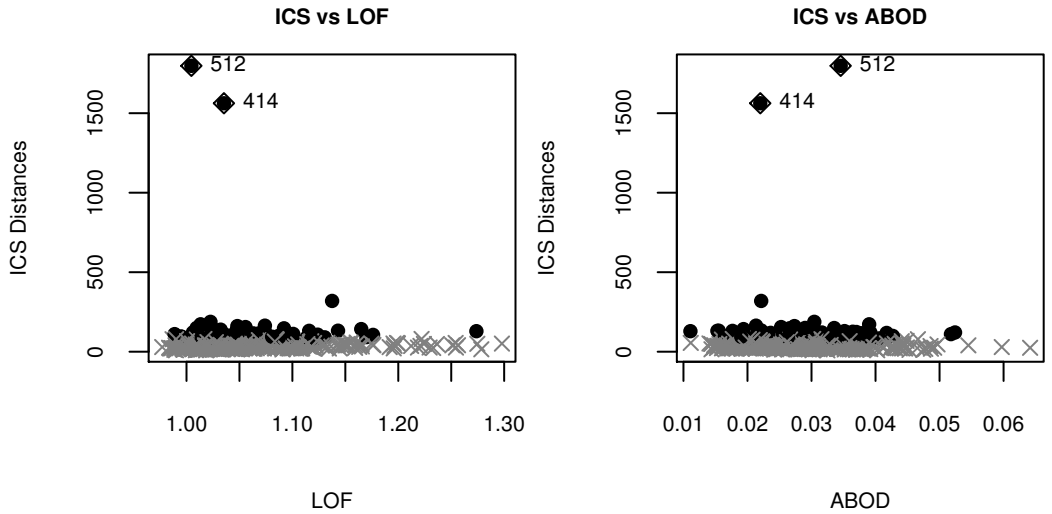



Figure 3.6 – Squared ICS distances with 39 components against LOF (left panel) and ABOD (right panel) measures of outlierness for Reliability data.

This analysis shows that ICS selects here quite many components and suggests 86 outliers at the level of 1%, represented by black points on Figure 3.6. When looking however at the distances, the observations 414 and 512 are clearly separated from the bulk of the data for the ICS method while the two observations are not detected using the other methods. As explained in Archimbaud et al. (2016), the automatic component selection may not be the best choice for this example with too many components selected. Using the screeplot on this example is again more appropriate than automatic selection.

### 3.4.4 HBK data set

The `hbk` data set is included in the robustbase package (Rousseeuw et al., 2016). It is an artificial data set created by Hawkins et al. (1984) for illustrating the so-called masking effect of outliers. It contains two groups of outliers with observations 1-10 in the first group and observations 11-14 in the second group, over the 75 observations characterized by the three explanatory variables. Usually with non-robust methods only the observations 12, 13 and 14 are identified as outliers. For this example the percentage of outliers is around 19% and so is larger than the percentage we recommend for the current version of the package. Our aim is to illustrate that in such a situation, the choice of the scatter matrices may have a large impact on the results.

First, the combination of the highly robust MCD estimates with the mean vector and regular covariance matrix is studied. Note that in order to be able to use the MCD via the function `CovMcd` a wrapper needs to be written around the function so that it returns the proper list as expected by `ics2`. Using then the D'Agostino test for skewness for selecting the invariant components leads to select two components at the default level of 5%. The level of the quantile used for deriving the cut-off for outlier identification is the default, 2.5%. For this combination, ICS performs equally well as the MCD-based Mahalanobis distances and as ROBPCA, as illustrated in the first three plots of the Figure 3.7. All outliers are correctly identified, no false positive is detected and the two groups of outliers are clearly separated. However it is important to note that for ROBPCA it is necessary to compare two distances (SD and OD) instead of only one for the ICS and Mahalanobis distances.

Next, we consider the combination of two non-robust location vectors and scatter matrices which are the default in ICS, with the same parameters for selecting the components and for identifying the outliers as previously. The results are presented on the fourth plot of Figure 3.7.

```
library("rrcov")
set.seed(123)
# HBK data set
data(hbk)

# ICS with MCD estimates and the usual estimates
# Need to create a wrapper for the CovMcd function to return
# first the location estimate and the scatter estimate secondly.
```

```
myMCD <- function(x,...){
  mcd <- CovMcd(x,...)
  return(list(location = mcd@center, scatter = mcd@cov))
}
icsHBK_mcd <- ics2(hbk[, 1:3], S1 = myMCD, S2 = MeanCov,
                   S1args = list(alpha = 0.75))

# Outlier detection with selection of components based on
# the D'Agostino test for skewness
icsOutlierDA.MCD <- ics.outlier(icsHBK_mcd, mEig = 5000,
                                level.dist = 0.025, mDist = 5000)
icsOutlierDA.MCD

R> [1] "2 components were selected and 14 outliers were detected."

# Robust Mahalanobis distance with MCD estimates
# with a breakdown point of 25%
MCD <-  covMcd(hbk[, 1:3], alpha = 0.75)
RD <- mahalanobis(hbk[, 1:3],  MCD$center, MCD$cov)
cutoff.chi.sq <- qchisq(0.975, df = ncol(hbk[, 1:3]))

# ROBPCA with two components
robpca <- PcaHubert(hbk[,1:3], k = 2, crit.pca.distances = 0.975)

# ICS with non robust estimates
icsHBK <- ics2(hbk[,1:3], S1 = MeanCov, S2 = Mean3Cov4)

# Outlier detection with selection of components based on
# the D'Agostino test for skewness
icsOutlierDA <- ics.outlier(icsHBK, level.dist = 0.025)
icsOutlierDA

R> [1] "2 components were selected and 2 outliers were detected."

# Visualization of the results
par(mfrow = c(2, 2))
par(mar = c(4, 4, 2, 0.2))

# Robust Mahalanobis distance with MCD estimates
colPoints <- ifelse(RD >= cutoff.chi.sq, 1, grey(0.5))
pchPoints <- ifelse(RD >= cutoff.chi.sq, 16, 4)
plot(RD, col = colPoints, pch = pchPoints,
     cex.lab = 0.7, cex.axis = 0.7, main = "Robust MD", cex.main = 0.7)
abline(h = cutoff.chi.sq)

# ICS with MCD estimates and regular covariance
plot(icsOutlierDA.MCD,  cex.lab = 0.7, cex.axis = 0.7,
     main = "ICS MCD-COV", cex.main = 0.7)

# ROBPCA
colPoints <- ifelse(robpca@flag==FALSE, 1, grey(0.5))
pchPoints <- ifelse(robpca@flag==FALSE, 16, 4)
plot(robpca, col = colPoints, pch = pchPoints,
     cex.lab = 0.7, cex.axis = 0.7, cex.main = 0.6)

# ICS with default non-robust estimates
plot(icsOutlierDA, cex.lab = 0.7, cex.axis = 0.7, main = "ICS COV-COV4",
     cex.main = 0.7)
```

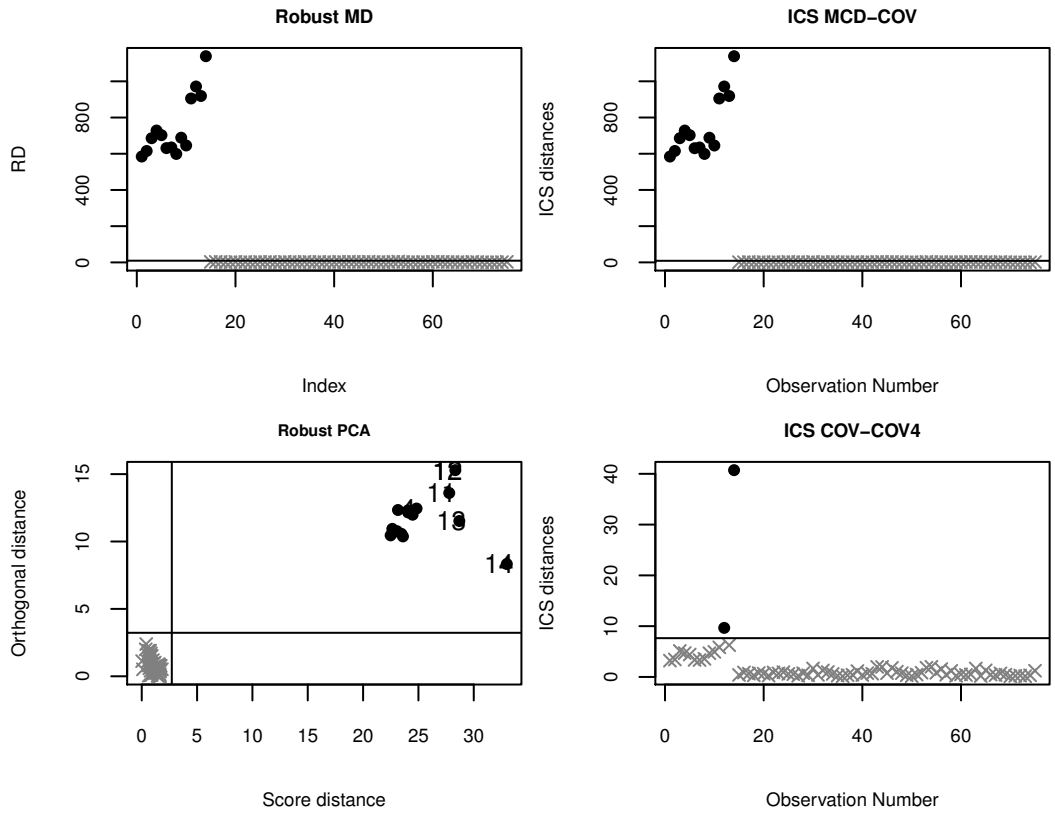



Figure 3.7 – (1st row, 1st column) Squared robust Mahalanobis distance based on the MCD , (1st row, 2nd column) Squared ICS distances with MCD and Mean-Cov, (2nd row, 1st column) ROBPCA with two components, (2nd row, 2nd column) Squared ICS distances with default scatters.

Here the default `ics2` scatter combination does not perform well. Only three outliers from the second group are detected, although the distance plot indicates three distance levels but the cut-off value is not good and the separation not clear especially when compared with the MCD-COV combination. However, when looking at the invariant coordinates on Figure 3.8,

```
plot(icsHBK, col = rep(3:1, c(10, 4, 61)))
```

we can see that the default scatter combination reveals the 14 outliers and the masking effect appears only when computing the ICS distances.

## 3.5 Conclusion and future developments

The efficiency of ICSOutlier for outlier detection has been illustrated on several examples and different situations. The main advantage of the package is that it provides a completely automated method of outlier detection to the user. By using the default values, the function `ics.outlier` is very easy to use even by non-statisticans with only a few parameters to tune. As illustrated on several examples, the results are already pretty good with the default parameters as soon as the percentage of outliers is small. Note

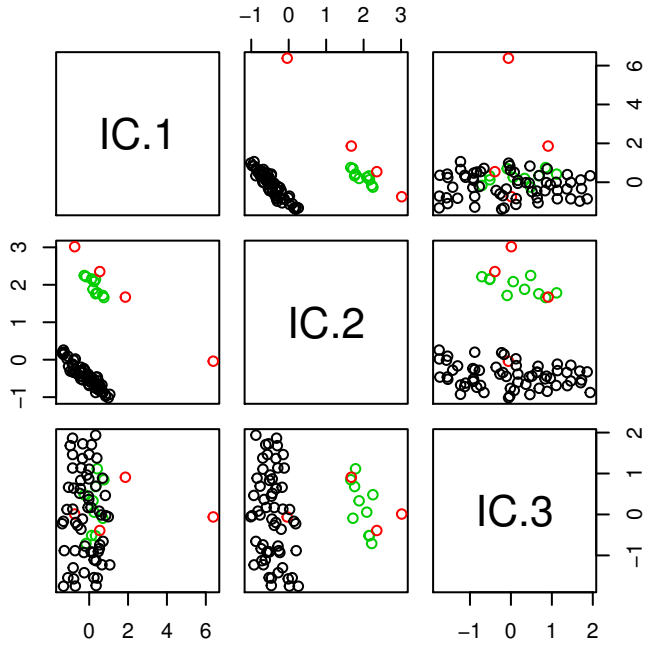



Figure 3.8 – Scatter plot of the invariant coordinates of the hbk data for the default scatter combinations. The two outlier groups are marked with different colors.

however that the components selection procedure can be improved sometimes by a visual inspection of the screeplot and the cut-off for outlier identification can be adjusted by looking at the ICS distances plot. The ability of the method to check for the absence of outliers is an advantage compared with methods such as the Mahalanobis distance. In the same vein, the number of false positives, i.e. the number of non-outliers declared as outliers, is restricted thanks to the use of the cut-off derived from simulations. Moreover, the real HTP data set included in the package shows that the ICS method is useful for outlier detection in the context of quality control, among possible applications. Finally the ICS method competes with common approaches based on distances (the Mahalanobis distance, its robust version, the PCA methods, its variants and improvements) as well as other methods based on the density (LOF) or on angles (ABOD).

The computation of cut-offs for the selection of the non-gaussian invariant components, on the one hand, and for the threshold for labeling the outliers, on the other hand, are time consuming. This can be a serious drawback for the use of the function on a large data sets, so a solution based on interpolations is under preparation. Moreover, for now, the function can only select the first invariant components which is relevant when the percentage of outliers is small. For a large percentage, the invariant components associated with the smallest eigenvalues may also be of interest and this is a current perspective of the present package.

## 3.6 Complements on Chapter 3: the ICSShiny package

Following the development of the ICSOutlier package, we decide to create a shiny application, the ICSShiny package, to perform ICS and especially ICS outlier identification in a user-friendly way. This new tool for exploring a multivariate data was developed in collaboration with an intern (May, 2017) and it is inspired from the Factoshiny application of the FactoMineR package dedicated to factorial analysis. The following explanations are a reprint of the help manual.

The ICSShiny package performs ICS via a shiny app where the user can change the scatter matrices, explore the output and download graphs and components. Also the ICS outlier detection framework, from the ICSOutlier package is available. The function `ICSShiny` returns several tabs on the navigator:

— **Choice of the parameters**
The scatterplot matrix of an ICS object for the parameters chosen on the left part (variables included/excluded, location vectors and scatter matrices).

— **Component selection**
Three different subtabs to help the user to choose the interesting components. The first sub-tab is the screeplot of the eigenvalues of the ICS object followed by the summary of the analysis. The second sub-tab plots the kernel density of the ICS components. The third sub-tab suggests which components to select, starting from the highest and/or the lowest kurtosis, through different normality tests or simulations.
The default values of the slidebar in the left are obtained from `agostino.test` at 5%.

— **Matrix scatterplot of invariant components**
The two sub-tabs aim at identifying groups or outliers by using pairwise plots of invariant coordinates. It offers two ways of plotting them: only two invariant components or a scatterplot matrix with up to six invariant components. The left panel allows to color the groups identified by the user and label the observations.

— **Outlier identification**
This tab plots outlierness values for each observation based on the selected components. These squared ICS distances are computed through the `ics.distances` function as the Euclidian distance of the observations to the origin using the selected centered components. The identification of the outliers can be based on different cut-offs: from Monte Carlo simulations as in `dist.simu.test` or by giving a percentage or a number of observations to identify.

— **Descriptive statistics**
This tab gives some descriptive statistics on different subsets of the data (for all the observations, for the observations from a given cluster, for the outlying observations) and enables to compare the sub-populations. The application includes a

boxplot, a kernel density, an histogram and some basic statistics: Min, Q1, Mean, Median, Q3 and Max.

— **Data Table**

This tab contains the dataset with a nice display and the possibility to choose different sub-populations of the data: all the observations, the observations from a given cluster or the outlying observations.

— **Save**

This tab allows to display and save the data table of components and the summary of operations. The data frame contains the components kept in the analysis as well as the distance generated by these components. It also includes the cluster the observation belongs to whether the observation is defined as an outlier, as well as the variables used for labelling and categorizing the data. The data are saved in a csv format. The summary of operations contains a summary of all parameters that were used to obtain the current result, it may be useful for another user who may want to get the same result as the original user. It is saved in a txt format.

The "Close the session" button closes the application and saves the `icsshiny` object into the global environment.

# Chapter 4

# ICS with positive semi-definite scatter matrices for data not in general position

*This chapter focuses on a way to adapt the ICS method to the case when the scatter matrices are not necessarily positive definite. Indeed, in the classical presentation of ICS, Tyler et al. (2009) define a scatter matrix as a symmetric positive definite matrix, equivariant under affine transformations. However, if the data contain collinear variables, or if the number of dimensions p is greater than the number of observations n, then even the empirical variance-covariance matrix becomes singular. This case is therefore worth considering as these data are typical in the context of semiconductors for the automotive industry and characteristic of the aerospace integrated circuits. In this context, Tyler (2010) demonstrates that, as long as the data is in general position, any affine equivariant scatter matrices is proportional to the variance-covariance matrix. So, the ICS method cannot reveal the outlierness structure as it diagonalizes a matrix proportional to the identity. The paradigm taken in this chapter is however that the data is not in general position. In our industrial context, observations are usually lying on a subspace of smaller dimension, which ensures the assumption to be realistic. For generalizing the ICS method to singular estimates, three approaches are proposed: computing the generalized inverse of one of the scatter matrices, pre-processing the data by a reduction of dimension or performing a Generalized Singular Value Decomposition (GSVD). These solutions are investigated from a theoretical point of view under various aspects. Is the criterion of the method modified? Are the scores still affine invariant? Do the scatter estimators play a symmetric role? It occurs that the nice properties of the classical ICS only remain valid for the method based on the generalized singular value decomposition. The affine invariance property only holds for affine equivariant semi-definite positive scatter matrices. But, in high dimension or in presence of collinear features, the most common scatter estimates are only orthogonally equivariant, except for the variance-covariance matrix. A theoretical example illustrates the first part while a real case focuses on the GSVD method.*

## Sommaire


**4.1 Introduction . . . . . 103**

**4.2 The classical ICS method . . . . . 104**

4.2.1 Principle . . . . . 105

4.2.2 Interpretation . . . . . 105

4.2.3 Properties . . . . . 106

4.2.4 Other statistical methods solving a GEP of two scatter matrices 107

**4.3 ICS with semi-definite positive scatter matrices . . . . . 109**

4.3.1 Challenges with singular scatter matrices: an example . . . . . 110

4.3.2 ICS with a Moore-Penrose pseudo-inverse . . . . . 112

4.3.3 A dimension reduction as preprocessing . . . . . 115

4.3.4 ICS with a Generalized Singular Value Decomposition . . . . . 119

**4.4 GSVD on a practical case: an industrial example with collinearity . . . . . 125**

4.4.1 Context . . . . . 126

4.4.2 Computation of the two scatter matrices COV and $COV_4$ . . . . 126

4.4.3 Implementation of the GSVD procedure . . . . . 126

4.4.4 Some results . . . . . 127

**4.5 Conclusion and perspectives . . . . . 129**

## 4.1 Introduction

The Invariant Coordinate Selection (ICS) method is a powerful tool for multivariate outlier detection, as explained in Chapter 2. Its affine invariance property comes from the simultaneous diagonalization of two scatter matrices that are affine equivariant and positive definite. In this chapter, we focus on the case where at least one of these scatter matrices is singular. From our experience, such a case is becoming a common issue. Indeed, the variance-covariance matrix can be semi-definite positive and so can be singular if: (i) some of the initial variables are collinear or (ii) the number of variables exceeds the number of observations. The collinearity is a long-standing issue, that can be (not ideally) dealt by deleting some of the collinear features, taking the risk of potentially losing information. The second reason is a more recent issue emerging with the exponential increase of measurements, leading to more features than observations. This latter case is referred in the literature as the High Dimension/Small Sample Size (HDLSS) situation.

Performing the ICS method in this context is very challenging because of three main issues. First, as stated by Tyler (2010), many of the well-known affine equivariant scatter statistics, such as the M-estimators (Maronna, 1976) or the MCD (Rousseeuw, 1986), are not defined in this case. Second, for those which can be defined such as the variance-covariance matrix or the projection based estimators (Donoho and Gasko, 1992; Maronna et al., 1992; Tyler, 1994), they are only guaranteed to be positive semi-definite. Note that the Stahel-Donoho estimator (Stahel, 1981; Donoho, 1982), presented in Section 4.3.3, is a particular case of the projection based estimators (Tyler, 1994). Third, Tyler (2010) shows that if the data is in general position [1], then all affine equivariant scatter statistics, symmetric in the observations [2], are proportional to the variance-covariance matrix. So, in this case, the ICS method cannot reveal the outlierness structure as it diagonalizes a matrix proportional to the identity.

The paradigm taken in this chapter is however that the data is not in general position. Concretely, when the number of observations $n$ is larger than the number of variables $p$, this phenomenon happens as soon as some of the variables are perfectly multicollinear. From our experience in an industrial context, this problem of multicollinearity is very frequent (see Section 5.3.1 for details). In the HDLSS case, if the $n$ observations span a space with dimension $n-1$, they are in general position. But if the dimension is smaller than $n-1$, they are not in general position. If $n << p$, this last situation is quite unlikely to appear but it happens sometimes and, in such a situation, ICS may be useful. To be more precise, from our experience, it happens more frequently in the automotive field than in the spatial industry (see Chapter 5 for details).

In addition, in the classical presentation of ICS, Tyler et al. (2009) define the scatter matrices as symmetric positive definite matrices, equivariant by affine transformation. So,

1. Data is in general position if there is no subset of $k$ observations lying on a subspace of dimension $k-2$, with $k \leq p+1$ and $p$ denotes the number of variables.

2. i.e. $\mathbf{V}(\mathbf{Q}\mathbf{X}_n) = \mathbf{V}(\mathbf{X}_n)$ for any permutation matrix $\mathbf{Q}$ of order $n$ and $\mathbf{X}_n \in \Re^{n \times p}$, the initial data containing $n$ observations, characterized by $p$ variables.

it becomes necessary to adapt the definition for the case of singular scatter matrices. The non-singularity of scatter matrices is needed for many statistical methods. As mentioned by Branco and Pires (2015), "Statistical methods are not prepared to work directly with data where $p \geq n - 1$". In order to circumvent the problem, several adaptations exist. A simple idea is to perform a variable selection if many variables are known to be non-pertinent. However, this procedure can lead to delete a substantial amount of variables to obtain the convenient number of dimensions on which scatter estimators can be defined. Another way of handling this situation is to regularize the scatter estimates (Ollila and Tyler, 2014; Verbanck et al., 2015). For example, as already mentioned in Section 1.4.2, Ro et al. (2015) assume an easier data structure and analyze a diagonal estimator. In this section, dedicated to the outlier detection methods in the HDLSS case, it is also proposed to pre-process the data by a reduction of dimension. The goal of this procedure is to select a subspace spanned by $k$ linear combinations of the initial variables. The choice of $k$ is however a difficult task.

All the previous propositions can also be applied for adapting the ICS method. However, as we focus only on unsupervised methods, we have no a priori information on the variables and so the selection is not an easy task. Moreover, regularized scatter estimators are not affine equivariant. Thus, instead of investigating procedures as variable selection or regularization, we rather propose to adapt ICS which considers the structure of correlation between variables. It means that we have to assume that the data is not in general position. First, we recall the definition and the properties of the classical ICS method. Then, we propose three different methods to adapting ICS to the case of semi-definite positive scatter estimates and we investigate theoretically their properties in terms of: (i) the criterion to optimize, (ii) the affine invariance of the scores and (iii) the symmetry of the roles of the two estimates. Finally, we applied the method we consider as the best to a real world example.

## 4.2 The classical ICS method

Let us first recall the ICS method using a functional approach. Let $\mathbf{X}$ be a $p$-multivariate random vector $\mathbf{X} \in \Re^p$, $F_{\mathbf{X}}$ its cumulative distribution function and $\mathbf{m}(F_{\mathbf{X}})$ an affine equivariant location estimator. Let $\mathcal{P}_p$ be the set of all symmetric positive definite matrices of order $p$, $\mathcal{SP}_p$ be the set of all symmetric positive semi-definite matrices of order $p$. Suppose that both $\mathbf{V}_1(F_{\mathbf{X}}) \in \mathcal{P}_p$ and $\mathbf{V}_2(F_{\mathbf{X}}) \in \mathcal{P}_p$ are scatter functionals, uniquely defined at $F_{\mathbf{X}}$. A scatter functional is defined as a matrix $\mathbf{V}(F_{\mathbf{X}}) \in \mathcal{P}_p$ which is affine equivariant in the sense that: $\mathbf{V}(F_{\mathbf{AX}+\boldsymbol{\gamma}}) = \mathbf{AV}(F_{\mathbf{X}})\mathbf{A}'$, for all $p \times p$ non-singular matrices $\mathbf{A}$ and all $\boldsymbol{\gamma} \in \Re^p$. For convenience, the dependence on $F_{\mathbf{X}}$ is dropped from the different scatter functionals when the context is obvious.

### 4.2.1 Principle

The Invariant Co-ordinate Selection (ICS) method (Tyler et al., 2009) compares two scatter matrices $\mathbf{V}_1$, $\mathbf{V}_2$ in order to highlight some interesting multivariate structure from the random vector $\mathbf{X}$. More specifically, the method finds the $p \times p$ matrix $\mathbf{B} = (\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_p)'$ of eigenvectors, which simultaneously diagonalizes two scatter matrices $\mathbf{V}_1 \in \mathcal{P}_p$ and $\mathbf{V}_2 \in \mathcal{P}_p$. Typically, this simultaneous diagonalization corresponds to a generalized eigenvalue problem (GEP) or solving a linear matrix pencil, which can be simplified to a non-symmetric eigenvalue problem (EVP) by diagonalizing one scatter relatively to the other. To fix the order of the components and their normalization, we choose as the standard definition of the diagonalization of $\mathbf{V}_2$ relative to $\mathbf{V}_1$:

$$\mathbf{B}\mathbf{V}_1\mathbf{B}' = \mathbf{I}_p \quad \text{and} \quad \mathbf{B}\mathbf{V}_2\mathbf{B}' = \mathbf{D}, \tag{4.1}$$

where $\mathbf{D}$ is a diagonal matrix with decreasing diagonal elements $\rho_1 \geq \cdots \geq \rho_p > 0$, which correspond to the eigenvalues of $\mathbf{V}_1^{-1}\mathbf{V}_2$ and $\mathbf{B}$ contains the corresponding eigenvectors as its rows. This problem can be re-written as:

$$\mathbf{V}_2\mathbf{b}_i = \rho_i\mathbf{V}_1\mathbf{b}_i \Leftrightarrow \mathbf{V}_1^{-1}\mathbf{V}_2\mathbf{b}_i = \rho_i\mathbf{b}_i, \text{ for } i = 1, \ldots, p. \tag{4.2}$$

with the following normalization:

— $\mathbf{b}_i'\mathbf{V}_1\mathbf{b}_j = 0$ for $i \neq j$ and $\mathbf{b}_j'\mathbf{V}_1\mathbf{b}_j = 1$ for $i = j$, with $i, j = 1, ..., p$.

— $\mathbf{b}_i'\mathbf{V}_2\mathbf{b}_j = 0$ for $i \neq j$ and $\mathbf{b}_j'\mathbf{V}_2\mathbf{b}_j = \rho_j$ for $i = j$, with $i, j = 1, ..., p$.

Equivalently, as stated in Tyler et al. (2009), the eigenvalues $\rho_i$, for $i = 1, \ldots, p$ and the eigenvectors $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_p$ can also be sequentially defined by solving the successive maximization or minimization problems of the ratio

$$\mathcal{K}(\mathbf{b}) = \frac{\mathbf{b}'\mathbf{V}_2\mathbf{b}}{\mathbf{b}'\mathbf{V}_1\mathbf{b}} \tag{4.3}$$

where $\rho_1$ is the maximal possible value of $\mathcal{K}(\mathbf{b})$ over $\mathbf{b} \in \Re^p$ which is achieved in the direction of the eigenvector $\mathbf{b}_1$. Similarly, the minimal value of $\mathcal{K}(\mathbf{b})$ is $\rho_p$, which is achieved in the direction of $\mathbf{b}_p$. More generally, as stated in the equations (15) and (16) in Tyler et al. (2009), we have:

$$\sup\{\mathcal{K}(\mathbf{b}) | \mathbf{b} \in \Re^p, \mathbf{b}'\mathbf{V}_1\mathbf{b}_j, j = 1, \ldots, m-1\} = \rho_m$$

with the supremum being obtained at $\mathbf{b}_m$, and

$$\inf\{\mathcal{K}(\mathbf{b}) | \mathbf{b} \in \Re^p, \mathbf{b}'\mathbf{V}_1\mathbf{b}_j, j = m+1, \ldots, p\} = \rho_m$$

with the infimum being obtained at $\mathbf{b}_m$.

### 4.2.2 Interpretation

As $\mathcal{K}(\mathbf{b})^2$ can be viewed as a generalized measure of kurtosis, it follows that $\rho_i$ has a useful interpretation. Thus analyzing the new co-ordinates $\mathbf{Z} = \mathbf{B}(\mathbf{X} - \mathbf{m}(F_\mathbf{X}))$, resulting from the projection of the data, centered with respect to an affine equivariant location

estimator $\mathbf{m}(F_{\mathbf{X}})$, into the eigenspace, can reveal interesting structures. Indeed, the first new co-ordinate is the one with the maximal value of a generalized kurtosis measure and so, if some outlying observations are present, they should be revealed. Likewise, the last new co-ordinate is the one with the minimal value, so if some observations form some approximate equal size clusters, they should be highlighted here (see Darlington (1970)).

More specifically, Tyler et al. (2009) show in Theorem 4 that under a mixture of $k$ elliptically symmetric distributions with possibly different location and scatter parameters, the structure of the data can be highlighted by the span of some subset of the first and/or last invariant co-ordinates. Their Theorem 3 restricts the property to the case of two groups, and then only the first or the last component is of interest. These two properties imply that a subset of the first and/or last invariant components is expected to recover the Fisher's discriminant subspace under some elliptical mixture models, even if the class identifications are unknown and for any pair of scatter matrices $\mathbf{V}_1$ and $\mathbf{V}_2$. Although these properties are restricted to elliptically symmetric distributions, they advocate in favor of analyzing these first and/or last new co-ordinates $\mathbf{Z}$ for highlighting some interesting structure from the data. In the Chapter 2, Section 2.8.2, some conditions, in favor of maximizing or minimizing the ICS ratio (4.3) with $\mathbf{V}_1 = \text{COV}$ and $\mathbf{V}_2 = \text{COV}_4$, are derived for different models.

### 4.2.3 Properties

It is important to note that the new scores $\mathbf{Z}$ are called the invariant components (IC) because of their affine invariance properties proposed by Tyler et al. (2009) and recalled below. Note that the theorems have been slightly adapted to the case where we center the components. The proofs are derived in Tyler et al. (2009) in Appendix A.1.

**Property 1.** ***Affine invariance with distinct roots****, Tyler et al. (2009), Theorem 1. If the roots $\rho_1, \ldots, \rho_p$ are all distinct, then for the affine transformation $\mathbf{X}^* = \mathbf{AX} + \boldsymbol{\gamma}$, with $\mathbf{A}$ being non-singular and $\boldsymbol{\gamma} \in \Re^p$, $\mathbf{Z}^* = \mathbf{B}(F_{\mathbf{X}^*})(\mathbf{X}^* - \mathbf{m}(F_{\mathbf{X}^*}))$ and $\mathbf{Z} = \mathbf{B}(F_{\mathbf{X}})(\mathbf{X} - \mathbf{m}(F_{\mathbf{X}}))$, then*

$$\mathbf{Z}^* = \mathbf{JZ}$$

*where $\mathbf{J}$ is a $p \times p$ diagonal matrix with diagonal elements $\pm 1$, which means the invariant coordinates $\mathbf{Z}^*$ and $\mathbf{Z}$ change at most their signs.*

**Property 2.** ***Affine invariance with multiple roots****, Tyler et al. (2009), Theorem 2. If the roots $\rho_1, \ldots, \rho_p$ consist of $m$ distinct values, $\rho_{(1)}, \ldots, \rho_{(m)}$, with $\rho_{(l)}$ having multiplicity $p_l$ for $l = 1, \ldots, m$. Then for the affine transformation $\mathbf{X}^* = \mathbf{AX} + \boldsymbol{\gamma}$, with $\mathbf{A}$ being non-singular, $\boldsymbol{\gamma} \in \Re^p$ and having partitioned $\mathbf{Z}' = (\mathbf{Z}'_{(1)}, \ldots, \mathbf{Z}'_{(m)})$, where $\mathbf{Z}_{(l)} \in \Re^{p_l}$, the components $\mathbf{Z}_{(l)} = \mathbf{B}(F_{\mathbf{X}})\mathbf{X}$ and $\mathbf{Z}^*_{(l)} = \mathbf{B}(F_{\mathbf{X}^*})\mathbf{X}^*$ span the same space. More formally, for some non-singular $\mathbf{C}_{(l)}$ of dimension $p_l \times p_l$,*

$$\mathbf{Z}^*_{(l)} = \mathbf{C}_{(l)}\mathbf{Z}_{(l)}, \quad \text{for } l = 1, \ldots, m$$

These affine invariance properties are attractive. Contrarily to the principal components in PCA, the invariant components remain the same after standardization for example. Indeed, the PCA method is only orthogonal invariant leading to different results depending on the scaling of the data.

Finally, we recall that the Euclidian norm of an observation using all invariant components leads to the Mahalanobis distance of this observation to the location estimator in the sense of $\mathbf{V}_1$ with the normalization (4.1).

**Property 3.** ***Link with the Mahalanobis distance****, Chapter 2, Proposition 2.*
*Let us consider an affine equivariant location estimator* $\mathbf{m}$ *and two scatter matrices* $\mathbf{V}_1$ *and* $\mathbf{V}_2$*. The Euclidian norm of an observation using its invariant coordinates corresponds to the Mahalanobis distance of this observation from* $\mathbf{m}$ *in the sense of* $\mathbf{V}_1$*. Formally, it means that for observation* $i = 1, \ldots, n$*,*

$$\mathbf{z}_i'\mathbf{z}_i = (\mathbf{x}_i - \mathbf{m})'\mathbf{V}_1^{-1}(\mathbf{x}_i - \mathbf{m})$$

*Proof.* With the chosen standardization (4.1) we have $\mathbf{V}_1^{-1} = \mathbf{B}'\mathbf{B}$, so the proof is immediate. □

As noticed at the end of the Section 2.3.2, the normalization (4.1) is important and has some consequences. In this case, if we exchange the roles of $\mathbf{V}_1$ and $\mathbf{V}_2$, then the eigenvalues are the inverse of the others, the eigenvectors are in reverse order and more importantly the scale of the invariant coordinates are not the same. More specifically, it implies that the Euclidian distance using all invariant components leads to the Mahalanobis distance in the sense of $\mathbf{V}_2$. So, the choice of the order of $\mathbf{V}_1$ and $\mathbf{V}_2$ can be based on this expected property: recovering the Mahalanobis distance in the sense of $\mathbf{V}_1$ or in the sense of $\mathbf{V}_2$ when all components are taken into account. In Chapter 2 and in Alashwali and Kent (2016), $\mathbf{V}_1$ is taken as the more "robust" scatter matrix by convention, but the previous property gives an argument in favor of this choice.

All these properties corroborate the idea that ICS is a really powerful method for exploring the structure of a data set. Indeed, if the interesting subspace is of lower dimension then analyzing only the first and/or last components is satisfactory. On the other hand, even if the full space is of interest, the ICS method is also able to recover the Mahalanobis distance (in a robust way or not). So, solving a Generalized Eigenvalue Problem (GEP) of two scatter matrices is an efficient tool for statistical purposes.

### 4.2.4 Other statistical methods solving a GEP of two scatter matrices

In fact, several other well-known statistical methods are also solving a GEP of two scatter matrices, mostly for linear dimension reduction purposes. Two unsupervised methods are reviewed below: Principal Component Analysis (PCA), already presented in Section 1.2.3, and Linear Discriminant Analysis (LDA).

### PCA: Principal Component Analysis

Unlike the standard definition given in Jolliffe (2002) for example, the PCA method can also be defined as a diagonalization of two scatter matrices, as detailed in Liski et al. (2014) and Virta (2014). In fact, PCA diagonalizes the empirical covariance matrix or the correlation matrix, if the data is standardized first. This problem is equivalent to solve the GEP of $\mathbf{V}_1 = \mathbf{I}$ and $\mathbf{V}_2 = \text{COV}$ or the GEP of $\mathbf{V}_1 = \text{diag(COV)}$ and $\mathbf{V}_2 = \text{COV}$ in the second case. Actually, $\mathbf{I}$ and diag(COV) are not "true" scatter matrices as they are not affine equivariant. And so, PCA is not exactly a particular case of ICS as it fails the affine invariance property.

Note that ICS can be interpreted as a PCA on data whitened with respect to $\mathbf{V}_1$. Indeed, the initial $\mathbf{X}$ are transformed to $\mathbf{X}^* = \mathbf{V}_1(F_{\mathbf{X}})^{-1/2}\mathbf{X}$ and PCA diagonalizes $\mathbf{V}_2(F^*_{\mathbf{X}}) = \mathbf{V}_1(F_{\mathbf{X}})^{-1/2}\mathbf{V}_2(F_{\mathbf{X}})\mathbf{V}_1(F_{\mathbf{X}})^{-1/2}$ which is the usual symmetric transformed eigenvalue problem (EVP) of $\mathbf{V}_1(F_{\mathbf{X}})^{-1}\mathbf{V}_2(F_{\mathbf{X}})$. In terms of interpretation, if $\mathbf{V}_1$ is chosen as a robust estimate of the covariance matrix, then the method performs a PCA on robustly whitened data.

### LDA: Linear Discriminant Analysis

The well-known Linear Discriminant Analysis maximizes the separability between groups based on the between and the within-class covariance matrices $\mathbf{\Sigma_B}(\mathbf{X}_n)$ and $\mathbf{\Sigma_W}(\mathbf{X}_n)$, as originally presented by Fisher (1936). The within-group covariance matrix can be defined as $\mathbf{\Sigma_W} = \frac{1}{n}(\mathbf{X}_n - \mathbf{G}_K\mathbf{M}'_K)'(\mathbf{X}_n - \mathbf{G}_K\mathbf{M}'_K)$ and the between-group covariance matrix as $\mathbf{\Sigma_B} = \frac{1}{n}(\mathbf{G}_K\mathbf{M}'_K - \mathbf{1}_n\bar{\boldsymbol{\mu}}'_n)'(\mathbf{G}_K\mathbf{M}'_K - \mathbf{1}_n\bar{\boldsymbol{\mu}}'_n)$, where $\mathbf{G}_K$ is a $n \times K$ matrix with $\mathbf{G}_{i,k}$ being an indicator of whether the observation $i$ is in class $k$, $\mathbf{M}_K$ is a $p \times K$ matrix s.t $\mathbf{M}_K = (\bar{\boldsymbol{\mu}}_{n,1} \dots \bar{\boldsymbol{\mu}}_{n,k} \dots \bar{\boldsymbol{\mu}}_{n,K})$ with $\bar{\boldsymbol{\mu}}_{n,k}$ be the mean $p$-vector of the observations inside the class $k$, $\mathbf{1}_n$ a $n$-vector of ones and $\bar{\boldsymbol{\mu}}_n$ the empirical $p$-mean vector. It is important to note that the between-class covariance matrix $\mathbf{\Sigma_B}$ is singular if the number of groups $K$ is lower than the minimum between the number of dimensions and observations. More specifically, searching for the Linear Fisher Discriminant subspace which optimizes the separability of the groups is equivalent to solve the $\mathbf{V}_1^{-1}(\mathbf{X}_n)\mathbf{V}_2(\mathbf{X}_n)$ eigenvalue problem.

### Other methods

In addition, other methods are based on solving a Generalized Eigenvalue Problem. First, Safo et al. (2016) define the Canonical Correlation Analysis as a GEP and solve it to compare two groups of quantitative variables both measured on the same individuals. This kind of analysis is notably used in bio-statistics to see if the groups describe the same phenomenon or not in order to reduce the further analysis to only one group of variables. Then, the Independent Component Analysis (ICA) is a special case of blind source separation. Contrary to PCA or ICS, even if it is also an unsupervised dimension reduction method, ICA is a model-based approach. For more details, see Hyvärinen et al.

(2004) and the master's thesis of Virta (2014). Finally, the Sliced Inverse Regression (SIR) is a supervised method of dimension reduction which regresses each explanatory variable against the response in order to find the best linear combinations in terms of explained univariate response. See among others Saracco et al. (1999), Liquet and Saracco (2012), Liski et al. (2014) or Oja et al. (2006) for more details.

To conclude, the simultaneous diagonalization of two scatter matrices is a common approach in multivariate data analysis. In the next section, we investigate solutions to adapt ICS to the case of singular scatter matrices.

## 4.3 ICS with semi-definite positive scatter matrices

With the previous definition of the ICS method, the two scatter matrices $\mathbf{V}_1$ and $\mathbf{V}_2$ should be definite positive to find a finite and nonzero eigenvalue $\rho$ to the eigen-problem (4.2): $\mathbf{V}_2\mathbf{b} = \rho\mathbf{V}_1\mathbf{b} \Leftrightarrow \mathbf{V}_1^{-1}\mathbf{V}_2\mathbf{b} = \rho\mathbf{b}$. The positive definiteness of $\mathbf{V}_1$ is required to compute its inverse whereas the positive definiteness of $\mathbf{V}_2$ ensures nonzero eigenvalues. In addition, in the previous Section 4.2, we define a scatter functional as belonging to $\mathcal{P}_p$. However, if the data contains collinear variables, or if the number of dimensions $p$ is greater than the number of observations $n$ (HDLSS data), then even the empirical variance-covariance matrix becomes singular. So, in this section, we define a scatter functional as a matrix $\mathbf{V}(F_{\mathbf{X}}) \in \mathcal{SP}_p$ which is affine equivariant in the sense that: $\mathbf{V}(F_{\mathbf{AX}+\boldsymbol{\gamma}}) = \mathbf{A}\mathbf{V}(F_{\mathbf{X}})\mathbf{A}'$, for all $p \times p$ non-singular matrices $\mathbf{A}$ and all $\boldsymbol{\gamma} \in \Re^p$. In this context, most of the affine equivariant scatter estimators are not well-defined and Tyler (2010) states that any affine equivariant scatter functional is proportional to the variance-covariance matrix if we assume that the data is in general position. So, as explained in the introduction, the paradigm taken in this chapter is that the data is not in general position.

Under this assumption, if $\mathbf{V}_1$ is singular, then the equivalence (4.2) is not true anymore since $\mathbf{V}_1$ is not invertible. The problem can not anymore be simplified and so, we have to solve the initial Generalized Eigenvalue Problem (GEP):

$$\mathbf{V}_2\mathbf{b}_i = \rho_i\mathbf{V}_1\mathbf{b}_i \qquad \text{for } i = 1, \ldots, p. \tag{4.4}$$

with $\mathbf{V}_1 \in \mathcal{SP}_p$ and $\mathbf{V}_2 \in \mathcal{SP}_p$ which are not necessarily of full ranks. If $\mathbf{V}_1$ and/or $\mathbf{V}_2$ is singular, that means that the null spaces of the two scatter matrices are not empty and they do not necessarily span the same subspace. Concretely, in this context, solving the GEP of $\mathbf{V}_1$ and $\mathbf{V}_2$, leads to consider the following cases:

— if $\mathbf{b} \in \text{range}(\mathbf{V}_1) \cap \text{range}(\mathbf{V}_2)$ then $\rho \in \Re^{+*}$.

— if $\mathbf{b} \in \text{null}(\mathbf{V}_2) - \text{null}(\mathbf{V}_1)$ then $\rho = 0$.

— if $\mathbf{b} \in \text{null}(\mathbf{V}_1) - \text{null}(\mathbf{V}_2)$ then $\rho = \infty$.

— if $\mathbf{b} \in \text{null}(\mathbf{V}_1) \cap \text{null}(\mathbf{V}_2)$ then any $\rho \in \Re$ is a solution of the GEP. The corresponding eigenvectors are not well defined and so, we do not consider further this case.

So, contrary to the classical ICS, the directions $\mathbf{b}$ associated with infinite or zero eigenvalues should also be analyzed as they might highlight some of the structure of the data.

This kind of issue also happens when searching the Fisher Linear Discriminant subspace as explained in Tebbens and Schlesinger (2007) or Howland and Park (2004) among others. As mentioned in 4.2.4, the classical Fisher Linear Discriminant Analysis (FLDA) maximizes the separation ratio of the between and the within-group covariance matrices $\mathbf{\Sigma}_B$ and $\mathbf{\Sigma}_W$. In this context, the rank of the between class covariance matrix $\mathbf{\Sigma}_B$ is equal to the number of groups, leading to a singular matrix. In addition, with HDLSS data, the within-group covariance matrix $\mathbf{\Sigma}_W$ can also be singular. Typically, the maximization problem cannot be performed by solving the $\mathbf{\Sigma}_W^{-1}\mathbf{\Sigma}_B$ eigenvalue problem anymore. However, solving a GEP of two scatter matrices with a common null space is particularly challenging. Indeed, the presence of this null space makes procedures like the well-known QZ-algorithm, introduced by Moler and Stewart (1973), very unstable. So, Tebbens and Schlesinger (2007) review some solutions to deal with these issues that can be adapted to ICS. The methods exploit the Moore-Penrose pseudo-inverse or the Generalized Singular Value Decomposition (GSVD) among others.

First, we illustrate the challenges of considering semi-definite scatter matrices for ICS on an artificial example. Then we adapt some well-known solutions to the singularity issues to ICS and we investigate theoretically and practically its new properties. Finally, we apply the method we consider as the best to a real data set.

### 4.3.1 Challenges with singular scatter matrices: an example

Let $\mathbf{X} = (X_1, \ldots, X_p)'$ be a $p$-multivariate real random vector and assume the distribution of $\mathbf{X}$ is a mixture of two Gaussian distributions with different covariance matrices:

$$\mathbf{X} \sim (1-\epsilon)\,\mathcal{N}\left(\mathbf{0}_p, \begin{pmatrix} \boldsymbol{W}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\right) + \epsilon\,\mathcal{N}\left(\mathbf{0}_p, \begin{pmatrix} \boldsymbol{W}_1 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{W}_2 \end{pmatrix}\right) \tag{4.5}$$

with $\epsilon < 1/2$, $\boldsymbol{W}_1 \in \mathcal{P}_{r_1}$ and $\boldsymbol{W}_2 \in \mathcal{SP}_{p-r_1}$ with $\text{rank}(\mathbf{W}_2) = r_2 \leq p - r_1$.
Such a distribution illustrates a model containing two clusters: the majority of the data and the outliers. The first cluster follows a Gaussian distribution such that the majority of the data is contained in a $r_1$-dimensional subspace spanned by the range of $\boldsymbol{W}_1$. The observations from the second cluster behave the same as previously on the $r_1$-dimensional subspace but they are also present in $r_2$ directions not spanned by the majority of the data. The goal of the ICS method is to find this $r_2$-dimensional subspace where the observations of the second cluster are outlying. Here this subspace is spanned by the range of $\mathbf{W}_2$ which is the orthogonal complement of the range of $\mathbf{W}_1$, i.e. the null space of $\mathbf{W}_1$.

Let us try to recover this subspace using the ICS method with a theoretical "perfectly robust" scatter functional $\mathbf{V}_1 = \begin{pmatrix} \boldsymbol{W}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$ and a theoretical "non-robust" scatter functional, the covariance of $F_{\mathbf{X}}$, $\mathbf{V}_2 = \begin{pmatrix} \boldsymbol{W}_1 & \mathbf{0} \\ \mathbf{0} & \epsilon \boldsymbol{W}_2 \end{pmatrix}$. We have $\mathbf{V}_1 \in \mathcal{SP}_p$ and $\mathbf{V}_2 \in \mathcal{SP}_p$, with $\text{rank}(\mathbf{V}_1) = r_1 < \text{rank}(\mathbf{V}_2) \leq p$. In addition, $\text{range}(\mathbf{V}_1) = \text{range}(\mathbf{W}_1)$ and $\text{range}(\mathbf{V}_2) = \text{range}(\mathbf{W}_1) \oplus \text{range}(\mathbf{W}_2)$.

Several of the aforementioned cases arise on this example. First, the intersection of the spaces spanned by the two scatter functionals $\mathbf{V}_1$ and $\mathbf{V}_2$ corresponds to the $r_1$-dimensional subspace spanned by $\mathbf{W}_1$, so $r_1$ nonzero eigenvalues should be found. Then, since $\text{null}(\mathbf{V}_1) - \text{null}(\mathbf{V}_2) \neq \{0\}$, a new direction associated to an $\infty$ eigenvalue should also be analyzed. In fact, this is the one which reveals the outliers. Finally, if $\text{rank}(\mathbf{V}_2) < p$, then the two scatter functionals share a part of their null subspaces. This subspace is not important since it contains no structure. However, we consider this phenomenon in our analysis because it is common in practice that the data is not of full rank. In addition, this feature could also make some algorithms like the QZ-algorithm unstable.

Practically, to illustrate the model (4.5), we generate 1000 observations with exactly 20 outliers, $\boldsymbol{W}_1 = \boldsymbol{I}_2$ and $\boldsymbol{W}_2 = \text{diag}(2, 0)$.

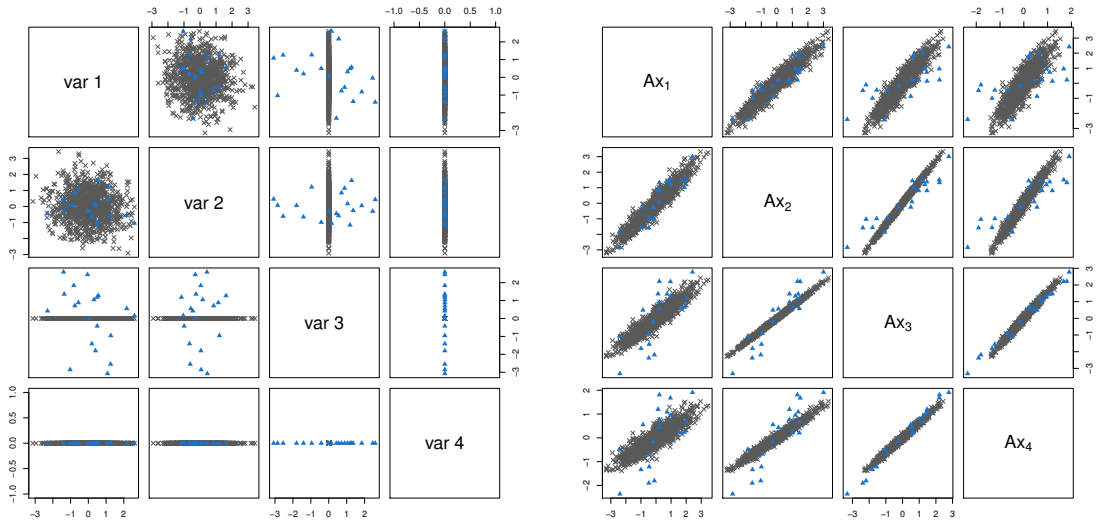



Figure 4.1 – Left: Scatterplot matrix of the simulated observations with 1000 observations, 20 outliers, $\boldsymbol{W}_1 = \boldsymbol{I}_2$ and $\boldsymbol{W}_2 = \text{diag}(2, 0)$. Right: Scatterplot matrix of the simulated observations transformed by the non-singular matrix $\mathbf{A}$.

On the left scatterplot matrix of the Figure 4.1, we can see that the outliers represented by some blue triangles behave differently than the majority of the data only on the third variable. The subspace spanned by this third variable is the only one interesting to identify these observations as outliers. This example can be seen as tricky since the outliers are well-identified on the third variable and that no observations lie on the fourth one. In addition, we apply an affine transformation so that the data looks more like real data. For

that, we choose a non-singular $p \times p$ matrix $\mathbf{A}$ as a particular Toeplitz matrix

$$\mathbf{A} = \begin{pmatrix} 1 & \frac{p-1}{p} & \frac{p-2}{p} & \frac{1}{p} \\ \frac{p-1}{p} & 1 & \frac{p-1}{p} & \frac{p-2}{p} \\ \frac{p-2}{p} & \frac{p-1}{p} & 1 & \frac{p-1}{p} \\ \frac{1}{p} & \frac{p-2}{p} & \frac{p-1}{p} & 1 \end{pmatrix} \tag{4.6}$$

to transform the initial vector $\mathbf{X}$ to $\mathbf{X}^* = \mathbf{A}\mathbf{X}$. We can notice on the right scatterplot matrix of the Figure 4.1, that the outliers are no longer as well separated on the third transformed variable as they were initially. In addition, we are no longer able to see that the observations lie on a three-dimensional subspace. However, the structure of outlierness of the data is still contained in one dimension only. The challenge is to be able to recover the direction spanned by the outliers with ICS.

### 4.3.2 ICS with a Moore-Penrose pseudo-inverse

The initial definition (4.2) of ICS leads to inverting $\mathbf{V}_1$ for solving the standard eigenvalue problem $\mathbf{V}_1^{-1}\mathbf{V}_2$. As we consider that $\mathbf{V}_1$ may be singular, the idea is to use its Moore-Penrose pseudo-inverse, $\mathbf{V}_1^+$ instead of $\mathbf{V}_1^{-1}$. In this case, it becomes possible to solve the $\mathbf{V}_1^+\mathbf{V}_2$ standard eigenvalue problem (EVP).

#### Principle and Interpretation

Computation of $\mathbf{V}_1^+$

The pseudo-inverse of $\mathbf{V}_1 \in \mathcal{SP}_p$ with $\text{rank}(\mathbf{V}_1) = r_1 < p$ can be defined based on its spectral decomposition: $\mathbf{V}_1 = \mathbf{P}\mathbf{\Lambda}\mathbf{P}'$, where $\mathbf{\Lambda}$ is a diagonal matrix with decreasing eigenvalues of $\mathbf{V}_1$, $\lambda_1 \geq \ldots \lambda_{r_1} > \lambda_{r_1+1} = \cdots = \lambda_p = 0$ and $\mathbf{P}$ is an orthogonal matrix containing the eigenvectors of $\mathbf{V}_1$. $\mathbf{P}$ can be partitioned as $\mathbf{P} = [\mathbf{P}_{r_1}\ \mathbf{P}_{p-r_1}]$, with $\mathbf{P}_{r_1}$, the $p \times r_1$ matrix containing the first $r_1$ eigenvectors associated to the $r_1$ nonzero eigenvalues of $\mathbf{V}_1$, which is an orthonormal basis for the range space of $\mathbf{V}_1$. Similarly, the $p \times p - r_1$ matrix $\mathbf{P}_{p-r_1}$ spans the null space of $\mathbf{V}_1$. $\mathbf{P}_{r_1}$ and $\mathbf{P}_{p-r_1}$ are semi-orthogonal matrices such that: $\mathbf{P}'_{r_1}\mathbf{P}_{r_1} = \mathbf{I}_{r_1}$ and $\mathbf{P}'_{p-r_1}\mathbf{P}_{p-r_1} = \mathbf{I}_{p-r_1}$.
$\mathbf{V}_1$ can be rewritten as $\mathbf{V}_1 = \mathbf{P}_{r_1}\mathbf{\Lambda}_{r_1}\mathbf{P}'_{r_1}$ and so its pseudo-inverse is $\mathbf{V}_1^+ = \mathbf{P}\mathbf{\Lambda}^+\mathbf{P}' = \mathbf{P}_{r_1}\mathbf{\Lambda}_{r_1}^{-1}\mathbf{P}'_{r_1}$, with $\mathbf{\Lambda}_{r_1}^{-1}$ containing only the inverse of the $r_1$ nonzero eigenvalues of $\mathbf{V}_1$.

Interpretation

To compare the ICS criterion, obtained when using a general inverse, to the initial one (4.3), we have to derive theoretically the solution of $\mathbf{V}_1^+\mathbf{V}_2$.

Usually, to simplify the computation, the eigenvalue problem $\mathbf{V}_1^+\mathbf{V}_2$ is transformed to the symmetric eigenvalue problem $\mathbf{V}_1^{+1/2}\mathbf{V}_2\mathbf{V}_1^{+1/2}$:

$$(\mathbf{\Lambda}_{r_1}^{-1/2}\mathbf{P}'_{r_1}\mathbf{V}_2\mathbf{P}_{r_1}\mathbf{\Lambda}_{r_1}^{-1/2} - \rho\mathbf{I}_{r_1})\mathbf{b}^* = 0 \tag{4.7}$$

with $\mathbf{b} = \mathbf{P}_{r_1}\mathbf{\Lambda}_{r_1}^{-1/2}\mathbf{b}^*$. By multiplying by $\mathbf{P}_{r_1}\mathbf{\Lambda}_{r_1}^{1/2}$, and because $\mathbf{P}_{r_1}$ is only semi-orthogonal, the equation (4.7) can be rewritten as:

$$(\mathbf{P}_{r_1}\mathbf{P}'_{r_1}\mathbf{V}_2\mathbf{P}_{r_1}\mathbf{P}'_{r_1} - \rho\mathbf{P}_{r_1}\mathbf{\Lambda}_{r_1}^{-1}\mathbf{P}'_{r_1})\mathbf{b} = 0 \tag{4.8}$$

which leads to the following modified ICS criterion for the eigenvector associated with the largest eigenvalue:

$$\max_{\mathbf{b}\in\Re^p,\mathbf{b}\neq 0} \quad \frac{\mathbf{b}'\mathbf{P}_{r_1}\mathbf{P}'_{r_1}\mathbf{V}_2\mathbf{P}_{r_1}\mathbf{P}'_{r_1}\mathbf{b}}{\mathbf{b}'\mathbf{P}_{r_1}\mathbf{\Lambda}_{r_1}\mathbf{P}'_{r_1}\mathbf{b}} \tag{4.9}$$

with $\mathbf{P}_{r_1}\mathbf{P}'_{r_1} = \text{Proj}_{\mathbf{V}_1}$, an orthogonal projection matrix onto the range($\mathbf{V}_1$), as $\mathbf{P}_{r_1}$ is an orthonormal basis for range($\mathbf{V}_1$). So range($\mathbf{P}_{r_1}\mathbf{P}'_{r_1}\mathbf{V}_2\mathbf{P}_{r_1}\mathbf{P}'_{r_1}$) $\subseteq$ range($\mathbf{V}_1$) and range($\mathbf{P}_{r_1}\mathbf{\Lambda}_{r_1}\mathbf{P}'_{r_1}$) = range($\mathbf{V}_1$). In addition, $\Re^p$ can be decomposed such that: $\Re^p$ = range($\mathbf{V}_1$) $\oplus$ null($\mathbf{V}_1$), thus the solution $\mathbf{b}$ of the criterion (4.9) can be expressed as:

$$\begin{aligned} &\mathbf{b} = \mathbf{v}_1 + \mathbf{v}_0 \quad \text{with} \quad \mathbf{v}_1 \in \text{range}(\mathbf{V}_1) \quad \text{and} \quad \mathbf{v}_0 \in \text{null}(\mathbf{V}_1), \\ &\text{with} \quad \mathbf{v}_1 = \underset{\mathbf{b}\in\Re^p,\mathbf{b}\neq 0}{\text{argmax}} \frac{\mathbf{b}'\text{Proj}_{\mathbf{V}_1}\mathbf{V}_2\text{Proj}_{\mathbf{V}_1}\mathbf{b}}{\mathbf{b}'\mathbf{P}_{r_1}\mathbf{\Lambda}_{r_1}\mathbf{P}'_{r_1}\mathbf{b}} \end{aligned} \tag{4.10}$$

As null($\mathbf{V}_1$) $\subseteq$ null($\text{Proj}_{\mathbf{V}_1}\mathbf{V}_2\text{Proj}_{\mathbf{V}_1}$), optimizing the new criterion (4.9) restricts the solutions to directions $\mathbf{b}$ only onto the subspace spanned by $\mathbf{V}_1$ and expressed by $\mathbf{v}_1$. If the structure of the data is only visible onto the subspace spanned by $\mathbf{V}_2$, in the null space of $\mathbf{V}_1$, then it is not possible to highlight it.

Example

Considering the example from Subsection 4.3.1, we compute the new ICS criterion.
Using the Moore-Penrose pseudo-inverse of $\mathbf{V}_1$ leads to optimize the criterion (4.9) that can be rewritten in our example as:

$$\max_{\mathbf{b}\in\Re^p,\mathbf{b}\neq 0} \quad \frac{\mathbf{b}'\text{Proj}_{\mathbf{W}_1}\mathbf{V}_2\text{Proj}_{\mathbf{W}_1}\mathbf{b}}{\mathbf{b}'\mathbf{W}_1\mathbf{b}} \quad = \quad \max_{\mathbf{b}\in\Re^p,\mathbf{b}\neq 0} \quad \frac{\mathbf{b}'\mathbf{W}_1\mathbf{b}}{\mathbf{b}'\mathbf{W}_1\mathbf{b}} = 1 \tag{4.11}$$

Clearly, in this case, any $\mathbf{b} \in \Re^p$ is solution of the maximization which implies that the structure of outlierness contained in $\mathbf{W}_2$ cannot be highlighted. If we go back to our simulated example, we obtain two eigenvalues equal to one as $\mathbf{V}_1$ is two-dimensional and two others equal to zero. The projection of the data onto the eigenvectors space is illustrated on the Figure 4.2. Definitely, the outliers cannot be identified because the eigenspace is restricted to the subspace spanned by $\mathbf{V}_1$ which does not contain the structure of outlierness defined by $\mathbf{W}_2$. So, the pseudo-inverse of $\mathbf{V}_1$ does not always give the correct solution to the singularity issue of the scatter matrices.

### Properties

In addition, let us determine the properties of ICS when using a generalized inverse.

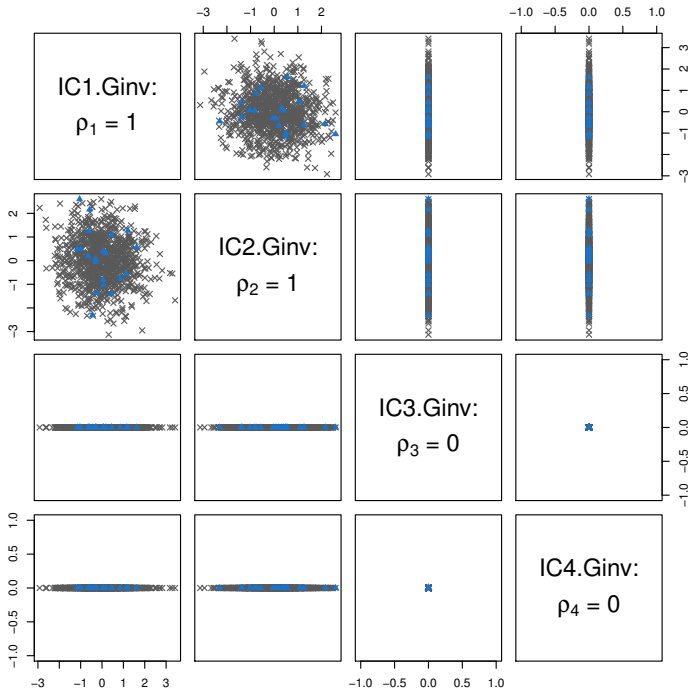



Figure 4.2 – Scatterplot matrix of the IC resulting of using a generalized inverse.

**Property 4.** ***Equivalence with the classical ICS.***
*If* $\mathbf{V}_1 \in \mathcal{P}_p$ *then solving the* $\mathbf{V}_1^{-1}\mathbf{V}_2$ *eigenvalue problem or using the Moore-Penroe pseudo-inverse of* $\mathbf{V}_1$ *is equivalent because* $\mathbf{V}_1^{+} = \mathbf{V}_1^{-1}$.

*Proof.* Let $\mathbf{X}^* = \mathbf{A}\mathbf{X}+\boldsymbol{\gamma}$, with $\mathbf{A}$ non-singular, $\boldsymbol{\gamma} \in \Re^p$ and $\mathbf{V}_1 \in \mathcal{SP}_p$ with $\text{rank}(\mathbf{V}_1) < p$. By definition of a scatter functional we have $\mathbf{V}_1(F_{\mathbf{X}^*}) = \mathbf{A}\mathbf{V}_1(F_{\mathbf{X}})\mathbf{A}'$ and if $\mathbf{V}_1(F_{\mathbf{X}^*})$ is not singular $\mathbf{V}_1^{-1}(F_{\mathbf{X}^*}) = (\mathbf{A}')^{-1}\mathbf{V}_1^{-1}(F_{\mathbf{X}})\mathbf{A}^{-1}$.
If $\mathbf{A}$ is an orthogonal matrix, let us prove that

$$\mathbf{V}_1(F_{\mathbf{X}^*})^{+} = (\mathbf{A}')^{-1}\mathbf{V}_1(F_{\mathbf{X}})^{+}\mathbf{A}^{-1}.$$

$\mathbf{V}_1(F_{\mathbf{X}^*})^{+}$ has to satisfy the four conditions:

- Condition 1: $\mathbf{V}_1(F_{\mathbf{X}^*})\mathbf{V}_1(F_{\mathbf{X}^*})^{+}\mathbf{V}_1(F_{\mathbf{X}^*}) = \mathbf{V}_1(F_{\mathbf{X}^*})$
- Condition 2: $\mathbf{V}_1(F_{\mathbf{X}^*})^{+}\mathbf{V}_1(F_{\mathbf{X}^*})\mathbf{V}_1(F_{\mathbf{X}^*})^{+} = \mathbf{V}_1(F_{\mathbf{X}^*})^{+}$
- Condition 3: $(\mathbf{V}_1(F_{\mathbf{X}^*})\mathbf{V}_1(F_{\mathbf{X}^*})^{+})' = \mathbf{V}_1(F_{\mathbf{X}^*})\mathbf{V}_1(F_{\mathbf{X}^*})^{+}$
- Condition 4: $(\mathbf{V}_1(F_{\mathbf{X}^*})^{+}\mathbf{V}_1(F_{\mathbf{X}^*}))' = \mathbf{V}_1(F_{\mathbf{X}^*})^{+}\mathbf{V}_1(F_{\mathbf{X}^*})$

The proof of conditions 1 and 2 can be generalized to any matrix $\mathbf{A}$ but conditions 3 and 4 rely on the assumption of orthogonality of $\mathbf{A}$.
Indeed, for condition 3, we have $(\mathbf{V}_1(F_{\mathbf{X}^*})\mathbf{V}_1(F_{\mathbf{X}^*})^{+})' = (\mathbf{A}')^{-1}\mathbf{V}_1(F_{\mathbf{X}})\mathbf{V}_1(F_{\mathbf{X}})^{+}\mathbf{A}'$. Since $\mathbf{A}$ is orthogonal $(\mathbf{A}')^{-1} = \mathbf{A}$ and $\mathbf{V}_1(F_{\mathbf{X}})^{+} = \mathbf{A}\mathbf{V}_1(F_{\mathbf{X}})^{+}\mathbf{A}'$, so we obtained the desired equality:

$$(\mathbf{A}')^{-1}\mathbf{V}_1(F_{\mathbf{X}})\mathbf{V}_1(F_{\mathbf{X}})^{+}\mathbf{A}' = \mathbf{A}\mathbf{V}_1(F_{\mathbf{X}})\mathbf{A}'\mathbf{A}\mathbf{V}_1(F_{\mathbf{X}})^{+}\mathbf{A}' = \mathbf{V}_1(F_{\mathbf{X}^*})\mathbf{V}_1(F_{\mathbf{X}^*})^{+}.$$

The proof is similar for the condition 4.
Then, we solve the eigenvalue problem $\mathbf{V}_1(F_{\mathbf{X}^*})^+\mathbf{V}_2(F_{\mathbf{X}^*})$ with $\mathbf{V}_1(F_{\mathbf{X}^*})^+ = \mathbf{A}\mathbf{V}_1(F_{\mathbf{X}})^+\mathbf{A}'$, which leads to the following modified ICS criterion:

$$\max_{\mathbf{b}\in\Re^p,\mathbf{b}\neq 0} \quad \frac{\mathbf{b}'\mathbf{A}\mathbf{P}_{r_1}\mathbf{P}'_{r_1}\mathbf{V}_2\mathbf{P}_{r_1}\mathbf{P}'_{r_1}\mathbf{A}'\mathbf{b}}{\mathbf{b}'\mathbf{A}\mathbf{P}_{r_1}\mathbf{\Lambda}_{r_1}\mathbf{P}'_{r_1}\mathbf{A}'\mathbf{b}} \tag{4.12}$$

Compared to the criterion (4.11), the eigenvectors are rotated by $\mathbf{A}'$ and so projecting the transformed data $\mathbf{X}^*$ onto $\mathbf{B}\mathbf{A}'$ or projecting $\mathbf{X}$ onto $\mathbf{B}$ leads to the same components. □

*Remark* 5. If $\mathbf{V}_1 \in \mathcal{SP}_p$ then the ICS coordinates are not necessarily invariant by an affine transformation since the the assumption of orthogonality is required in the proof (see the next counter-example 4.3.2).

Example
Let us consider the previous simulated example presented in 4.3.1 with 1000 observations, 20 outliers, $\boldsymbol{W}_1 = \boldsymbol{I}_2$, $\boldsymbol{W}_2 = \text{diag}(2, 0)$ and then transformed by the non-singular matrix $\mathbf{A}$ (4.6).
In this case, the structure of outlierness of the data is still contained only in one dimension. So, if the two scatter matrices $\mathbf{V}_1$ and $\mathbf{V}_2$ are of full ranks then, doing ICS on the initial data $\mathbf{X}$ or on the transformed $\mathbf{X}^*$ should lead to the same eigenvalues and the same scores. However, if $\text{rank}(\mathbf{V}_1) < p$ and if we use the pseudo-inverse $\mathbf{V}_1^+$ then we loose this affine invariance property of the Invariant Components (IC). Indeed, in the simulated example, we obtain two different eigenvalues, $\rho_1 = 1.1237$ and $\rho_2 = 1$ instead of the two equal to one, as mentioned on the Figure 4.3. Obviously, projecting the data onto the eigenvectors space leads to new scores and the affine invariance is lost.

To conclude, using the Moore-Penrose pseudo-inverse of $\mathbf{V}_1$ to find the maximal eigenvalue of the GEP of $\mathbf{V}_1$ and $\mathbf{V}_2$ presents three major drawbacks. First, the initial ICS criterion (4.3) may be modified to the criterion (4.9), which leads to find directions only on the subspace spanned by $\mathbf{V}_1$ and so the structure contained in the space spanned by $\text{null}(\mathbf{V}_1) - \text{null}(\mathbf{V}_2)$ cannot be highlighted. Second, if we use a generalized inverse, the scores are invariant up to an orthogonal transformation as for PCA. However, with the classical ICS method, the scores do not depend on the scale of the data. That is really unfortunate since it means that in this context there is an additional choice to make: standardize the data or not. Finally, the two scatter functionals $\mathbf{V}_1$ and $\mathbf{V}_2$ are not exchangeable anymore. Indeed, the directions found only span the range of the inverted scatter matrix. So, the ranks of the null spaces of $\mathbf{V}_1$ and $\mathbf{V}_2$ are now important. The results remain if $\mathbf{V}_2$ is singular or not.

### 4.3.3 A dimension reduction as preprocessing

As the use of a Moore-Penrose pseudo-inverse is not satisfactory to solve the ICS criterion (4.9) in the case when one or two of the scatter matrices is not of full rank, we present a well-known method which is a preprocessing step. This approach consists to get ride off the singularity issues by doing a reduction of dimension first, hoping that no

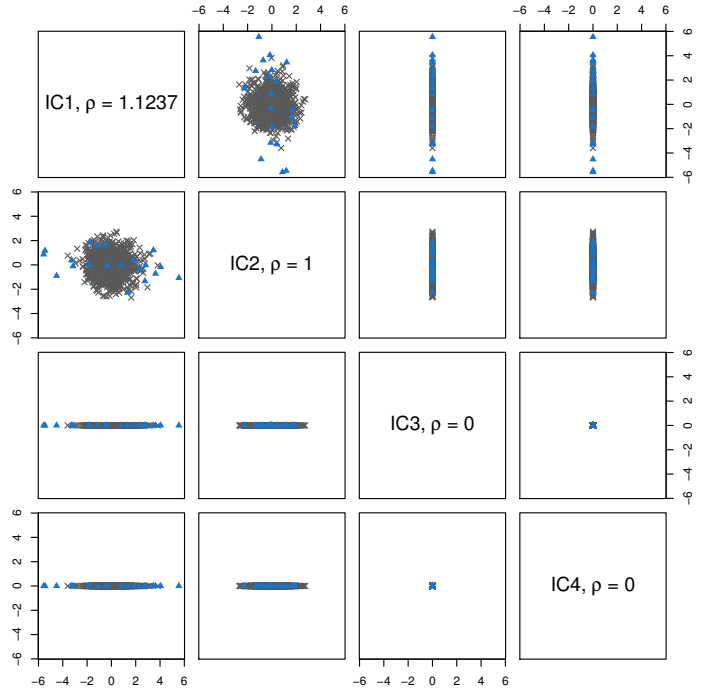



Figure 4.3 – Scatterplot matrix of the IC resulting of using a generalized inverse of $\mathbf{V}_1(\mathbf{X}_n^*)^+$.

information about the structure of the data will be lost, as mentioned in 1.4.2. The idea is to perform a Singular Value Decomposition (SVD) of the initial data and to project it onto the right-singular vectors associated to the non-zero singular values. Among others, Hubert et al. (2005) or Filzmoser et al. (2008) use this pre-processing step before applying their outlier detection algorithms based on PCA or Mahalanobis distances. This rank reduction is also used for the LDA method in the HDLSS context, as explained by Howland and Park (2004). However the performance of the preprocessing for the LDA method relies on the rank of the covariance matrix which has to fall into a specific range to ensure that the new within-covariance becomes non-singular. Another bad point is noted by She et al. (2016) whom advised against using a SVD before a robust PCA when outliers are present in the orthogonal complement subspace.

First, we recall the definition of a Singular Value Decomposition. Then, we demonstrate that this dimension reduction can also affect the ICS criterion and/or does not solve the singularity issues.

### Principle

In this section, $\mathbf{X}_n \in \Re^{n\times p}$ refers to the initial data containing $n$ observations, characterized by $p$ variables with all its columns properly centered by an equivariant location estimator and its rank $r_{\mathbf{X}_n} < \min(n, p-1)$. So, $\mathbf{V}_1 = \mathbf{V}_1(\mathbf{X}_n)$ and $\mathbf{V}_2 = \mathbf{V}_2(\mathbf{X}_n)$.

**Definition 1.** *Singular Value Decomposition, see for example Schott (2005), Chapter 4. It exists two $n \times n$ and $p \times p$ orthogonal matrices,* $\mathbf{U}$ *and* $\mathbf{P}$*, such that:*

$$\mathbf{X}_n = \mathbf{U}\mathbf{D}\mathbf{P}'$$

*with the $n \times p$ matrix* $\mathbf{D} = \begin{bmatrix} \mathbf{\Delta}^{1/2}_{r_{\mathbf{X}_n} \times r_{\mathbf{X}_n}} & \mathbf{0}_{r_{\mathbf{X}_n} \times (p - r_{\mathbf{X}_n})} \\ \mathbf{0}_{(n - r_{\mathbf{X}_n}) \times r_{\mathbf{X}_n}} & \mathbf{0}_{(n - r_{\mathbf{X}_n}) \times (p - r_{\mathbf{X}_n})} \end{bmatrix}$*, and* $\mathbf{U}\mathbf{U}' = \mathbf{U}'\mathbf{U} = \mathbf{I}_n$ *and* $\mathbf{P}\mathbf{P}' = \mathbf{P}'\mathbf{P} = \mathbf{I}_p$*. The elements of the diagonal matrix* $\mathbf{\Delta}^{1/2}$ *are the square roots of the positive eigenvalues of* $\mathbf{X}_n\mathbf{X}_n'$ *and* $\mathbf{X}_n'\mathbf{X}_n$*. The columns of* $\mathbf{P}$ *are also the eigenvectors of* $\mathbf{X}_n'\mathbf{X}_n$ *and the columns of* $\mathbf{U}$ *are the eigenvectors of* $\mathbf{X}_n\mathbf{X}_n'$*.*

$\mathbf{U}$ can be partitioned as $\mathbf{U} = [\mathbf{U}_1\ \mathbf{U}_2]$ where $\mathbf{U}_1$ is $n \times r_{\mathbf{X}_n}$ and $\mathbf{U}_2$ is $n \times (n - r_{\mathbf{X}_n})$. Similarly, $\mathbf{P}$ can be partitioned as $\mathbf{P} = [\mathbf{P}_1\ \mathbf{P}_2]$ where $\mathbf{P}_1$ is $p \times r_{\mathbf{X}_n}$ and $\mathbf{P}_2$ is $p \times (p - r_{\mathbf{X}_n})$. $\mathbf{U}_1$ and $\mathbf{P}_1$ are both semi-orthogonal matrices : $\mathbf{U}_1{}'\mathbf{U}_1 = \mathbf{P}_1{}'\mathbf{P}_1 = I_{r_{\mathbf{X}_n}}$. In addition, $\mathbf{U}_1$ (resp. $\mathbf{P}_1$) are orthonormal basis for the column space (resp. the row space) of $\mathbf{X}_n$. So, $\mathbf{X}_n$ can be decomposed as:

$$\mathbf{X}_n = \mathbf{U}_1\mathbf{\Delta}\mathbf{P}_1{}'$$

And its reduced form is obtained by projecting onto $\mathbf{P}_1$:

$$\mathbf{X}_n^* = \mathbf{X}_n\mathbf{P}_1 = \mathbf{U}_1\mathbf{\Delta}^{1/2}$$

where $\mathbf{X}_n^*$ is an $n \times r_{\mathbf{X}_n}$ matrix.

### Interpretation

Considering the projected data $\mathbf{X}_n^* \in \Re^{n \times r_{\mathbf{X}_n}}$ of rank $r_{\mathbf{X}_n}$, we have to solve the GEP of $\mathbf{V}_1(\mathbf{X}_n^*)$ and $\mathbf{V}_2(\mathbf{X}_n^*)$:

$$\mathbf{V}_2(\mathbf{X}_n^*)\mathbf{b} = \rho\mathbf{V}_1(\mathbf{X}_n^*)\mathbf{b}. \tag{4.13}$$

We can investigate several situations.

*(a) If* $\mathbf{X}_n$ *is of full rank.*
If $\mathbf{X}_n$ is of full rank then performing an SVD as a preprocessing step before ICS leads to the same components as if we directly compute the Invariant Coordinates from the initial data $\mathbf{X}_n$. Indeed, in this case, $\text{rank}(\mathbf{X}_n) = r = p$, the data is transformed by a $\mathbf{P}$ non-singular orthogonal $p \times p$ matrix and it is known that the Invariant Coordinates are invariant by an orthogonal transformation (see properties 1 and 2).

*(b) If* $\text{rank}(\mathbf{V}_1(\mathbf{X}_n^*)) = r_{\mathbf{X}_n} < p$.
If $\text{rank}(\mathbf{V}_1(\mathbf{X}_n^*)) = r_{\mathbf{X}_n}$, then the GEP (4.13) of $\mathbf{V}_1(\mathbf{X}_n^*)$ and $\mathbf{V}_2(\mathbf{X}_n^*)$ can be simplified to the classical EVP: $\mathbf{V}_1(\mathbf{X}_n^*)^{-1}\mathbf{V}_2(\mathbf{X}_n^*)\mathbf{b} = \rho\mathbf{b}$. However, doing the first step of dimension reduction and transforming the data onto $\mathbf{X}_n^*$ does not ensure the non-singularity of $\mathbf{V}_1(\mathbf{X}_n^*)$ and/or $\mathbf{V}_2(\mathbf{X}_n^*)$. So, it is possible to choose a $\mathbf{V}_1(\mathbf{X}_n^*)$ and/or $\mathbf{V}_2(\mathbf{X}_n^*)$ of lower rank than $\mathbf{X}_n$. In this case, the GEP (4.13) does not solve the problem since $\mathbf{V}_1(\mathbf{X}_n^*)$ is still singular.

**Property if** $\text{rank}(\mathbf{X}_n) = r_{\mathbf{X}_n} < p$ **and** $\mathbf{V}_1 = \text{COV}$**.**

Let us consider estimators of the form:

$$\mathbf{V}_2(\mathbf{X}_n) = \sum_{i=1}^{n} w_i(\mathbf{x}_i - \mathbf{m}(\mathbf{X}_n))(\mathbf{x}_i - \mathbf{m}(\mathbf{X}_n))' \qquad \text{with} \qquad \mathbf{m}(\mathbf{X}_n) = \sum_{i=1}^{n} w_i \mathbf{x}_i \tag{4.14}$$

where $w_i = w(\mathbf{X}_n)$ and $w$ is a nonnegative function of $\mathbf{X}_n$.
Note that if $w_i = 1$ for $i = 1, \ldots, n$, then the estimator $V(\mathbf{X}_n)$ corresponds to the empirical variance-covariance matrix. In fact, a lot of scatter matrices can be written under this weigthed form. In general, the weights $w_i$ are a function of the Mahalanobis distance. This is the case for the well-known scatter matrix based on the fourth moments $\text{COV}_4$, the one-step W-estimators, the M-estimators or the reweigthed MCD. However, the weights can also depend of other outlierness indexes, like in the Stahel-Donoho estimator (Stahel, 1981; Donoho, 1982). Formally, in this case the weigths are equal to:

$$w_i = \frac{w(r(\mathbf{x}_i, \mathbf{X}_n))}{\sum_{i=1}^{n} w(r(\mathbf{x}_i, \mathbf{X}_n))} \qquad \text{with} \qquad r(\mathbf{x}_i, \mathbf{X}_n) = \sup_{\mathbf{a} \in \Re^p} \frac{|\mathbf{a}'\mathbf{x}_i - \mu(\mathbf{X}_n a)|}{\sigma(\mathbf{X}_n a)} \tag{4.15}$$

where $\mu(.)$ and $\sigma(.)$ are equivariant univariate location and scale statistics.

**Proposition 7.**
*Performing ICS with the variance-covariance for* $\mathbf{V}_1$ *and any scatter matrix for* $\mathbf{V}_2$ *satisfying the condition (4.21) detailed below, after the dimension reduction pre-processing is equivalent to solve the ICS criterion using the Moore-Penrose pseudo-inverse of* $\mathbf{V}_1$*.*

*Proof.* Using the notations introduced in Section 4.3.2 for the Moore-Penrose pseudo-inverse, we obtain $\mathbf{V}_1(\mathbf{X}_n^*) = \mathbf{\Lambda}_{r_1}$. This is because the right eigenvectors from the singular value decomposition of $\mathbf{X}_n$ are the ones of $\mathbf{X}_n'\mathbf{X}_n = \mathbf{V}_1(\mathbf{X}_n) = \mathbf{P}_{r_1}\mathbf{\Lambda}_{r_1}\mathbf{P}_{r_1}'$ and $\mathbf{V}_1(\mathbf{X}_n^*) = \mathbf{P}_{r_1}'\mathbf{V}_1(\mathbf{X}_n)\mathbf{P}_{r_1}$ with $\mathbf{P}_{r_1}$ being a $p \times r_1$ semi-orthognal matrix and so $r_{\mathbf{X}_n} = r_1 < p$.

In this case, performing ICS on the projected data $\mathbf{X}_n^*$ leads to optimize the following criterion:

$$\max_{\mathbf{b} \in \Re^{r_1}, \mathbf{b} \neq 0} \quad \frac{\mathbf{b}'\mathbf{V}_2(\mathbf{X}_n^*)\mathbf{b}}{\mathbf{b}'V_1(\mathbf{X}_n^*)\mathbf{b}} \tag{4.16}$$

Using a generalized inverse of $\mathbf{V}_1$ leads to optimize the modified criterion (4.9):

$$\max_{\mathbf{a} \in \Re^p, \mathbf{a} \neq 0} \quad \frac{\mathbf{a}'\mathbf{P}_{r_1}\mathbf{P}_{r_1}'\mathbf{V}_2(\mathbf{X}_n)\mathbf{P}_{r_1}\mathbf{P}_{r_1}'\mathbf{a}}{\mathbf{a}'\mathbf{P}_{r_1}\mathbf{\Lambda}_{r_1}\mathbf{P}_{r_1}'\mathbf{a}} \tag{4.17}$$

We can notice that criterion (4.17) is equivalent to:

$$\max_{\mathbf{a} \in \Re^p, \mathbf{a} \neq 0} \quad \frac{\mathbf{a}'\mathbf{P}_{r_1}\mathbf{P}_{r_1}'\mathbf{V}_2(\mathbf{X}_n)\mathbf{P}_{r_1}\mathbf{P}_{r_1}'\mathbf{a}}{\mathbf{a}'\mathbf{P}_{r_1}\mathbf{P}_{r_1}'\mathbf{V}_1(\mathbf{X}_n)\mathbf{P}_{r_1}\mathbf{P}_{r_1}'\mathbf{a}} \tag{4.18}$$

since $\mathbf{P}_{r_1}'\mathbf{P}_{r_1} = \mathbf{I}_{r_1}$. In addition, we know that $\mathbf{a} = \mathbf{v}_1 + \mathbf{v}_0$ with $\mathbf{v}_1 \in \text{range}(\mathbf{V}_1)$ and $\mathbf{v}_0 \in \text{null}(\mathbf{V}_1)$, so $\mathbf{P}_{r_1}\mathbf{P}_{r_1}'\mathbf{a} = \mathbf{v}_1$, and the criterion (4.18) is equivalent to:

$$\max_{\mathbf{v}_1 \in \text{range}(\mathbf{V}_1), \mathbf{v}_1 \neq 0} \quad \frac{\mathbf{v}_1'\mathbf{V}_2(\mathbf{X}_n)\mathbf{v}_1}{\mathbf{v}_1'\mathbf{V}_1(\mathbf{X}_n)\mathbf{v}_1} \tag{4.19}$$

Since $\mathbf{v}_1 \in \text{range}(\mathbf{V}_1)$, $\exists\ \mathbf{b}_1$ s.t $\mathbf{v}_1 = \mathbf{P}_{r_1}\mathbf{b}_1$ and so the criterion (4.19) is equivalent to:

$$\max_{\mathbf{b}_1 \in \Re^{r_1}, \mathbf{b}_1 \neq 0} \quad \frac{\mathbf{b}_1' \mathbf{P}_{r_1}' \mathbf{V}_2(\mathbf{X}_n) \mathbf{P}_{r_1} \mathbf{b}_1}{\mathbf{b}_1' \mathbf{P}_{r_1}' \mathbf{V}_1(\mathbf{X}_n) \mathbf{P}_{r_1} \mathbf{b}_1} \tag{4.20}$$

and if we assume the next condition to be true:

$$\mathbf{V}_2(\mathbf{X}_n^*) = \mathbf{P}_{r_1}' \mathbf{V}_2(\mathbf{X}_n) \mathbf{P}_{r_1} \tag{4.21}$$

then, the criterion (4.20) is equivalent to:

$$\max_{\mathbf{b}_1 \in \Re^{r_1}, \mathbf{b}_1 \neq 0} \quad \frac{\mathbf{b}_1' \mathbf{V}_2(\mathbf{X}_n^*) \mathbf{b}_1}{\mathbf{b}_1' V_1(\mathbf{X}_n^*) \mathbf{b}_1} \tag{4.22}$$

which is exactly the criterion 4.16 that we maximize if we perform a dimension reduction before ICS.

□

Note that the condition (4.21) is true for the Stahel-Donoho estimator (defined in (4.15)) or any estimator of the previous weighted form if their weights are invariant by the transformation $\mathbf{P}_{r_1}$.

So in this case, doing the pre-processing leads to an additional step to the method which is not needed and which implies the same drawbacks as doing ICS with a generalized inverse.

To conclude, the preprocessing step of dimension reduction is not fulfilling all its promises. First, it cannot guarantee that it solves the singularity issues of the scatter matrices. Then, even if it does, if we choose $\mathbf{V}_1$ as the variance-covariance matrix, we recover exactly the same modified criterion to solve as when we use the generalized inverse and so the same drawbacks. Finally, if we choose $\mathbf{V}_2$ as the variance-covariance matrix, we might be unable to recover the structure of the data if it is only contained on the subspace spanned by $\mathbf{V}_1$.

### 4.3.4 ICS with a Generalized Singular Value Decomposition

Note that, in this section, we use either sample or functional versions of the scatter estimators.

#### Principle and properties

Contrary to the two previous methods which solve the GEP (4.4): $\mathbf{V}_2\mathbf{b}_i = \rho_i \mathbf{V}_1 \mathbf{b}_i$, for $i = 1, \ldots, p.$ , we propose to solve the following problem:

$$\beta_i^2 \mathbf{V}_2 \mathbf{h}_i = \alpha_i^2 \mathbf{V}_1 \mathbf{h}_i \qquad \text{for } i = 1, \ldots, p. \tag{4.23}$$

Re-writting the GEP (4.4) as in (4.23) presents some advantages compared to the other two methods. First, it allows us to find all the directions which can reveal some structure of the data in the general case where $\mathbf{V}_1 \in \mathcal{SP}_p$ and $\mathbf{V}_2 \in \mathcal{SP}_p$, as summarized

in the Table 4.1. Second, it is clear that $\mathbf{V}_1$ and $\mathbf{V}_2$ play a symmetric role. This is

| Direction | $\alpha_i$ | $\beta_i$ | $\rho$ | Optimization of the ICS ratio (4.3) |
|---|---|---|---|---|
| $\mathbf{h}_i \in \text{range}(\mathbf{V}_1) \cap \text{range}(\mathbf{V}_2)$ | $\alpha_i \neq 0$ | $\beta_i \neq 0$ | $\rho_i \in \Re^{+*}$ | Minimize or maximize |
| $\mathbf{h}_i \in \text{null}(\mathbf{V}_2) - \text{null}(\mathbf{V}_1)$ | $\alpha_i = 0$ | $\beta_i \in \Re^{+*}$ | $\rho_i = 0$ | Minimize |
| $\mathbf{h}_i \in \text{null}(\mathbf{V}_1) - \text{null}(\mathbf{V}_2)$ | $\alpha_i \in \Re^{+*}$ | $\beta_i = 0$ | $\rho_i = \infty$ | Maximize |

Table 4.1 – Summary of the different possible directions $\mathbf{h}_i$ depending on the values of $\alpha_i$ and $\beta_i$.

important since the other methods can miss the structure of the data if it is contained into the subspace spanned by $\mathbf{V}_2$ and in the null space of $\mathbf{V}_1$ in particular. Third, this formulation is still equivalent to the classical EVP (4.2) if the two scatter matrices are of full ranks with $\rho_i = \alpha_i^2/\beta_i^2$. In addition to these nice characteristics, the invariant coordinates remain invariant by affine transformation as we demonstrate below.

The affine invariance property is valid for simple or multiple eigenvalues as in Propositions 1 and 2

**Proposition 8.** ***Affine invariance property.***
*For two affine equivariant scatter matrices $\mathbf{V}_1$ and $\mathbf{V}_2$, and using the eigenvectors defined by (4.23), the invariant coordinates are invariant by affine transformation, see Propositions 1 and 2.*

*Proof.*
*(i) Adaptation of the proof from Tyler et al. (2009), appendix A.1, for distinct roots.*
Let $\mathbf{X}^* = \mathbf{G}\mathbf{X} + \boldsymbol{\gamma}$, with $\boldsymbol{\gamma} \in \Re^p$. Then $\mathbf{V}_1(F_{\mathbf{X}^*}) = \mathbf{G}\mathbf{V}_1(F_{\mathbf{X}})\mathbf{G}'$ and $\mathbf{V}_2(F_{\mathbf{X}^*}) = \mathbf{G}\mathbf{V}_2(F_{\mathbf{X}})\mathbf{G}'$.
By definition of ICS we have, for $i = 1, \ldots, p$:

$$\alpha_i^2(F_{\mathbf{X}^*})\mathbf{V}_2(F_{\mathbf{X}^*})\mathbf{h}_i(F_{\mathbf{X}^*}) = \beta_i^2(F_{\mathbf{X}^*})\mathbf{V}_1(F_{\mathbf{X}^*})\mathbf{h}_i(F_{\mathbf{X}^*})$$

$$\alpha_i^2(F_{\mathbf{X}^*})\mathbf{G}\mathbf{V}_2(F_{\mathbf{X}})\mathbf{G}'\mathbf{h}_i(F_{\mathbf{X}^*}) = \beta_i^2(F_{\mathbf{X}^*})\mathbf{G}\mathbf{V}_1(F_{\mathbf{X}})\mathbf{G}'\mathbf{h}_i(F_{\mathbf{X}^*})$$

Multiplying by $\mathbf{G}^{-1}$: $\alpha_i^2(F_{\mathbf{X}^*})\mathbf{V}_2(F_{\mathbf{X}})\mathbf{G}'\mathbf{h}_i(F_{\mathbf{X}^*}) = \beta_i^2(F_{\mathbf{X}^*})\mathbf{V}_1(F_{\mathbf{X}})\mathbf{G}'\mathbf{h}_i(F_{\mathbf{X}^*})$
If $\alpha_i^2(F_{\mathbf{X}})/\beta_i^2(F_{\mathbf{X}})$ is a distinct root, then $\alpha_i^2(F_{\mathbf{X}})/\beta_i^2(F_{\mathbf{X}}) = \alpha_i^2(F_{\mathbf{X}^*})/\beta_i^2(F_{\mathbf{X}^*})$ and $\mathbf{h}_i(F_{\mathbf{X}}) \propto \mathbf{G}'\mathbf{h}_i(F_{\mathbf{X}^*})$, so $\mathbf{h}_i(F_{\mathbf{X}^*}) \propto \mathbf{G}'^{-1}\mathbf{h}_i(F_{\mathbf{X}})$ :

$$\alpha_i^2(F_{\mathbf{X}})\mathbf{V}_2(F_{\mathbf{X}})\mathbf{h}_i(F_{\mathbf{X}}) = \beta_i^2(F_{\mathbf{X}})\mathbf{V}_1(F_{\mathbf{X}})\mathbf{h}_i(F_{\mathbf{X}})$$

Projection onto $\mathbf{h}_i(F_{\mathbf{X}^*})$: $\mathbf{z}_i^* = \mathbf{h}_i(F_{\mathbf{X}^*})'\mathbf{x}^* = (\mathbf{G}'^{-1}\mathbf{h}_i(F_{\mathbf{X}}))'\mathbf{G}\mathbf{x} = \mathbf{h}_i(F_{\mathbf{X}})'\mathbf{x} = \mathbf{z}_i$.

*(ii) Proof from Tyler et al. (2009), appendix A.1, for multiple roots.*
In case of a multiple root of multiplicity $p_l$, the eigenvectors are not uniquely defined and can be chosen as any linearly independent vectors spanning the corresponding $p_l$-dimensional eigenspace. However the roots are still the same and so the subspace spanned by the corresponding $p_l$-dimensional eigenspace is still the same. □

This case of multiple roots may appear when $\mathbf{V}_1 \in \mathcal{SP}_p$ and/or $\mathbf{V}_2 \in \mathcal{SP}_p$ as it means than $\text{null}(\mathbf{V}_1) \neq \{0\}$ and/or $\text{null}(\mathbf{V}_2) \neq \{0\}$. For example, if we only focus on the cases

where $\mathbf{h} \in \text{null}(\mathbf{V}_2) - \text{null}(\mathbf{V}_1)$ or $\mathbf{h} \in \text{null}(\mathbf{V}_1) - \text{null}(\mathbf{V}_2)$, if $\dim(\text{null}(\mathbf{V}_2) - \text{null}(\mathbf{V}_1)) > 1$ and/or $\dim(\text{null}(\mathbf{V}_1) - \text{null}(\mathbf{V}_2)) > 1$ then 0 and/or $\infty$ are multiple roots.

EXAMPLE

Let us go back to our simulated example with the observations transformed by the non-singular matrix $\mathbf{A}$ as illustrated on the Figure 4.1. Even if the outliers are not visible anymore on the third variable, this transformation does not affect the results of the ICS method based on solving (4.23). The eigenvalues are the same as previously, namely $\rho_1 = \infty$ and $\rho_2 = \rho_3 = 1$ and so, the scores are really invariant as it can be noticed on the Figure 4.4. As in Figure 4.5, the outliers are clearly highlighted on the component associated to the infinite eigenvalue.

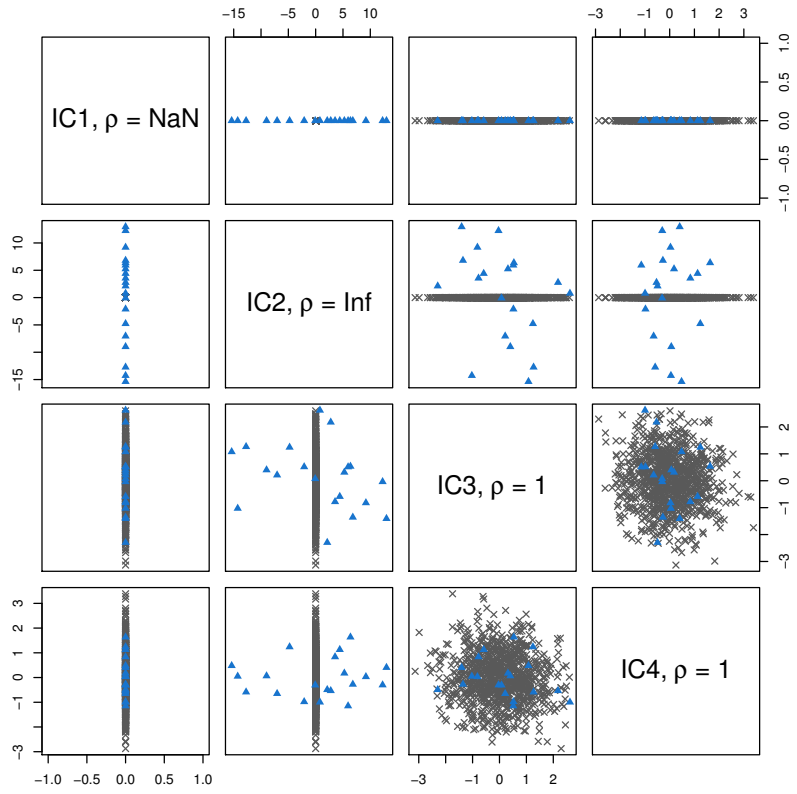



Figure 4.4 – Scatterplot matrix of the IC resulting from the GSVD of $\mathbf{X}^*_{\mathbf{V}_1}$ and $\mathbf{X}^*_{\mathbf{V}_2}$.

We present in the next section how we can implement a method to solve the problem (4.23).

### Implementation with a Generalized Singular Value Decomposition (GSVD)

In this section, we consider only sample location and scatter estimators based on $\mathbf{X}_n \in \Re^{n \times p}$.

Several methods exist to solve a Generalized Eigenvalue Problem as the GEP (4.23). Among others, it exists the the well-known QZ-algorithm introduced by Moler and Stew-

art (1973) or the procedure described by Schott (2005). In this section, we focus on an implementation based on a Generalized Singular Value Decomposition (GSVD) as proposed by Howland et al. (2003, 2006); Howland and Park (2004) and Kim et al. (2005) in the LDA context.

More specifically, they use a GSVD for computing eigenvectors to define the Fisher's discriminant subspace, when the between $\mathbf{\Sigma}_B$ and the within-group $\mathbf{\Sigma}_W$ covariance matrices, are susceptible to be singular. The only requirement with this method is to express $\mathbf{\Sigma}_B$ and $\mathbf{\Sigma}_W$ as cross-products matrices, which is easily obtained by their definition. Let us define $\mathbf{X}_{\mathbf{\Sigma}_B} \in \Re^{n\times p}$ s.t. $\mathbf{X}'_{\mathbf{\Sigma}_B}\mathbf{X}_{\mathbf{\Sigma}_B} = \mathbf{\Sigma}_B$ and $\mathbf{X}_{\mathbf{\Sigma}_W} \in \Re^{n\times p}$ s.t. $\mathbf{X}'_{\mathbf{\Sigma}_W}\mathbf{X}_{\mathbf{\Sigma}_W} = \mathbf{\Sigma}_W$. Then the GSVD of the $n \times p$ matrices $\mathbf{X}_{\mathbf{\Sigma}_B}$ and $\mathbf{X}_{\mathbf{\Sigma}_W}$ gives the eigenvalues and eigenvectors of the pencil $\mathbf{X}'_{\mathbf{\Sigma}_B}\mathbf{X}_{\mathbf{\Sigma}_B} - \rho\mathbf{X}'_{\mathbf{\Sigma}_W}\mathbf{X}_{\mathbf{\Sigma}_W}$ which is equivalent to solve the pencil $\mathbf{\Sigma}_B - \rho\mathbf{\Sigma}_W$. In this case, we define $\mathbf{X}_{\mathbf{\Sigma}_W}$ and $\mathbf{X}_{\mathbf{\Sigma}_B}$ s.t $\mathbf{X}_{\mathbf{\Sigma}_W} = \frac{1}{\sqrt{n}}(\mathbf{X}_n - \mathbf{G}_K\mathbf{M}'_K)$ and $\mathbf{X}_{\mathbf{\Sigma}_B} = \frac{1}{\sqrt{n}}(\mathbf{G}_K\mathbf{M}'_K - \mathbf{1}_n\bar{\boldsymbol{\mu}}'_n)$ with $\mathbf{G}_K$ as a $n\times K$ matrix with $\mathbf{G}_{i,k}$ an indicator of whether the observation $i$ is in class $k$, $\mathbf{M}_K$ is a $p \times K$ matrix s.t $\mathbf{M}_K = (\bar{\boldsymbol{\mu}}_{n,1} \dots \bar{\boldsymbol{\mu}}_{n,k} \dots \bar{\boldsymbol{\mu}}_{n,K})$ with $\bar{\boldsymbol{\mu}}_{n,k}$ be the mean $p$-vector of the observations inside the class $k$, $\mathbf{1}_n$ a $n$-vector of ones and $\bar{\boldsymbol{\mu}}_n$ the empirical $p$-mean vector.

This procedure, which uses a GSVD to solve a GEP, can be applied to other scatter matrices which can be expressed as crossproducts. Let us consider estimators of the same form as defined in Section 4.3.3. They can be expressed as crossproducts such that:

$$V(\mathbf{X}_n) = \mathbf{X}'_V\mathbf{X}_V \qquad \text{with} \qquad \mathbf{X}_V = \mathbf{W}(\mathbf{X}_n - \mathbf{1}_n t(\mathbf{X}_n)') \tag{4.24}$$

where $\mathbf{X}_V \in \Re^{n\times p}$ and $\mathbf{W} = \text{diag}(\sqrt{w_1}, \dots, \sqrt{w_n})$.

Defining a stable algorithm for the GSVD is very challenging and a lot of research was done regarding this topic, such as Paige and Saunders (1981); Paige (1986); Bai (1992); Bai and Demmel (1993); Bai and Zha (1993); Golub and Van Loan (1996) among others. In this section, we present the GSVD procedure as it is given in Hogben (2006) in Section 75-11. We retain this definition since it is already implemented into LAPACK and can be used directly in R through the geigen (Hasselman and Lapack authors, 2017) package. Then, we show how this decomposition can solve our generalized eigenvalue problem. Finally, we investigate its properties and illustrate them on an example.

### Definition

We follow the definition given in Hogben (2006), restricted to the case of real matrices. Let us define $\mathbf{X}_{\mathbf{V}_1} \in \Re^{n\times p}$ s.t. $\mathbf{X}'_{\mathbf{V}_1}\mathbf{X}_{\mathbf{V}_1} = \mathbf{V}_1 = \mathbf{V}_1(\mathbf{X}_n)$ and $\mathbf{X}_{\mathbf{V_2}} \in \Re^{n\times p}$ s.t. $\mathbf{X}'_{\mathbf{V_2}}\mathbf{X}_{\mathbf{V_2}} = \mathbf{V_2} = \mathbf{V}_2(\mathbf{X}_n)$. $\mathbf{V}_1, \mathbf{V}_2 \in \mathcal{SP}_p$ with $\text{rank}(\mathbf{X}_{\mathbf{V}_1}) = \text{rank}(\mathbf{V}_1) = r_1 \leq p$ and $\text{rank}(\mathbf{X}_{\mathbf{V}_2}) = \text{rank}(\mathbf{V}_2) = r_2 \leq p$.

**Definition 2.** *Generalized Singular Value Decomposition (GSVD)*
*The Generalized (or Quotient) Singular Value Decomposition (GSVD or QSVD) of two $n \times p$ matrices $\mathbf{X}_{\mathbf{V}_1}$ and $\mathbf{X}_{\mathbf{V}_2}$ is given by the pair of factorization as:*

$$\mathbf{X}_{\mathbf{V}_1} = \mathbf{U}\mathbf{D}_1[\mathbf{0}\ \mathbf{R}]\mathbf{Q}' \quad \textit{and} \quad \mathbf{X}_{\mathbf{V}_2} = \mathbf{V}\mathbf{D}_2[\mathbf{0}\ \mathbf{R}]\mathbf{Q}' \tag{4.25}$$

*The matrices in these factorizations have the following properties:*

— $\mathbf{U}$ *and* $\mathbf{V}$ *are* $n \times n$, $\mathbf{Q}$ *is* $p \times p$, *and all three matrices are orthogonal.*

— $\mathbf{R}$ *is* $r \times r$, *upper triangular and nonsingular with* $r = \operatorname{rank}([\mathbf{X}'_{\mathbf{V}_1}, \mathbf{X}'_{\mathbf{V}_2}]')$. $[\mathbf{0}\ \mathbf{R}]$ *is* $r \times p$ *(in other words, the* $\mathbf{0}$ *is an* $r \times (p-r)$ *zero matrix).*

— $\mathbf{D}_1$ *and* $\mathbf{D}_2$ *are* $n \times r$. *Both are real, nonnegative, and diagonal, satisfying* $\mathbf{D}'_1\mathbf{D}_1 + \mathbf{D}'_2\mathbf{D}_2 = \mathbf{I}_r$. *Write* $\mathbf{D}'_1\mathbf{D}_1 = \operatorname{diag}(\alpha_1^2, \ldots, \alpha_r^2)$ *and* $\mathbf{D}'_2\mathbf{D}_2 = \operatorname{diag}(\beta_1^2, \ldots, \beta_r^2)$, *the ratios* $\alpha_j/\beta_j$ *for* $j = 1, \ldots, r$ *are called the generalized singular values.*

$\mathbf{D}_1$ *and* $\mathbf{D}_2$ *have the following structure, depending on whether* $n - r \geq 0$ *or* $n - r < 0$.

— *In the first case, when* $n - r \geq 0$,

$$\mathbf{D}_1 = \begin{array}{c} \\ r - r_1 \\ r_1 \\ n - r \end{array} \begin{array}{c} \begin{array}{cc} r - r_1 & r_1 \end{array} \\ \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{C} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \end{array}, \quad \mathbf{D}_2 = \begin{array}{c} \\ r_1 \\ n - r_1 \end{array} \begin{array}{c} \begin{array}{cc} r - r_1 & r_1 \end{array} \\ \begin{pmatrix} \mathbf{0} & \mathbf{S} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \end{array}$$

$\mathbf{C}$ *and* $\mathbf{S}$ *are diagonal matrices satisfying* $\mathbf{C}^2 + \mathbf{S}^2 = I_{r_1}$ *and* $\mathbf{S}$ *is nonsingular. Let* $c_j$ *and* $s_j$ *be the diagonal entries of* $\mathbf{C}$ *and* $\mathbf{S}$, *respectively. Then we have* $\alpha_1 = \cdots = \alpha_{r-r_1} = 1, \alpha_{r-r_1+j} = c_j$ *for* $j = 1, \ldots, r_1$, $\beta_1 = \cdots = \beta_{r-r_1} = 0$, *and* $\beta_{r-r_1+j} = s_j$ *for* $j = 1, \ldots, r_1$. *Thus, the first* $r - r_1$ *generalized singular values* $\alpha_1/\beta_1, \ldots, \alpha_{r-r_1}/\beta_{r-r_1}$ *are infinite and the remaining* $r_1$ *generalized singular values are finite.*

— *In the second case, when* $n - r < 0$,

$$\mathbf{D}_1 = \begin{array}{c} \\ r - r_1 \\ n - r + r_1 \end{array} \begin{array}{c} \begin{array}{ccc} r - r_1 & n - r + r_1 & r - n \end{array} \\ \begin{pmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{C} & \mathbf{0} \end{pmatrix} \end{array}$$

$$\mathbf{D}_2 = \begin{array}{c} \\ n - r + r_1 \\ r - n \\ n - r_1 \end{array} \begin{array}{c} \begin{array}{ccc} r - r_1 & n - r + r_1 & r - n \end{array} \\ \begin{pmatrix} \mathbf{0} & \mathbf{S} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix} \end{array}$$

*Again,* $\mathbf{C}$ *and* $\mathbf{S}$ *are diagonal matrices satisfying* $\mathbf{C}^2 + \mathbf{S}^2 = I_{n-r+r_1}$, *and* $\mathbf{S}$ *is nonsingular. Let* $c_j$ *and* $s_j$ *be the diagonal entries of* $\mathbf{C}$ *and* $\mathbf{S}$, *respectively. Then we have* $\alpha_1 = \cdots = \alpha_{r-r_1} = 1, \alpha_{r-r_1+j} = c_j$ *for* $j = 1, \ldots, n-r+r_1$, $\alpha_{n+1} = \cdots = \alpha_r = 0$, *and* $\beta_1 = \cdots = \beta_k = 0$, $\beta_{r-r_1+j} = s_j$ *for* $j = 1, \ldots, n-r+r_1$, $\beta_{n+1} = \cdots = \beta_r = 1$. *Thus, the first* $r - r_1$ *generalized singular values* $\alpha_1/\beta_1, \ldots, \alpha_{r-r_1}/\beta_{r-r_1}$ *are infinite, and the remaining* $r_1$ *generalized singular values are finite.*

Interpretation

The GSVD of $\mathbf{X}_{\mathbf{V}_1}$ and $\mathbf{X}_{\mathbf{V}_2}$ allows to define the generalized eigenvalues and eigenvectors of the pencil $\mathbf{X}'_{\mathbf{V}_2}\mathbf{X}_{\mathbf{V}_2} - \rho\mathbf{X}'_{\mathbf{V}_1}\mathbf{X}_{\mathbf{V}_1}$:

$$\mathbf{H}'\mathbf{X}'_{\mathbf{V}_1}\mathbf{X}_{\mathbf{V}_1}\mathbf{H} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{D}'_1\mathbf{D}_1 \end{pmatrix} \quad \text{and} \quad \mathbf{H}'\mathbf{X}'_{\mathbf{V}_2}\mathbf{X}_{\mathbf{V}_2}\mathbf{H} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{D}'_2\mathbf{D}_2 \end{pmatrix}$$

or equivalently,

$$\mathbf{H}'\mathbf{V}_1\mathbf{H} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_1'\mathbf{D}_1 \end{pmatrix} \quad \text{and} \quad \mathbf{H}'\mathbf{V}_2\mathbf{H} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_2'\mathbf{D}_2 \end{pmatrix}$$

where $\mathbf{H} = \mathbf{Q}\begin{pmatrix} \mathbf{I}_{n-r} & \mathbf{0} \\ \mathbf{0} & \mathbf{R}^{-1} \end{pmatrix}$, $\mathbf{D}_1'\mathbf{D}_1 = \text{diag}(\beta_1^2, \ldots, \beta_r^2)$ and $\mathbf{D}_2'\mathbf{D}_2 = \text{diag}(\alpha_1^2, \ldots, \alpha_r^2)$. Vectorially, it is equivalent to solve the GEP (4.23):

$$\beta_i^2\mathbf{V}_2\mathbf{h}_i = \alpha_i^2\mathbf{V}_1\mathbf{h}_i \Leftrightarrow \mathbf{V}_2\mathbf{h}_i = \rho_i\mathbf{V}_1\mathbf{h}_i, \text{ for } i = 1, \ldots, p. \tag{4.26}$$

where $\rho_i = \alpha_i^2/\beta_i^2$ is real, nonnegative and possibly infinite. The columns of $\mathbf{H}$ are the eigenvectors of $\mathbf{X}_{\mathbf{V}_2}'\mathbf{X}_{\mathbf{V}_2} - \rho\mathbf{X}_{\mathbf{V}_1}'\mathbf{X}_{\mathbf{V}_1}$ or equivalently of $\mathbf{V}_2 - \rho\mathbf{V}_1$, and the "nontrivial" eigenvalues are the squares of the generalized singular values: $\rho_i = \alpha_i^2/\beta_i^2$, for $i = p - r + 1, \ldots, p$. The "trivial" eigenvalues are those corresponding to the leading $p - r$ columns of $\mathbf{H}$, which span the common null space of $\mathbf{X}_{\mathbf{V}_1}'\mathbf{X}_{\mathbf{V}_1}$ and $\mathbf{X}_{\mathbf{V}_2}'\mathbf{X}_{\mathbf{V}_2}$. These eigenvalues are not well defined and are not of interest.

Example

Let us take the same example as previously from Section 4.3.2 defined in (4.5) with $\mathbf{V}_1 = \begin{pmatrix} \boldsymbol{W}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$, $\mathbf{V}_2 = \begin{pmatrix} \boldsymbol{W}_1 & \mathbf{0} \\ \mathbf{0} & \epsilon\boldsymbol{W}_2 \end{pmatrix}$, $\mathbf{V}_1 \in \mathcal{SP}_p$ and $\mathbf{V}_2 \in \mathcal{SP}_p$, with $\text{rank}(\mathbf{V}_1) = r_1 < \text{rank}(\mathbf{V}_2) \leq p$. In addition, $\text{range}(\mathbf{V}_1) = \text{range}(\mathbf{W}_1)$ and $\text{range}(\mathbf{V}_2) = \text{range}(\mathbf{W}_1) \oplus \text{range}(\mathbf{W}_2)$.

Using the GSVD of $\mathbf{X}_{\mathbf{V}_1}$ and $\mathbf{X}_{\mathbf{V}_2}$ to solve the GEP (4.26) leads to investigate four different cases for the direction $\mathbf{h}$:

- if $\mathbf{h} \in \text{range}(\mathbf{V}_1) \cap \text{range}(\mathbf{V}_2) = \text{range}(\mathbf{W}_1)$, then the direction $\mathbf{h}$ is restricted to the subspace spans by $\mathbf{W}_1$ as when we use the Moore-Penrose pseudo-inverse.
- if $\mathbf{h} \in \text{null}(\mathbf{V}_2) - \text{null}(\mathbf{V}_1) = \{0\}$, then no direction $\mathbf{h}$ exists.
- if $\mathbf{h} \in \text{null}(\mathbf{V}_1) - \text{null}(\mathbf{V}_2) = \text{range}(\mathbf{W}_2)$, then $\rho = \infty$ because $\beta^2 = 0$ and so the direction $\mathbf{h}$ can highlight the structure of outlierness contained into the $\text{range}(\mathbf{W}_2)$, which is not the case when we use the Moore-Penrose pseudo-inverse.
- if $\mathbf{h} \in \text{null}(\mathbf{V}_1) \cap \text{null}(\mathbf{V}_2) = \text{null}(\mathbf{V}_2)$, then $\rho$ is a "trivial" eigenvalue and any direction $\mathbf{h} \in \Re^p$ is a solution.

However as explained previously in the beginning of the Section 4.3, only the "non-trivial" eigenvalues, corresponding to the first three cases, are interesting to highlight the structure of the data. More precisely, in this example only the eigenvector $\mathbf{h} \in \text{null}(\mathbf{V}_1) - \text{null}(\mathbf{V}_2) = \text{range}(\mathbf{W}_2)$ associated to the infinite eigenvalue, contains the structure of outlierness of the data. This is clearly visible on the Figure 4.5 which illustrates the projection of our simulated data onto the eigenvectors space. So, using the GSVD outperforms the use of a Moore-Penrose pseudo-inverse because it recovers the structure of outlierness of the data.

To conclude, solving the GEP of $\mathbf{V}_1$ and $\mathbf{V}_2$ through the GSVD of $\mathbf{X}_{\mathbf{V}_1}$ and $\mathbf{X}_{\mathbf{V}_2}$ presents three major advantages. First, it solves the possible singularity issues of $\mathbf{V}_1$ and/or $\mathbf{V}_2$ by searching in all directions, and remains equivalent to the EVP of $\mathbf{V}_1^{-1}\mathbf{V}_2$ if

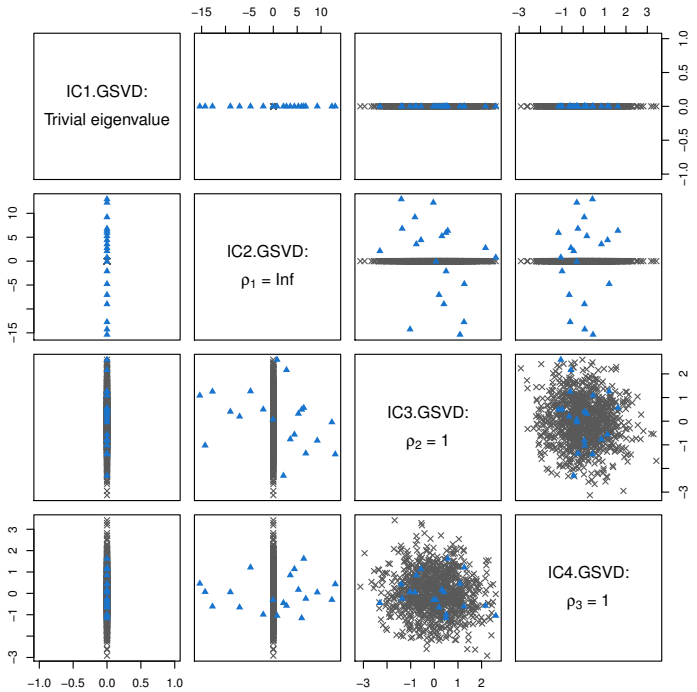



Figure 4.5 – Scatterplot matrix of the IC resulting of using the GSVD of $\mathbf{X}_{\mathbf{V}_1}$ and $\mathbf{X}_{\mathbf{V}_2}$.

the scatter matrices are of full ranks. In addition, it leads that $\mathbf{V}_1$ and $\mathbf{V}_2$ play a symmetric role since the symmetry of the problem is kept. The affine invariance property of the scores continue to be valid in the general case of semi-definite positive scatter matrices. Finally, it is interesting to note that this GSVD procedure is already implemented into the geigen R package, as presented in the next section. However, in practice, it is difficult to define another affine equivariant scatter estimator than the variance-covariance matrix. So, in the next section, we consider a modified version of the scatter matrix based on the fourth moments instead of an affine equivariant estimator such as the Stahel Donoho which requires too complex computations. Note that unfortunately, this modified version of $\text{COV}_4$ is no longer affine equivariant and so the scores are no longer affine invariant.

## 4.4 GSVD on a practical case: an industrial example with collinearity

In this section, we implement the GSVD method in R and analyze its properties on a practical case. We only focus on this approach as it is the one we consider as the best. Indeed, this method presents the same nice characteristics as the classical ICS. In addition, we assume that the data is not in general position. If this assumption is violated, then all the eigenvalues computed are equal and no structure can be highlighted by ICS.

### 4.4.1 Context

This confidential industrial real data set contains 457 devices characterized by 149 numerical measurements. All of them have been sold to a client but one was returned to the manufacturer due to default in use. In the industrial context, this device is labeled as a Customer Quality Incident (CQI). The main objective of the analysis is to illustrate whether it would have been possible to detect this default before selling the device.

All the measurements units are in the range $[10^{-3} : 10^{2}]$, so the data is not standardized. In addition, we also checked that the data is not in general position by estimating its rank (141) through the `Rank` function, from the pracma package (Borchers, 2017).

### 4.4.2 Computation of the two scatter matrices COV and $\text{COV}_4$

In this example, we choose to analyze the variance-covariance matrix COV and the one based on fourth moments, the so-called $\text{COV}_4$. However, on this real world example, the $\text{COV}_4$ scatter matrix is not well-defined since the data is not of full rank. Since some variables are collinear, the variance-covariance matrix COV is singular and so, it is not possible to compute the weights based on the Mahalanobis distance. One possibility, to solve this singularity issue, is to use its generalized inverse to determine the distances of each observation. Since the data is not in general position, the distances are not constant and we obtain an orthogonal equivariant semi-definite positive estimator of the covariance. This is not strictly speaking a true "scatter" as it is not affine equivariant but it still evaluates a multivariate measure of dispersion so we refer at it as if it were.

It is important to note that computing the generalized inverse of the variance-covariance matrix can be computationally tricky, as explained in the Section 5.3.2 of the next chapter. Below we use the well-known `ginv` function from the MASS (Venables and Ripley, 2002) package with default parameters.

### 4.4.3 Implementation of the GSVD procedure

For the implementation of ICS with the GSVD we use the geigen package. It is necessary to express $\mathbf{X}_{\text{COV}}$ and $\mathbf{X}_{\text{COV4}}$ as explained in Section 4.3.4 and the `gsvd` function computes the eigenvalues and the eigenvectors associated. We choose to normalize the resulting scores to obtain the same normalization (4.1) as in the classical ICS, but it is not mandatory. The main advantage of this method is to deal directly with $\mathbf{X}_{\text{COV}}$ and $\mathbf{X}_{\text{COV4}}$ and not on the crossproducts of the matrices which could contain some numerical errors (see Section 5.3.2 for more details).

### 4.4.4 Some results

The idea is to analyze the efficiency of the GSVD method and its properties on a practical case. More specifically, we investigate the effect of exchanging the roles of the two scatter matrices COV and $COV_4$. Indeed, we demonstrate in Section 4.3.4 that considering the combination COV-$COV_4$ or $COV_4$-COV, should have no impact on the results, except the standardization of the resulting scores. In addition, they should be invariant by any affine transformation if the two scatter matrices are affine equivariant (see Proposition 8). However, if the data is not of full rank, the $COV_4$ scatter matrix is only orthogonally equivariant, so the components should be only invariant by an orthogonal transformation. Finally, if the data is standardized then the components are theoretically not invariant in this case. We illustrate these different points on our real example.

Figure 4.6 plots the components associated to the maximum (or minimum) eigenvalue computed with the GSVD of COV-$COV_4$ (or $COV_4$-COV) in different scenarios: on initial data (1st row), on rescaled data by an orthogonal transformation [3] (2nd row) and on standardized data (3rd row). In this example, only one observation is a confirmed anomaly, the so-called "CQI 1", represented by a blue triangle. Obviously, this die is clearly identified as the only outlying observation in all cases.

More specifically, as expected, exchanging the roles of COV and COV4 gives a different scaling of the component since we do not use the same normalization in each case. Moreover, we can notice that the eigenvalues are equal when the data is rescaled by an orthogonal transformation ($\rho_1 = 1.572$ with COV-$COV_4$) but not when the data is standardized ($\rho_1 = 1.998$). So, as anticipated, the first component is invariant by an orthogonal transformation but not by standardization of the data. However, the CQI is clearly identified even in this case the data is standardized, with close values on the first component (18.607 against 18.668). In addition, the five observations with the highest values are ordered in the same way either the data is standardized or not.

In practice, the GSVD method presents good results. Indeed, even if it is not affine invariant anymore, the role of the two scatter matrices are symmetrical and the outlying observation is identified on initial and standardized data. Moreover, the affine invariance is lost because we consider here the $COV_4$ scatter matrix which is only orthogonal equivariant in presence of collinearity. It would be interesting to use an estimator such as the Stahel Donoho estimator. Unfortunately, the existing implementation in R of this estimator does not permit to treat collinear variables.

3. generated by the `randortho` function from the pracma package

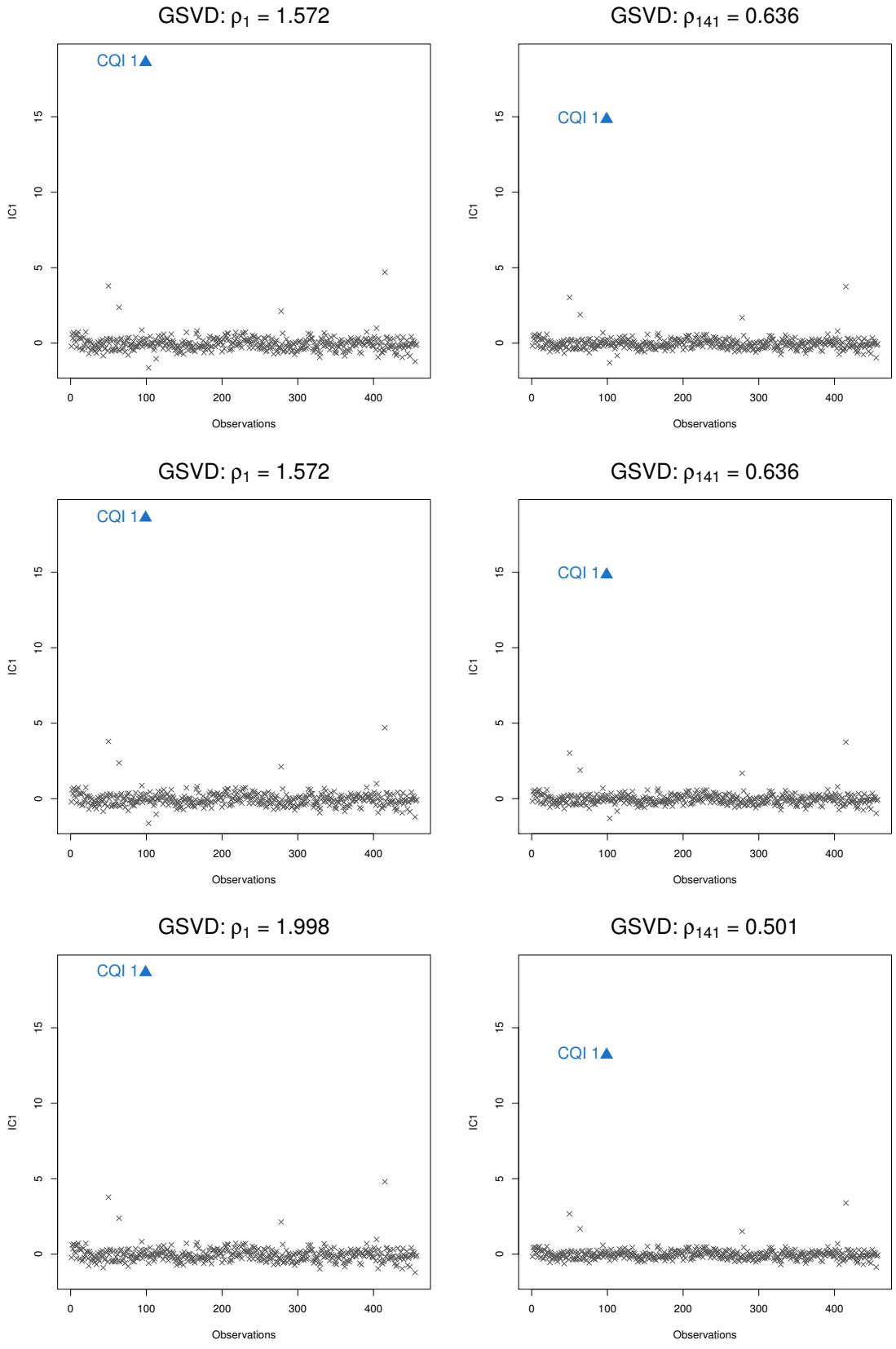



Figure 4.6 – Left: GSVD of COV-$COV_4$. Right: GSVD of $COV_4$-COV. 1st row: on initial data , 2nd row: on data rescaled by an orthogonal transformation, and 3rd row: on standardized data.

## 4.5 Conclusion and perspectives

In this chapter, we focus on data that present some collinearity and/or count more dimensions than observations. This context is really common in practice and it requires a suitable solution. However, we have to assume that the data is not lying in general position. Indeed, as already demonstrated by Tyler (2010), in this case, any scatter matrix is proportional to the variance-covariance matrix and so no structure can be found by comparing them.

First, from a theoretical point of view, we investigate the properties of three different ways of adapting the ICS method to this context. Our analysis shows that using a generalized singular value decomposition (GSVD) allows to keep all the nice properties of the classical ICS. On the contrary, when we use a generalized inverse, the choice of the scatter matrix which is inverted is important and depending on it some structure of the data can not be recovered. Another inconvenient is that even if two affine equivariant scatter matrices can be estimated, then the new coordinates are only invariant by an orthogonal transformation. In addition, preprocessing the data by a dimension reduction turns out to be equivalent to the use of the generalized inverse of the variance-covariance matrix and shares therefore the same disadvantages. More disturbing, this method only solves the issue of considering a singular covariance-matrix as the first scatter $\mathbf{V}_1$. If another scatter matrix is chosen for $\mathbf{V}_1$, of rank lower that the data, then the scatter matrix would still be singular. Further, it is easier to interpret the components resulting from a GSVD than the ones based on a two-steps procedure.

Then, in practice, as most of the scatter estimates are not well-defined in this particular context, we investigate which properties still hold on a real industrial data set with some collinear features. We only focus on the GSVD method, as it is the one with the best theoretical properties. It turns out that the scatter matrix based on the fourth moments is not affine equivariant in this context and so, the method is only invariant by an orthogonal transformation. This is a very common fact in multivariate analysis when the data is not of full rank, and, as proposed by Serfling and Mazumder (2013) we should maybe rather focus on the relative ranking of the outlierness measures of the observations instead of requiring exactly the same value after an affine transformation.

Finally, if the data is not lying in general position, as long as it is possible to assume that the structure of outlierness is contained in a subspace of lower dimension, other solutions are feasible. One idea is to penalize the ICS method, as Witten and Tibshirani (2011) do for LDA for example. This is one of my current topics of interest. It could also be possible to consider other scatter statistics which ideally be well-defined equivariant scatter matrices even in HDLSS context. Further attention needs to be given to this point.

# Chapter 5

# A new outlier detection solution for HDLSS data in an industrial context

*This chapter focuses on the industrial context considered throughout this research work. The main objective was the development of a marketable solution for unsupervised outlier detection in HDLSS data (more variables than production items). Since this particular data are common in the field of aerospace, ippon innovation cooperated with Microchip-Atmel, a cutting-edge company for the electronic components used in the space industry, within the framework of an European project. This partnership allowed us to have access to real data with actual reliability issues to identify in an early stage of the quality control process. In addition, as part of our collaboration, we wrote a conference paper for the $10^{th}$ International Conference on Mathematical Methods in Reliability (MMR) in 2017. This chapter includes a reprint of the paper, which is entitled "High dimensional outlier screening of small dice samples for aerospace IC reliability". It presents in more details the objectives as well as the general idea of the developed algorithm and some results in terms of efficiency compared to other statistical methods. Although the final algorithm is confidential, this chapter describes the different challenges encountered during its development as well as the benefits for the company. Currently, several licenses of the solution are installed at customers' sites in order to identify in advance future reliability problems. This algorithm saves time in detection cycle, and improves the reliability of the space components.*

## Sommaire


**5.1 Context and objectives . . . . . 133**

**5.2 High dimensional outlier screening of small dice samples for Aerospace IC reliability . . . . . 134**

5.2.1 Introduction . . . . . 134

5.2.2 Available Statistical tools . . . . . 135

5.2.3 Advanced Tools for High-Dimensional (HD) data . . . . . 137

5.2.4 Case study on space components . . . . . 138

5.2.5 Conclusion . . . . . 139

**5.3 Challenges in the development . . . . . 140**

5.3.1 Available data . . . . . 140

5.3.2 Numerical errors . . . . . 144

5.3.3 Methodological challenges . . . . . 148

**5.4 Conclusion . . . . . 150**

## 5.1 Context and objectives

This PhD work was financed by the ippon innovation company, headed by François Bergeret, within a "Cifre" agreement with the TSE-R laboratory of Toulouse 1 Capitole. Based in Toulouse (France), this firm develops, among other activities, some advanced outlier detection tools mainly for the automotive field and mostly in cooperation with the universities of Toulouse. Then, their partner Mentor Graphics (part of Siemens), industrializes and commercializes these solutions for the semiconductor industry.

Three years ago, ippon innovation decided to expand its scope to aerospace and so to collaborate with Microchip-Atmel, one of the leading firms in the integrated circuit market for aerospace. This partnership was initiated within the RESIST (RESilent Integrated SysTems) project, co-funded by the European CATRENE (Cluster for Application and Technology Research in Europe on NanoElectronics) program, to promote the resilience of electronics in the avionic, automotive and aerospace industries. The objective of this cooperation was to develop an unsupervised fault detection solution for aerospace products. These electronic devices are characterized by a very large number of measurements compared to the number of manufactured components. This situation is not really frequent in the automotive field, so a new solution should be created specifically for it. However, in statistics, dealing with HDLSS data is really challenging as most of the usual methods for outlier detection require more observations than variables. So, as this topic was worth considering from both a practical and a theoretical perspectives, ippon innovation decided to finance this PhD work.

The main part of my assignment involves the development of a statistical algorithm for unsupervised outlier detection on numerical features in high dimension, i.e. with more variables than units. The new designed solution, named GAT (Good Average Testing) is a confidential complex procedure which includes several stages and a suitable ICS is one of them. The next Section 5.2 is a reprint of a proceedings from a conference on reliability, co-authored with Microchip-Atmel's manufacture which presents the general idea of the algorithm. In addition, based on the real data production of Microchip-Atmel, some results demonstrate the efficiency of GAT compared to other statistical methods. This new procedure is considered as a major progress inside Microchip-Atmel's firm for ensuring the critical aerospace quality levels. Currently, the statistical tool GAT has been fully implemented in the production tool of Microchip-Atmel and it can be used at probe and final test levels. However, the development of a such complex algorithm had to deal with some challenges, mostly numerical, as presented in Section 5.3. Finally, in a concluding part, we discuss the benefits of the theoretical research work for the company.

## 5.2 High dimensional outlier screening of small dice samples for Aerospace IC reliability

This is a reprint of Archimbaud, A., Soual, C., Bergeret, F., D'Alberto, S., Thebault, T., and Bonin, C. (2017d). High dimensional outlier screening of small dice samples for aerospace IC reliability. *The 10th international conference on mathematical methods in reliability, Grenoble, France.*

In this section, the word "parameter" alternatively refers to a test measure on an electronic component to a parameter of a statistical method or a distribution.

**Abstract**
*Aerospace Integrated Circuit (IC) reliability increasingly demands a high level of performance. This article is based on a collaborative project between a production company in the aerospace industry (Microchip-Atmel) and a statistical company (ippon innovation). The main objective is to develop an innovative advanced tool to detect multivariate outliers in small samples, based on measurements of thousands of tests, which is called a high dimensional situation in statistics. After presenting the context and current computational methods used in this industry for the screening of abnormal dice, the article introduces two methods that are designed for dealing with the special case of high- dimensional datasets: the ROBPCA and the GAT algorithms. The two methods are compared in a case study. GAT has a definite advantage over other methods in detecting atypical instances. Finally, the integration of this algorithm with the production tool provides the ability to go back to the real measurements involved in the revealed anomalies. The use of a sound statistical method to adress small samples and high dimension data are needed to detect reliability issues in the space industry.*

***Keywords:*** *Outlier Detection, Multivariate Analysis, High Dimension - Low Sample Size, Integrated Circuits.*

### 5.2.1 Introduction

#### The context

The aerospace market is characterized by a very high level of reliability, and complex qualification steps with the Qualified Manufacturers List & European Space Components Coordination (QML & ESCC) standard specifications. Today a low cost aerospace market is going to emerge, that maintain a high level of reliability. Highrel electronic dice delivery requires the implementation of a new strategy to carry out the qualification steps on the production lots. For the aerospace market, each lot of silicon is probed at the wafer level (a wafer is a manufacturing unit with usually hundreds of dice). Good dice are assembled, and during the final test each die is measured at three temperatures, -55°C, 25°C, 125°C for the three supply voltages, Vcc ±10%. For each die, the testing program generates more than 500-1500 parameter measurements. Each die undergoes a burn-in (Early Failure rate) of 240 hours at 125°C, which is a short-time temperature stress. A maximum drift of 10% in critical data before and after burn-in is acceptable and a maximum percentage defect of

5% is authorized for delivery of the dice to the customer. Some parts of the lot also undergo an HTOL long-time temperature stress (LFR) for up to 2000 hours at 125°C. No rejection is authorized after HTOL and such an accident induces a very high manufacturing cost and a very long delay in delivery to the customer. To avoid this catastrophic scenario, all abnormal dice have to be detected before HTOL. Currently, many sorting and validation checks exist, but they are resource-intensive in silicon, packaging, testers and man-power time. The well-known statistical univariate sorting methods, such as PAT are impractical because they lead to the rejection of almost all dice. Therefore, a plan to develop a statistical tool to deal with the high-dimensionality of aerospace Integrated Circuits (IC) has been proposed inside the RESIST project, co-funded by the European CATRENE program. This project proposes to develop software, methodology, electronic models, IC design solutions and electronic-circuit architectures that are able to promote resilience in electronic devices and applications. It covers domains in which very high reliability is mandatory: avionic, automotive and aerospace domains.

#### The project objective

The project involves the development of a statistical algorithm named GAT (Good Average Testing) that deals with high-dimensional data. Our goals are to: deal with small samples of high-dimensional data and to find optimal projections. At the same time, this new sorting solution must be integrated into the test-management software used by the production company. Then, for wafers that have undergone a long failure-rate aging process and have been found to contain some dice with reliability weaknesses, the statistical tool is used on measurements before burn-in. The algorithm has to detect these abnormal dice by: (i) dealing with high-dimensional aerospace data for: less than one hundred dice and with a few thousand parameters, (ii) being efficient on probe and final test measurements, (iii) returning an outlierness value and the identities of the outlying observations in a few seconds and, finally, (iv) being linked to initial parameters for an easy interpretation. The usual statistical tools used on semi-conductors are first recalled, and then the ROBPCA method and the GAT algorithm are introduced as solutions for dealing with issues of high dimensionality. Finally, some performance results are presented on real samples.

### 5.2.2 Available Statistical tools

#### Univariate Methods

The semiconductor industry uses statistical JEDEC JESD50 standards to detect abnormal dice:

— PAT (Part Average Testing): to identify and reject dice that stand out from the standard test distribution for a given parameter (but are within the test specifica-

tion limits, see Fig.5.1). Usually, based on the Gaussian hypothesis, limits are set at $\mu \pm 6\sigma$. Alternative methods exist for non-Gaussian parameters.

— NNR (Nearest Neighbor Residual): to detect and reject a die when its measured value is statistically different from the average value of its neighboring dice, for a given parameter, see Fig.5.1.

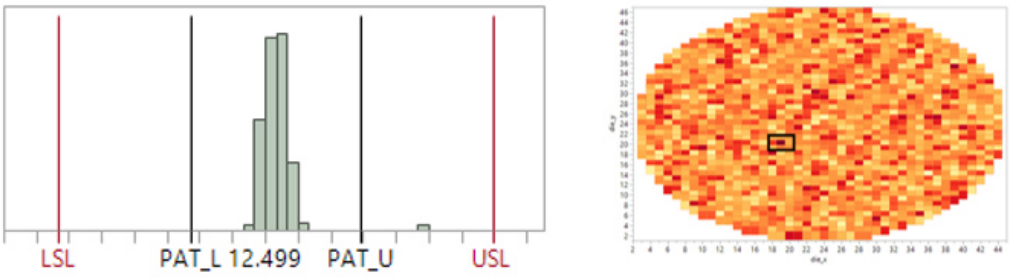



Figure 5.1 – Left: PAT_L and PAT_U are PAT limits calculated at $\mu \pm 6\sigma$, LSL and USL are the specification limits. Right: Example of NNR, the dark red die in the center of the black rectangle is different from its neighbors in a leakage test.

— GPAT (Geographic Part Average Testing) or GDBC (Good Die in Bad Cluster): the idea is to identify and reject a good die (which passed all tests) located in a cluster of bad dice.

A good summary of these methods can be found in Moreno-Lizaranzu and Cuesta (2013). These methods enable some improvements but their costs quickly become prohibitive. Except for GPAT, applying each method on each test independently multiplies the false alarm rate by almost $p$ and only reveals univariate outliers.

### Multivariate Methods: Hotelling's $T^2$ statistic

A widely used method in statistical process control is Hotelling's $T^2$ statistic (see Jensen et al. (2007)). This statistic is equivalent to the squared Mahalanobis distance, which computes the distance of each die from the center of the distribution. Formally, let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be $n$ observations characterized by $p$ quantitative variables. Each $\mathbf{x}_i$ is assumed to be generated by a multivariate normal distribution with a location parameter $\boldsymbol{\mu}$ and a positive definite variance-covariance scatter matrix $\boldsymbol{\Sigma}$: $\mathbf{x}_i \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Usually the parameters are estimated by the empirical mean $\boldsymbol{\mu}_n$ and $\boldsymbol{\Sigma}_n = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu}_n)(\mathbf{x}_i - \boldsymbol{\mu}_n)^t$ (or a robust version).

$$T^2_{\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n}(\mathbf{x}_i) = MD^2_{\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n}(\mathbf{x}_i) = (\mathbf{x}_i - \boldsymbol{\mu}_n)^t \boldsymbol{\Sigma}_n^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_n) \tag{5.1}$$

Then, all distances are compared to a quantile of a $\chi^2$ distribution at the level $\alpha\%$. However, this affine equivariant statistic is neither able to address singularity issues nor able to find the relevant subspace for the analysis.

### 5.2.3 Advanced Tools for High-Dimensional (HD) data

**ROBPCA: ROBust Principal Component Analysis**

Hubert et al. (2005) introduce the ROBPCA method, which is designed for small samples of high-dimensional data. We propose to use this orthogonally-invariant method, which is different from a standard robust PCA method, for detecting outlying dice in the aerospace context. The key steps are:

(i) Reducing the initial dimension to at most $n-1$ without losing any information. An affine transformation is performed using singular value decomposition: $X_n - 1_n\mu_n^t = U\Delta V^t$, where $U^tU = V^tV = I_r$ are semi-orthogonal matrices and $r = \text{rank}(X_n - 1_n\mu_n^t) \leq n-1$. Then the initial variables are projected: $Y_n = (X_n - 1_n\mu_n^t)V^t = U\Delta$.

(ii) Searching for a subspace of smaller dimension containing only the $h$ observations with the lowest modified Stahel-Donoho orthogonally-invariant outlierness value based on Projection Pursuit (PP) techniques. A spectral decomposition of the new covariance matrix allows the determination of the dimension $k$ for consideration in further analysis. Then the observations are projected on this subspace leading to $\tilde{X}_n$.

(iii) Performing a spectral decomposition of a robust reweighted MCD (Minimum Covariance Determinant) scatter estimator of $\tilde{X}$: $S_{2,n} = PDP^t$, where $D$ is the diagonal matrix with the eigenvalues $\gamma_1 \geq \cdots \geq \gamma_p$ of $S_{2,n}$, and the columns of the (orthonormal) matrix $P$ contain the corresponding eigenvectors. Now the potential outliers can be revealed by a diagnostic plot based on two orthogonally equivariant distances: the Score Distance (SD) and the Orthogonal Distance (OD). For each observation $\tilde{x}_i$ and $k$ components:

$$\begin{aligned} \text{SD}^2_{S_{2,n}}(\tilde{x}_i, k) &= ||\text{diag}(\frac{1}{\sqrt{\gamma_1}}, \ldots, \frac{1}{\sqrt{\gamma_k}})P_k^t(\tilde{x}_i - \mu_{2,n})||^2 \\ \text{OD}^2_{S_{2,n}}(\tilde{x}_i, k) &= ||(I_d - P_kP_k^t)(\tilde{x}_i - \mu_{2,n})||^2 \end{aligned} \tag{5.2}$$

with $||X||^2 = X^tX$. An observation $\tilde{x}_i$ is identified as an outlier if:

$$\text{SD}(\tilde{x}_i) > c_{p,1-\alpha} \text{ and/or } \text{OD}(\tilde{x}_i) > (\text{med}_{\text{OD}} + \text{MAD}_{\text{OD}}\ z_{1-\alpha})^{3/2} \tag{5.3}$$

where $c_{p,1-\alpha}$ is the $(1-\alpha)^{th}$ quantile of a $\chi^2_p$ distribution and $z_{1-\alpha}$ is the $(1-\alpha)^{th}$ quantile of a Gaussian distribution.

**GAT: Good Average Testing**

The GAT algorithm is based on:

(i) Pre-processing of the data set (in collaboration with Microchip-Atmel) to control potential noisy parameters and univariate outliers.

(ii) Reducing the initial dimension as in ROBPCA leading to $Y_n$.

(iii) Performing a multivariate analysis optimized for outlier detection based on the analysis of two symmetric and positive definite scatter estimates: $V_1(Y_n)$ and $V_2(Y_n)$. This involves looking for the $p \times p$ matrix $B(Y_n)$ and the diagonal matrix $D(Y_n)$ such that:

$$V_1(Y_n)^{-1}V_2(Y_n)B(Y_n)' = B(Y_n)'D(Y_n) \tag{5.4}$$

s.t $B(Y_n)V_1(Y_n)B'(Y_n) = I_p$ and $B(Y_n)V_2(Y_n)B'(Y_n) = D(Y_n)$.
$V_1(Y_n) = (\sum_{i=1}^{n} w(||y_i - \mu_n||^2_{V_{1,n}^{-1}})(y_i - \mu_n)(y_i - \mu_n)^t)/(\sum_{i=1}^{n} w(||y_i - \mu_n||^2_{V_{1,n}^{-1}})$,
$||X||^2_{V_{1,n}^{-1}} = X^t V_{1,n}^{-1} X$, $w$ is a positive decreasing function and $V_2(Y_n) = \Sigma_n$. $D(Y_n)$ contains the eigenvalues of $V_1(Y_n)^{-1}V_2(Y_n)$ in decreasing order and $B(Y_n) = (b_1, \ldots, b_p)'$ the corresponding eigenvectors as its rows.
We denote by $Z_{n,k}$ the projection of the data set onto the $k$ selected eigenvector directions and an outlierness value is computed for each observation $y_i$. Based on $k$ components, high values reveal possible outliers:

$$GAT_{measure}(y_i, k) = ||g(Z_{n,k})||^2 \tag{5.5}$$

where $g$ is a transformation function. A test is performed to identify outliers at a given risk level of $\alpha\%$. Lastly, parameters that contribute most to the abnormal behavior of the dice are returned.

### 5.2.4 Case study on space components

#### An example on one sample

First, we focus on a product characterized by 2026 parameters (temperature, voltages, frequency). After having undergone a long failure rate aging process, it appears that two dice have shown reliability issues. The goal is to detect these rejects at an early stage of final testing and before the EFR step. Applying the pre-processing step leads to 462 parameters and 57 dice.
The PAT method fails to detect the rejects on any of the parameters. A robust version identifies them as outlying on four (resp. three) parameters but with a too-costly rejection rate of 50%. The GAT algorithm detects the two rejects while only mislabeling three dice (see Fig.5.2). The ROBPCA algorithm (in the R package *rrcov*), applied on standardized measurements, fails to detect the second reject and it mislabels more than twice as many dice as GAT.
The GAT algorithm offers the possibility of easily identifying the parameters that explain the bad behavior of the dice (see Fig.5.3). Based on these contributing parameters, the engineers can quickly determine whether there is a reliability issue. This is a very helpful tool, as it allows the reduction of the false alarm-rate and saves engineers a lot of time.

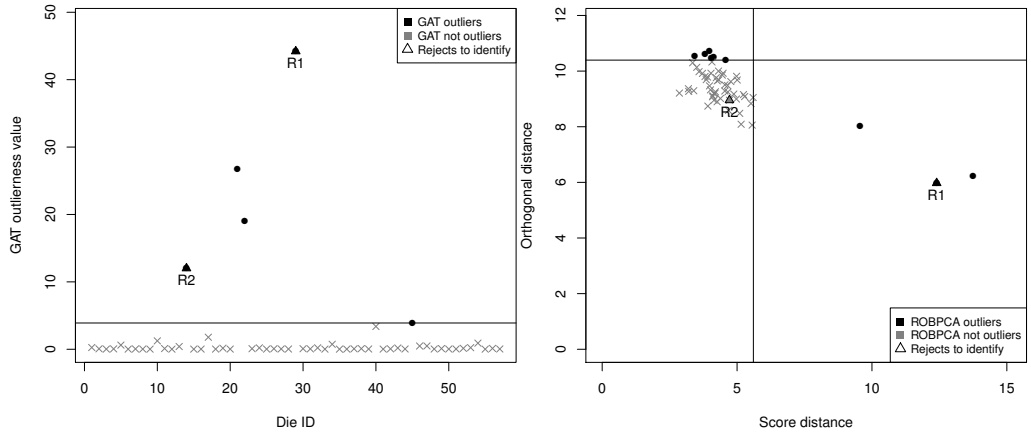



Figure 5.2 – Left: GAT outlierness value (performed at a risk level of 10%). Right: SD vs OD distances returned by ROBPCA (with cut-offs at a level of 5% for each).

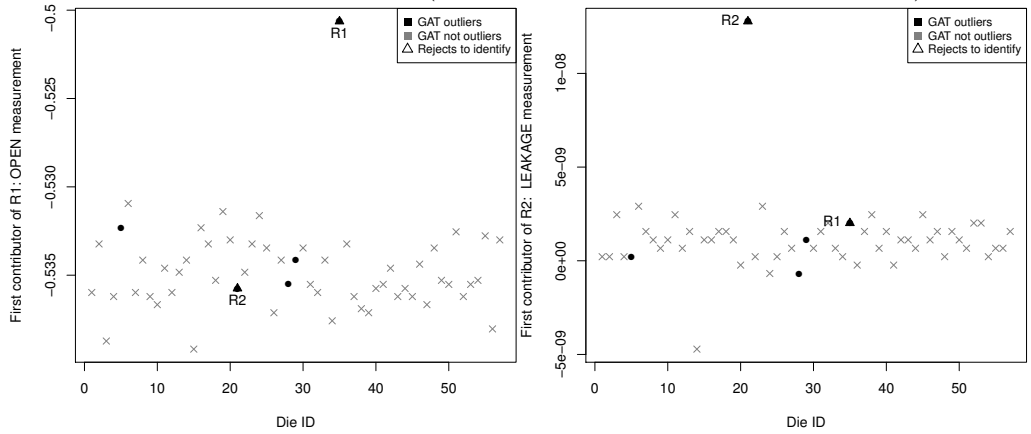



Figure 5.3 – 1*st* contributor measurements (from the initial data set) for the R1 die (left) and the R2 die (right).

### A global comparison on several samples

A more general analysis is conducted on several insertions for five lots at final testing and one lot at the probe level. The number of tests varies from a few hundred to a few thousand, and the sample size varies from 20 to 200. The goal is to identify the 15 dice with reliability issues with a rejection rate of less than 10%. First, the PAT method detects two rejects but not in an early stage of the qualification steps. Then, the ROBPCA algorithm achieves the detection of three rejects out of 15, which leads to a performance rate of 20% (the usual detection rate in this context). Finally, the GAT algorithm outperforms ROBPCA in terms of efficiency by revealing nine rejects out of 15, which leads to a performance rate of 60%. All algorithms finish running in a few seconds.

## 5.2.5 Conclusion

To ensure the high level of reliability required for the aerospace market, the electronic dice are being subjected to increased testing, leading to more parameters than observations. Even though new methods for high-dimensional data have been developed by statisticians, they have not yet been incorporated into the test-management processes of

the industrial sector. Here, we propose the application of the ROBPCA method, which is based on the ideas of dimension reduction and robust statistics with a Projection Pursuit approach. Then, we present the GAT algorithm, which is designed for low-volume products with large numbers of tests (space components, defense products, medical devices, etc). In our evaluation study, GAT outperforms the ROBPCA method by leading to a detection rate of more than 50%. However, detecting reliability issues in such noisy data sets is still very challenging. Therefor, to reduce the false-alarm rate, we have developed, in collaboration with Microchip-Atmel, a pre-processing procedure for the data set to control potentially noisy parameters. Then, the experts analyze the contributors to the outlying dice to decide if the dice are reliability rejects. To conclude, the RESIST project, with the collaboration of experts in aerospace and in statistics, has led to the development of the GAT algorithm, which fully meets the expectations. The next step is to test it on more wafers at the probe level as soon as the data are available and to continue to improve the detection and the ability to deal with missing values.

**Acknowledgments**
This publication is based upon work supported by the Catrene (Cluster for Application and Technology research in Europe on nanoElectronics) European program under RESIST project n° CT217. Many thanks to B. Azalbert, E. Facchi, M. Gomez, C. Legrand, S. Malbranche, A. Ruiz-Gazen, J. Savournin, S. Tapie and B. Thomas for their technical advices and actions in project advancement.

## 5.3 Challenges in the development

As mentioned in the previous section, the development of GAT fully meets its expectations and this solution is already integrated in Microchip-Atmel's quality control tool. However, from a practical point of view, its development had to deal with some challenges that we introduce here. First, one of the main difficulties was to have access to real data to evaluate the efficiency of the developed method. We briefly describe below the kind of data sets we used for this validation step. Then, working with such complex aerospace integrated circuits, requires to deal with a high level of value accuracy. However, software computations can suffer from numerical errors. The main causes of this instability are presented in Section 5.3.2, though the solutions used in practice are confidential. Finally, we present shortly the main challenges in the development of GAT from a methodological point of view.

### 5.3.1 Available data

At the beginning of my thesis in 2015, the RESIST project was just initiated. The insertion of GAT tool in the production tool of Microchip-Atmel, has necessitated a complete overhaul of the data collection through the probe and the test programs. Some major

modifications have been done to get analog measurements instead of pass/fail information. At beginning, no reliability data was available in Microchip-Atmel because no high temperature stress was available on the new datasets. Indeed, in aerospace, the production of electronic die is a process of several months. Including also the testing steps at the probe and the final test levels, it takes months before having useable data. So, we decided first to consider the significant database of ippon innovation from the semiconductor industry (mostly in automotive) to start the statistical research. With my colleague Carole Soual, statistician engineer at ippon innovation, we conceived a new structure of our data bank containing the data sets together with the solutions proposed by the company and their performance results. As ippon innovation develops unsupervised methods, it was important to keep apart which electronic die fails from the data set. In addition, because of the diversity in the sources of the data sets, a substantial work had to be done to automatically format them in the same way. In the end, with this new architecture, benchmarking different methods on all the data sets becomes possible quite easily.

**Data from the semiconductor industry**

This data bank includes 201 data files coming from different leading companies of the semiconductor industry between 2009 and 2017. All this real world data sets contain some confirmed reliability issues, so-called CQI (Customer Quality Incident). The available measurements are made at the probe and/or final test (FT) levels, i.e. before and/or after dice are assembled. The Table 5.1 briefly describes the characteristics of the data sets. It can be noted that almost two-thirds of the files concern measures at the probe level. Indeed, in an industrial context, identifying reliability issues at the probe level costs less than at the FT level. However, the number of confirmed CQI is nearly the same at the two levels, between one and five dice, which corresponds to less than 2% of the all observations by file.

| Type | Number | Number of CQI | | | Dimension Standard | | | HD | All |
|---|---|---|---|---|---|---|---|---|---|
| | of files | Min | Mean | Max | S1 | S2 | S3 | $n < p$ | |
| Probe | 144 | 1 | 1.1 | 3 | 20.9% | 39.3% | 6.0% | 5.4% | 71.6% |
| FT | 57 | 1 | 1.7 | 5 | 9.5% | 3.5% | 7.4% | 8.0% | 28.4% |
| All | 201 | 1 | 1.3 | 5 | 30.4% | 42.8% | 13.4% | 13.4% | 100.0% |

Table 5.1 – Description of the ippon's database from the semiconductor industry.

The files are clustered depending on their dimensions: S1, S2 and S3 if the data is in standard dimension with $n > p$ or in HD if $p > n$. More precisely, the first three groups are: S1 if $n/p > 5$, S2 if $5 > n/p > 2$ and S3 if $2 > n/p > 1$. The first class S1 was created based on a piece of advice from Rousseeuw and Van Zomeren (1990) to avoid the curse of dimensionality when computing robust estimators as the MCD. However, in practice the `covMcd` function (from the robustbase R package), warns the user about this phenomenon only when $n < 2p$. Hence, we decided to create the S2 class of datasets with $5 > n/p > 2$.

The boxplots in Figure 5.4 give a more precise idea about the homogeneity of the data sets in terms of number of variables and observations. In general the number of dimensions is at least of a few hundreds. So, one of the first challenges of the outlier detection methods, is to deal with this amount of variables within a few seconds. Second, it is well-known that in real world data sets, the variables are often collinear. Indeed, it appears that less than one-half of the files in S1 have no collinearity issue, i.e. for which the empirical variance-covariance matrix is not singular. In total, 86% of the data sets are concerned by the context considered in Chapter 4 of singular scatter matrices so, this situation is really common in practice. In addition, we can notice that around 30 files already contain more variables than units. Actually, with the exponential increase in the number of measurements on electronic components these last few years, the semiconductor industry is as concerned by dealing with HDLSS data as the aerospace industry is. The major differences between the two areas mainly lie in the complexity and the number of manufactured products. So, it appears that this database is legit to benchmark methods and test new solutions.

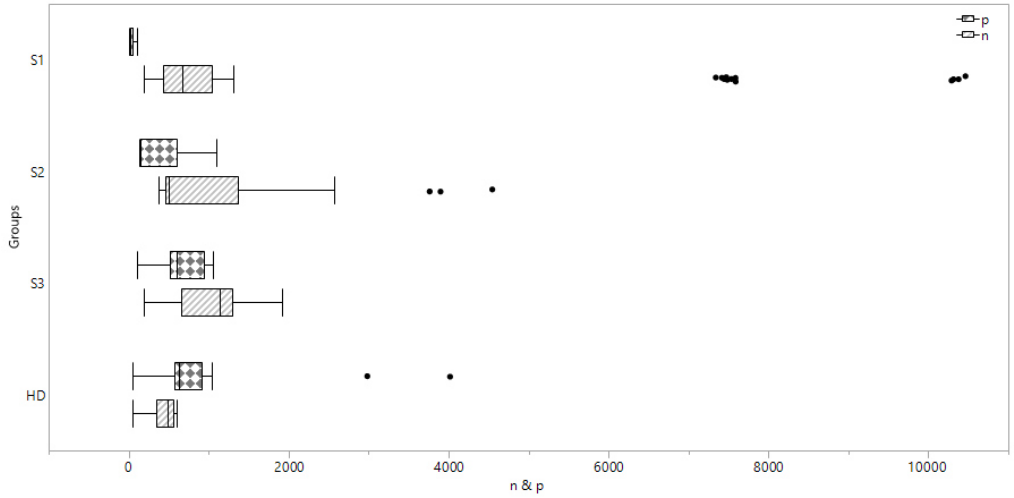



Figure 5.4 – Boxplots of the number of variables ($p$) and the number of observations ($n$) by classes of dimension.

For your information, the real data set HTP presented in Sections 1.2.5, 2.6.3 and 3.4.2 come from this database. This is also the case for the real Reliability data set introduced in Section 3.4.3 and the industrial data set in Section 4.4.

### Data from the spatial industry

At the beginning of the partnership, Microchip-Atmel offered us to consider former lots of dice. But, after some months of analysis, it appears that these lots were not completely relevant compared to the new products they designed. Nevertheless, the extensive study of these data sets allowed us to better understand their products and their qualification steps. Mostly, it was crucial to find out when our solution will be applied. Indeed, depending

on the stage of the process, the data sets have not the same characteristics and they can contain univariate outliers or missing values for example.

After settling this part, we noticed that some measurements seem to have too much variation on the reference units and affect the results of the analysis. These variations are related to the measurement system and can be evaluated through an uncertainty analysis, called a "gage R&R", in terms of Reproducibility and Repeatability. As an expert of the industrial context, ippon innovation worked with Microchip-Atmel to measure this source of variation and informed them on the level at which the statistical analysis could be influenced. Then, as explained in Section 5.2.3, Microchip-Atmel did a considerable effort to evaluate this measurement variation and to remove any test which could impact our analysis. However, integrating these new steps has been possible only for new products as they required additional measures. This extended the time before we had usable and cleaned data sets.

Table 5.2 presents the six lots of electronic components considered in the proceedings presented in Section 5.2. As we can notice, there is only one lot at the probe level and the other five are at the FT step. In addition, the measurements of the dice are not necessarily available at the three temperatures (hot, cold and room) because some of the lots are older and the filtering step was not possible. The reliability rejects considered here were identified by Microchip-Atmel at a later stage of the quality control process.

| Lots | Received | Number of reliability rejects | Probe | Final Test | | |
|---|---|---|---|---|---|---|
| | | | | 125°C | -55°C | 25°C |
| Lot 1 | 09/2016 | 3 ($\approx 2\%$) | | | 171×50 | |
| Lot 2 | 09/2016 | 2 ($\approx 2\%$) | | | | 110×4747 |
| Lot 3 | 09/2016 | 2 ($\approx 4\%$) | | | 57×462 | 51×347 |
| Lot 4 | 10/2016 | 1 ($\approx 2\%$) | | 62×585 | 61×434 | 60×434 |
| Lot 5 | 10/2016 | 3 ($\approx 5\%$) | | 55×1246 | 55×946 | 55×737 |
| Lot 6 | 01/2017 | 3 ($\approx 11\%$) | 28×985 | | | |

Table 5.2 – Description of the cleaned files of measurements from the aerospace industry and their dimensions (number of observations × number of variables).

First, it is interesting to note that even after the filtering step, almost all the data sets contain more variables than units, except for the first lot. The reason for this one is that it is an old product compared to the others. Then, compared to the automotive industry, we can notice that the number of reliability rejects to detect is relatively higher, from two to eleven percent at the probe level. This is mostly due to the fact that the number of observations is smaller than in the previous context. Indeed, in the aerospace, only a few decade of dice are manufactured because this is the quantity needed in the spatial industry. Moreover, we can notice that the lots have really different number of observations and dimensions since the integrated circuits are not necessarily designed for the same product. Consequently, the method we developed can deal with HDLSS data as well as data in standard dimension ($n > p$), with dimensions between fifty to a few thousand and really

small sample sizes. All these objectives become even more challenging in practice, mostly because of numerical errors which causes the instability of the algorithm.

### 5.3.2 Numerical errors

We consider that an algorithm is unstable if the results of the analysis are not the same by permuting randomly rows or columns whereas it should be. Contrary to the computation of MCD estimators for example, in our case, the algorithm does not rely on some random generator and so the results should be exactly the same if we permute rows or columns. However, we found out that the oultyingness indexes of the observations computed with our first version of the algorithm, were not necessary the same whether we performed it on the initial data set or after permuting rows or columns. More annoying, the ranks of the observations ordered by these outlierness indexes were not always consistent. As an illustration, the Figure 5.5 plots the rejection rates [1] needed to identify the CQIs in the 27 data sets in HDLSS context from the automotive industry. It appears that the algorithm is unstable by rows or columns permutations in 60% of the cases. This instability is noteworthy as it can be really important. For example, the file 1 has an initial rejection rate of approximately 10% which increases to 80% when some rows are permuted. Still, the columns permutations are also meaningful to consider as it can impact the results as for the file 8. In addition, this instability was even more severe with the aerospace products.

After further analysis, it occurs that the preprocessing step of dimension reduction was the one which causes the instability of the whole algorithm. After checking that this problem still remains using other programming languages, we tried to understand the reasons of such an instability. We found out that the floating point arithmetic and the condition number were of great importance when we perform a dimension reduction. We describe below the reasons why.

#### Floating point arithmetic

The instability of an algorithm is due to the floating point arithmetic. So, we had to take an interest in the field of numerical analysis. As we are not an expert of this area, this section only introduces basic notions on numerical computations, based on the books from Trefethen and Bau III (1997) and Ueberhuber (2012a,b).

First, it is important to realize that in programming, the numbers are represented in a binary format, usually with the IEEE 754 standard. This standard defines a format which includes the representation of the floating-point numbers and special values as the infinite or the *NaN* (Not a Number), together with a definition of the specific arithmetic operations for these numbers. In fact, these digital representations suffer from some limitations and so the usual arithmetic operations can not be applied as in theory. Indeed, as the numbers

1. Rejection rate: percentage of observations to discard for identifying the CQI as an outlier.

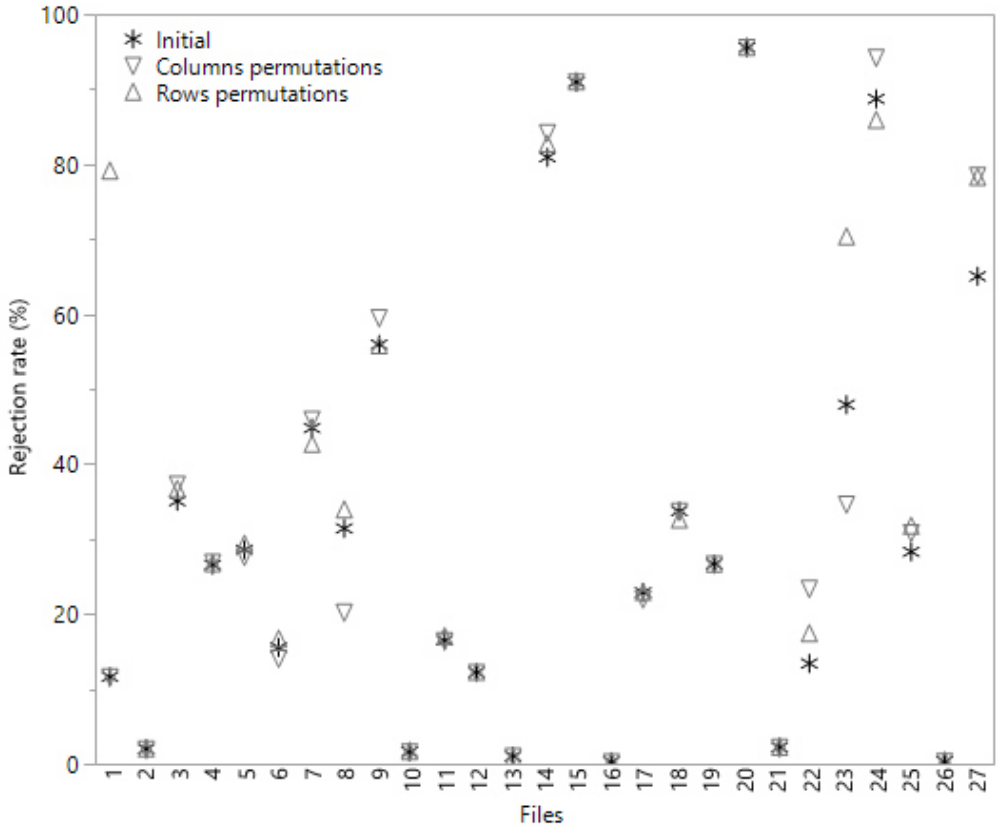



Figure 5.5 – Rejection rates with the first version of our algorithm applied in HD files from the automotive industry, on initial data sets and after randomly permuting rows or columns (with the `sample` R function).

are represented by a finite number of bits, it can happen that it is not possible to have an exact representation of it. In this case, some specific rules are necessary to define the common arithmetic operations to ensure the accuracy of the result. However, this precision is only guaranteed up to a "machine epsilon". For example, in R, this value $\epsilon_m$ is around $2.2e^{-16}$ with a 64 bits double digits representation, i.e. that, as long as we take a number $|x| > \epsilon_m$, then $1 + x \neq 1$, but if $|x| < \epsilon_m$ then $1 + x = 1$.

As a consequence, considering a high level of accuracy can be challenging in programming and can make unstable basic arithmetic operations, like sums or matrix computations. Thus, the commutative and associative properties of the sum are no longer guaranteed in practice. So, this is one of the instability causes when analyzing such complex electronic components. Indeed, to ensure the high level of reliability required in this domain, the integrated circuits suffered a lot of different tests with strong value accuracy, from the pico ($10^{-12}$) or the femto ($10^{-15}$) to the mega ($10^{6}$) or the tera ($10^{12}$). In electronics, the features are measured in diverse units but the pico is really common since the electrical capacity is valued in picofarad in the International System (SI). However, it is clear that considering these disparate units is quite challenging in computations.

### Condition number

In addition, the diversity of units impacts the condition number of a matrix $\mathbf{A}$, denoted by $\kappa(\mathbf{A})$. This characteristic gives an idea about the maximum inaccuracy that may result in computations. If the condition number is small, say around $10^k$ with $k = 1, 2, 3$, then the matrix is well-conditioned. If $\kappa(\mathbf{A})$ is much higher, with values of $k$ higher than 6, then the matrix is really ill-conditioned. In this case, inverting this matrix is a challenging issue, and we can lose up to $k$ digits of precision when doing so. In addition, this estimation of accuracy does not even take into account the loss from numerical instabilities due to floating-point arithmetic and round-off errors presented in the previous section.

More formally, the condition number of a matrix is relative to the choice of a norm:

$$\kappa(\mathbf{A}) = ||\mathbf{A}||\ ||\mathbf{A}^{-1}||$$

Typically, it is the 2-norm which is considered and thus:

$$\kappa(\mathbf{A}) = \frac{\rho_{max}}{\rho_{min}}$$

with $\rho_{max}$ the maximal singular value of $\mathbf{A}$ and $\rho_{min}$ the minimal value. If the matrix is singular, the condition number $\kappa(\mathbf{A}) = \infty$. However, in this case Trefethen and Bau III (1997) explains that we can use the pseudo-inverse of $\mathbf{A}$, since the matrix is not invertible. In this case, the minimal singular value is the smallest one which is not equal to zero.

### Singular Value Decomposition (SVD)

In practice, we consider only HDLSS data sets which are not of full rank by definition. The GAT algorithm integrates a singular value decomposition as preprocessing to reduce the initial dimension to the value of the rank of the data. With this reduction dimension technique, the new coordinates are now of full rank. So, the challenge is to estimate correctly the rank of the initial data and thus to determine which singular eigenvalues are non-zero. Indeed, as we just explained, if the minimal non-zero value and the maximal one are too different, then the matrix is ill-conditioned and so the singular vectors might not be well estimated. In addition of this inaccuracy, it is usually difficult to know which singular values to consider. Actually, because of round-off errors due to floating-point arithmetic, no values are exactly zero, but approximately $10^{-10}$, $10^{-13}$, $10^{-16}$, $10^{-20}$. So the challenge is to decide when we can consider these values to be "true" zeros.

In fact, this is a tricky question in programming and a lot of rules are used in practice. However, it seems to exist a consensus about the use of a relative criterion, to decide when a singular value $\rho_i$ should be considered as zero:

$$\rho_i < \rho_{max} * tolerance$$

for $i = 1, \ldots, p$ and $\rho_i$ the i$^{th}$ singular value of the $n \times p$ matrix $\mathbf{A}$. There is no general agreement on the tolerance value and it mostly depends on the algorithm used to compute the singular values. Indeed, in theory, the singular values of a matrix $\mathbf{A}$ are the square

roots of the eigenvalues of $\mathbf{A}'\mathbf{A}$ or $\mathbf{A}\mathbf{A}'$. So, depending on the dimensions of the data, it is usual to calculate the spectral decomposition of $\mathbf{A}$ first and then deducing its singular value decomposition. In this case, the tolerance value considered can also be different as we study now the square of the previous values. Table 5.3 summarizes some of the values used in some R functions, usually based on the epsilon machine $\epsilon_m =$ `.Machine$double.eps` $\approx 2.2e^{-16}$.

| Tolerance | Purpose | R functions (Packages) |
|---|---|---|
| $\sqrt{\epsilon_m}$ | singular values | `ginv` (MASS) |
| $\epsilon_m^{2/3}$ | singular values | `pinv` (pracma) |
| $10^{-6}$ | singular values | `WhitenSVD` (REPPlab) |
| $\max(n,p)\epsilon_m$ | singular values | `rankMM` (rrcov), |
| | | `Rank` (pracma) |
| $n\epsilon_m$ | eigenvalues | `kernelEVD`, `PcaHubert` (rrcov) |
| | | `ClassPC` (robustbase) |

Table 5.3 – Tolerance values used by default in some R functions.

So, the choice of the tolerance value is really a critical step as it can have a huge impact on the accuracy of an SVD. Indeed, if we consider too many singular values as non-zero, then the singular vectors associated to them are not accurate. In addition, since they estimate a part of the null space of the data, we should not consider them at all. However, if the tolerance value is too restrictive, then the rank of the data is under-evaluated. In this case, the original data are projected onto a subspace of smaller dimension than it should be and some information about the structure of the data may be lost.

#### Example on the effect of the tolerance value

Let us return to our simulated example in Chapter 4.3.1, with $\mathbf{W}_1 = \text{diag}(10^4, 10^2)$, $\mathbf{W}_2 = \text{diag}(1, 0)$ and a percentage of outliers of $\alpha = 2\%$. We choose $\mathbf{V}_2$ as the theoretical variance-covariance matrix, so the rank of $\mathbf{X}$ can be determined based on its eigenvalues. Since $\mathbf{V}_2 = \begin{pmatrix} \boldsymbol{W}_1 & \mathbf{0} \\ \mathbf{0} & \alpha\boldsymbol{W}_2 \end{pmatrix}$, its eigenvalues are: $10^8$, $10^4$, $\alpha^2$ and 0, with $0 < \alpha < 0.5$. Hence, compared to the maximal eigenvalue, the $\alpha^2 < 0.25$ eigenvalue can be considered as a "true" zero if we take a tolerance value of $10^{-8}$ for example. In this case, the rank of $\mathbf{X}$ is evaluated as two instead of three. Projecting the data only onto the subspace spans by these two first components eliminates the structure of the data associated with the third component on which the outlying observations lie. None of the outlier detection methods can then identify the outliers anymore.

#### Example on the effect of the algorithm to compute the SVD

In theory, the singular values of a matrix $\mathbf{X}$ are the square roots of the eigenvalues of $\mathbf{X}'\mathbf{X}$ so, the variance-covariance matrix $\text{COV}(\mathbf{X})$ and the matrix $\mathbf{X}$ should have the same rank. However, in practice, because of the numerical errors, the ranks can be really different. Considering the automotive data sets, the Figure 5.6 illustrates this by plotting

the estimated ranks of the matrix $\mathbf{X}$ against the ones of COV($\mathbf{X}$), by the `Rank` function from the pracma package.

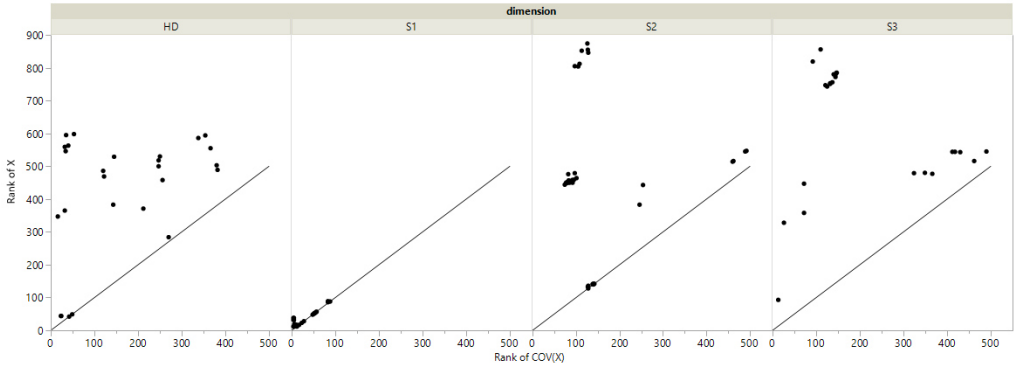



Figure 5.6 – Estimated ranks of the data sets from the automotive industry by dimensions.

### 5.3.3 Methodological challenges

The first challenge to take care was the singularity issue of the HDLSS data sets. Theoretically, this problem was solved by using an SVD to reduce the dimension of the data as already explained. However, we showed that it is not easy to obtain accurate new coordinates because of precision errors. In addition, this technique is not invariant by affine transformation (see Section 4.3), so depending on the scale of the initial data it may be interesting to apply first a rescaling.

#### Preprocessing the data

In our context, we observed that the data sets are in disparate units. This is a major issue since that means the unit values of the measurements impact the results of the singular value decomposition. Together with the numerical errors, computing stable new components is really challenging.

A lot of solutions exist to improve the accuracy of the decomposition algorithms. One of the main idea is to precondition the initial data to simplify the following computations. Chen (2005) presents some of these methods but usually they are already integrated to the so-called "preconditioned solvers". Moreover, a consequent literature focuses on computing accurate eigenvalues or singular values, see among others Parlett (1998), Golub and Van Loan (1996); Golub and Van der Vorst (2000), Higham (2000, 2002), Stewart (1998), Press et al. (1996), DO Q (2012). Nevertheless, it is also possible to improve the accuracy of computations by some "equilibratium" process (see among others Livne and Golub (2004) and Van der Sluis (1969)). For example, just standardizing the data first can reduce the instability of the algorithm.

So, we conduct an extensive study to investigate the impact of different rescaling of the data, in terms of stability and efficiency. In particular, we try to standardize the data

or to rescale all the values between -1 and 1, as it is usual in machine learning algorithms. We also transform the values in the International System (SI) to not take into account the units afterward. The idea is to make "comparable" values of different kind of tests which may be in different scale of units. For example, if one measure values 5 kV (kilo-volts) and another 2.3 pF (pico-farads), they are encoded by $5e^3$ and $2.3e^{-12}$ in the data set. They can be complicated to compare, so we decide to transform these values into 5 and 2.3, to take off the influence of the units. The results of the all analysis are confidential but it appeared that rescaling the data improved the stability of the algorithm.

#### Adapted ICS and selection of components

As presented in Section 5.2.3, the GAT algorithm includes an adapted ICS step. This is a modified version of the classical ICS as it occurs that the data sets are sometimes in general position [2], contrary to the context considered in Chapter 4. In fact, the automotive data sets are usually not in general position and so the methods presented in Chapter 4 can be applied but this is not the case for the spatial data sets.

In addition, as in the classical method, only the relevant components have to be selected. Some theoretical research work was developed for some particular models, as presented in Section 2.8.2. However, as we modified the ICS method, an additional empirical analysis was conducted to develop a new way of consistently selecting the relevant components in practice. This method is not presented as it is confidential.

#### Interpretation of the outliers

In the industrial context, the understanding of the outlying behavior is crucial. Indeed, as mentioned in Section 1.2.4, if we identify some outlying dice, we have to give some explanations to the tests engineers who are the ones who decide to discard these dice from the production or not. Aggarwal (2013, 2017) refereed to this important step as "Intensional Knowledge". This stage is maybe more investigated in the computer science community than in the statistical one, but some statistician take an interest in it as Debruyne et al. (2017).

In fact, because of the complexity of our algorithm, calculating the contributing values of each variable to the abnormal behavior of the anomalies was quite difficult. So a specific part of our algorithm deals with this issue. However, because of confidentiality issues, the computations can again not be presented.

2. Data is in general position if there is no subset of $k$ observations lying on a subspace of dimension $k-2$, with $k \leq p+1$. See Tyler (2010).

## 5.4 Conclusion

From a practical point of view, a new marketable solution of unsupervised outlier detection was developed for the ippon innovation company, within the RESIST project. This stable algorithm is efficient to detect outlying observations in numerical HDLSS data, more precisely the defects in aerospace electronic components. Moreover, it identifies the tests explaining the abnormal behavior of the dice. So, it appears that this solution meets all the expectations from Microchip-Atmel and is already integrated in their production tool at probe and final test levels. In addition, since the algorithm is implemented in R, we developed a shiny application in order to have a more user-friendly demonstrator for the company.

From a theoretical point of view, dealing with HDLSS data was really challenging. In addition, because of the characteristics of the aerospace measurements, a lot of numerical difficulties occurred. First, the diversity of the units of these measurements was problematic in programming and made unstable the preprocessing step of dimension reduction. Then, the algorithm had to deal with data sets in general position or not and so, adapts itself depending on these different cases.

A more extensive reviewing part than the one presenting in Chapter 1 was also actually valuable for ippon innovation. Indeed, I benchmarked a lot of existing methods on their database and showed that none of them met all the expectations of the company. Moreover, this knowledge was important on other assignments, and so I helped ippon in different practical cases as drifts analysis or cyber security intrusion monitoring. Finally, all the theoretical work from the Chapters 2 and 4 allowed a better understanding of the conditions under which a given method works in practice. As a consequence, each step of the GAT algorithm has its own purpose which can now be statistically and computationally explained.

# Conclusion et Perspectives

## Version française

Ce travail de doctorat, initié par l'entreprise ippon innovation, apporte une contribution au domaine de la détection non-supervisée d'observations atypiques. Plus spécifiquement, il se focalise sur l'identification de composants électroniques défectueux dans les industries automobile et spatiale. Dans ces domaines, de plus en plus de mesures numériques sont effectuées afin de garantir un niveau élevé de fiabilité, au point que le nombre de tests excède souvent la quantité de pièces produites. Cette situation est très intéressante d'un point de vue statistique car encore peu de méthodes sont capables de traiter ce type de données en grande dimension (HDLSS).

L'état de l'art réalisé a permis, dans un premier temps, de constater que les standards dans le contexte automobile n'incluent principalement que des méthodes de détection univariées pour détecter des atypiques. Quant aux acteurs de l'aérospatiale, la majorité ne semble pas encore prête à recourir à des analyses statistiques, bien que certains reconnaissent leur potentiel. Pour preuve, ippon innovation s'est associée à l'entreprise Microchip-Atmel dans le cadre du projet RESIST, qui a pour but d'améliorer la fiabilité de l'électronique dans les secteurs de l'avionique, de l'automobile et de l'aérospatiale.
Dans un second temps, il est apparu que la littérature sur la détection d'observations atypiques est abondante dans les communautés statistique et informatique mais assez dispersée. En se restreignant seulement aux approches non-supervisées, on propose ici une synthèse des principales méthodes en fonction de leurs caractéristiques. Afin de faciliter leur application, on précise également leur mise en œuvre avec le logiciel R, si elle est connue au moment de l'écriture de ce manuscrit. Après avoir testé certaines de ces solutions sur des données industrielles, il s'est avéré qu'elles ne répondaient pas à toutes les attentes requises.

Partant du constat que la distance de Mahalanobis (MD) semble être une des méthodes multivariées les plus utilisées dans le secteur automobile, nous nous sommes intéressés plus en détail à son comportement quand le nombre de variables augmente. Puisqu'il apparaît qu'elle se retrouve en difficulté dans ce cas particulier, nous avons proposé comme alternative une méthodologie d'identification des observations qui repose sur la méthode ICS. Cette analyse, affine invariante, se base sur la diagonalisation simultanée de deux estima-

teurs multivariés de dispersion. Une sélection pertinente des nouvelles composantes permet alors de révéler la structure d'atypicité des données. Nos travaux de recherche se sont principalement focalisés sur cette étape de sélection, autant d'un point de vue empirique que théorique. Toutefois, seul le contexte industriel, caractérisé par un faible pourcentage de contamination a été investigué. Il serait donc pertinent d'étendre notre étude au cas où ce pourcentage augmente. Enfin, nous pourrions également chercher à dériver les résultats théoriques obtenus dans ce manuscrit pour des mélanges à deux composantes pour des mélanges à trois composantes ou plus.
Afin de démocratiser son utilisation, nous avons mis en œuvre la procédure de détection que nous proposons dans deux packages R. Le premier, ICSOutlier, permet d'appliquer la méthode de manière automatique tandis que le deuxième, ICSShiny, est une application shiny qui rend son utilisation plus simple et plus attractive.

Dans la réalité industrielle, les mesures sont souvent colinéaires ou excèdent le nombre d'observations. Dans ce cas, les estimateurs de dispersion ne sont plus définis positifs et la méthode d'ICS classique n'est plus applicable. De plus, il est nécessaire de supposer que les données ne sont pas en position générale pour réussir à définir deux estimateurs de dispersion non proportionnels à la matrice de variance covariance. C'est donc sous cette hypothèse que nous avons adapté la méthode ICS en utilisant trois idées plus ou moins bien connues dans le contexte HDLSS, caractérisé par un nombre plus important de variables que d'observations. En analysant théoriquement les propriétés de ces approches, il est apparu que seule l'utilisation d'une GSVD permettait de garder au mieux les caractéristiques de la méthode classique.

Notons toutefois qu'en pratique la méthode développée n'est qu'invariante par transformation orthogonale puisque les estimateurs de matrice de dispersion considérés ne sont plus qu'orthogonalement equivariants. Pour aller plus loin et résoudre ce problème, il faudrait être en mesure de définir et calculer des estimateurs affines équivariants en HDLSS. Or, du fait de la structure même des données dans ce contexte, cette idée semble difficile à mener à bien. D'autres approches sont envisageables. Ainsi, de nombreux chercheurs proposent directement de régulariser les estimateurs de dispersion afin de les rendre inversibles, au prix d'une hypothèse sur la structure globale des données. De notre côté, avec mes directeurs de thèse et un chercheur chilien nous commençons à réfléchir à une manière de pénaliser le critère ICS, comme le font Witten and Tibshirani (2011) pour l'analyse discriminante. De cette manière, il serait également possible d'adapter la méthode ICS dans le cas où les données ne sont pas en position générale en utilisant un paramètre de pénalisation.

Enfin du point de vue d'ippon innovation, cette collaboration de trois ans a débouché sur le développement d'une solution efficace de détection non-supervisée d'anomalies. L'algorithme confidentiel proposé remplit toutes les attentes de l'entreprise : il s'exécute en quelques secondes, s'adapte au cas où le nombre d'observations est très faible comparé au nombre de tests, s'ajuste si les données sont en position générale et surtout détecte des problèmes de fiabilité en avance. Concrètement, sa mise en œuvre en R a nécessité de

résoudre des problèmes numériques majeurs afin d'assurer la stabilité des résultats lors de permutations de lignes ou de colonnes notamment.
Actuellement, une application shiny a été développée afin d'aider à la stratégie commerciale de l'entreprise qui a déjà vendu plusieurs licences. Le but est de promouvoir l'outil dans le domaine de l'aérospatiale où la statistique a un rôle primordial à jouer. Lors d'une première réunion à l'ESA (*European Space Agency*), plusieurs acteurs de l'industrie spatiale ont montré un intérêt certain pour notre algorithme de détection.

## English version

This PhD work, initiated by the ippon innovation company, contributes to the field of unsupervised outlier detection. More specifically, it focuses on the identification of defective electronic components in the automotive and space industries. In these areas, more and more numerical measurements are carried out in order to guarantee a high level of reliability. As a consequence, the number of tests often exceeds the amount of manufactured parts. This situation is very interesting from a statistical point of view because only a few methods can deal with high dimensional and low sample size (HDLSS) data.

On the one hand, the state of the art established that the standards to detect outliers in the automotive context, mainly include univariate detection methods. As for the aerospace, most of the industries do not seem ready to use statistical analysis, although some recognize their potential. As proof, ippon innovation collaborated with Microchip-Atmel within the RESIST project, which aims to improve the reliability of electronics in the avionics, automotive and aerospace domains.
On the other hand, it appeared that the outlier detection literature is particularly rich in the statistical and computer science communities but rather spread. Restricting only to unsupervised approaches, we propose here a summary of the main methods according to their characteristics. In order to simplify their application, their implementation within the R software is also specified, if it is known at the writing time of this manuscript. After testing some of these solutions on industrial data, it turned out that they did not meet all the required expectations.

Since the Mahalanobis distance (MD) seems to be one of the most widely used multivariate methods in the automotive sector, we focused on its behavior when the number of variables increases. In this particular case, the MD often seems to have troubles to identify the outliers, so we proposed an alternative methodology for outlier detection based on the ICS method. This invariant affine analysis consists of simultaneously diagonalizing two multivariate scatter matrices. A relevant selection of the new components then reveals the outlierness structure of the data. Our research mainly focused on this selection step, both from an empirical and theoretical point of view. However, only a low percentage of contamination, typical in the industrial context, was investigated. It would therefore be appropriate to extend our study to the case where this percentage increases. Finally, we could also try to derive the theoretical results obtained in this manuscript for two component mixtures for mixtures with three or more components.
In order to broaden its use, we implemented this outlier detection procedure that we propose in two packages R. The first one, ICSOutlier, allows to apply the method automatically while the second, ICSShiny is a shiny application that makes it easier and more attractive to use.

In the industrial context, the measurements are often collinear or exceed the number of observations. In this case, the scatter matrices are no longer positive definite and

the classical ICS method is no longer usable. Moreover, it is necessary to assume that the data is not in general position to successfully define two scatter matrices that are not proportional to the variance-covariance matrix. It is therefore under this hypothesis that we adapted the ICS method by using three ideas more or less well-known in the HDLSS context (High-Dimension and Low Sample Size). After theoretically analyzing the properties of these approaches, it turned out that only the use of an GSVD maintains, as well as possible, the characteristics of the classical method.

However, it should be noted that in practice the developed method is only invariant by orthogonal transformation since the considered scatter matrix are only orthogonally equivariant. To go further and solve this problem, one would have to be able to define and compute equivariant affine scatter matrices in HDLSS. However, due to the structure of the data in this context, this idea seems difficult to achieve. Other approaches are possible. Thus, many researchers propose to regularize the scatter matrices to make them invertible, at the cost of an assumption on the overall structure of the data. On our side, with my thesis directors and a Chilean researcher, we begin to think of a way to penalize the ICS criterion, as Witten and Tibshirani (2011) do for discriminant analysis. Considering a penalty parameter would also allow for adapting the ICS method in the case where the data are not in general position.

Finally, from the point of view of ippon innovation, this three-year collaboration resulted in the development of an effective solution for the unsupervised outlier detection. The proposed confidential algorithm fulfills all the expectations of the company: it runs within seconds, it adapts itself to the case where the number of observations is very low compared to the number of tests, it adjusts if the data is in general position and it detects reliability issues in advance. Major numerical problems were solved during the implementation step in R to ensure the stability of the results, in particular when rows or columns are permuted.
Currently, a shiny application has been developed to help the business strategy of the company that has already sold several licenses. The aim is to promote the tool in the aerospace field to favor the use of statistics in this critical area. At a first meeting at the ESA (*European Space Agency*), several players in the space industry have shown some interest in our outlier detection algorithm.

# Bibliography


Aggarwal, C. C. (2013). *Outlier Analysis*. Springer Publishing Company, Incorporated.

Aggarwal, C. C. (2017). *Outlier Analysis, 2nd edition*. Springer Publishing Company, Incorporated.

Aggarwal, C. C., Hinneburg, A., and Keim, D. A. (2001). On the surprising behavior of distance metrics in high dimensional space. In *International Conference on Database Theory*, pages 420–434. Springer.

Aggarwal, C. C. and Sathe, S. (2017). *Outlier Ensembles: An Introduction*. Springer.

Aggarwal, C. C. and Yu, P. S. (2001). Outlier detection for high dimensional data. In *ACM Sigmod Record*, pages 37–46. ACM.

Agostinelli, C., Leung, A., Yohai, V. J., and Zamar, R. H. (2015). Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination. *Test*, 24(3):441–461.

Agyemang, M., Barker, K., and Alhajj, R. (2006). A comprehensive survey of numeric and symbolic outlier mining techniques. *Intelligent Data Analysis*, 10(6):521–538.

Alashwali, F. and Kent, J. (2016). The use of a common location measure in the invariant coordinate selection and projection pursuit. *Journal of Multivariate Analysis*, 152:145–161.

An, J. and Cho, S. (2015). Variational autoencoder based anomaly detection using reconstruction probability. *SNU Data Mining Center - Special Lecture on IE*.

Archimbaud, A., May, J., Nordhausen, K., and Ruiz-Gazen, A. (2017). *ICSShiny: Invariant Coordinate Selection With a Shiny App*. R package version 0.4.

Archimbaud, A., Nordhausen, K., and Ruiz-Gazen, A. (2016). ICS for multivariate outlier detection with application to quality control. *submitted*.

Archimbaud, A., Nordhausen, K., and Ruiz-Gazen, A. (2016). *ICSOutlier: Outlier Detection Using Invariant Coordinate Selection*. R package version 0.3-0.

Archimbaud, A., Nordhausen, K., and Ruiz-Gazen, A. (2017a). Unsupervized outlier detection with ICSOutlier. *accepted under revision*.

Archimbaud, A., Soual, C., Bergeret, F., D'Alberto, S., Thebault, T., and Bonin, C. (2017b). High dimensional outlier screening of small dice samples for aerospace IC reliability. *The 10th international conference on mathematical methods in reliability, Grenoble, France*.

Automotive Electronic Council (2011). Guidelines for part average testing. *AEC-Q001, rev-D*.

Bai, Z. (1992). The CSD, GSVD, their applications and computations. *IMA Preprint Series*, (958).

Bai, Z. and Demmel, J. W. (1993). Computing the generalized singular value decomposition. *SIAM Journal on Scientific Computing*, 14(6):1464–1486.

Bai, Z. and Zha, H. (1993). A new preprocessing algorithm for the computation of the generalized singular value decomposition. *SIAM Journal on Scientific Computing*, 14(4):1007–1012.

Barnett, V. and Lewis, T. (1994). *Outliers in Statistical Data*. Wiley.

Bay, S. D. and Schwabacher, M. (2003). Mining distance-based outliers in near linear time with randomization and a simple pruning rule. In *Proceedings of the ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 29–38. ACM.

Becker, C. and Gather, U. (1999). The masking breakdown point of multivariate outlier identification rules. *Journal of the American Statistical Association*, 94(447):947–955.

Beckman, R. J. and Cook, R. D. (1983). Outliers. *Technometrics*, 25(2):119–149.

Bernard, A. and Saporta, G. (2013). Analyse en composantes principales sparse pour données multiblocs et extensiona l'analyse des correspondances multiples sparse. In *45emes Journées de Statistique*.

Beyer, K., Goldstein, J., Ramakrishnan, R., and Shaft, U. (1999). When is "nearest neighbor" meaningful? In *International Conference on Database Theory*, pages 217–235. Springer.

Billor, N., Hadi, A. S., and Velleman, P. F. (2000). BACON: blocked adaptive computationally efficient outlier nominators. *Computational Statistics & Data Analysis*, 34(3):279–298.

Bonett, D. G. and Seier, E. (2002). A test of normality with high uniform power. *Computational Statistics & Data Analysis*, 40(3):435–445.

Bookstein, F. L. and Mitteroecker, P. (2014). Comparing covariance matrices by relative eigenanalysis, with applications to organismal biology. *Evolutionary Biology*, 41(2):336–350.

Borchers, H. W. (2017). *pracma: Practical Numerical Math Functions*. R package version 2.0.7.

Branco, J. A. and Pires, A. M. (2015). High dimensionality: the trouble with Mahalanobis distance. WOMAT: Workshop On Multivariate Analysis Today.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J. (1999). Optics-of: Identifying local outliers. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 262–270. Springer.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J. (2000). LOF: identifying density-based local outliers. In *ACM Sigmod Record*, pages 93–104. ACM.

Campbell, N. A. (1980). Robust procedures in multivariate analysis I: Robust covariance estimation. *Applied Statistics*, 29(3):231–237.

Campos, G. O., Zimek, A., Sander, J., Campello, R. J., Micenková, B., Schubert, E., Assent, I., and Houle, M. E. (2015). On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study. *Data Mining and Knowledge Discovery*, pages 1–37.

Candès, E. J., Li, X., Ma, Y., and Wright, J. (2011). Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11.

Cardot, H. and Godichon, A. (2015). Robust principal components analysis based on the median covariation matrix. *arXiv preprint arXiv:1504.02852*.

Cateni, S., Vannucci, M., and Colla, V. (2008). *Outlier detection methods for industrial applications*. INTECH Open Access Publisher.

Cator, E. A. and Lopuhaä, H. P. (2012). Central limit theorem and influence function for the MCD estimators at general multivariate distributions. *Bernoulli*, 18(2):520–551.

Caussinus, H., Fekri, M., Hakam, S., and Ruiz-Gazen, A. (2003a). A monitoring display of multivariate outliers. *Computational Statistics & Data Analysis*, 44(1):237–252.

Caussinus, H., Hakam, S., and Ruiz-Gazen, A. (2003b). Projections révélatrices contrôlées: Groupements et structures diverses. *Revue de Statistique Appliquée*, 51(1):37–58.

Caussinus, H. and Ruiz-Gazen, A. (1990). Interesting projections of multidimensional data by means of generalized principal component analyses. In *Proceedings of COMPSTAT'1990*, pages 121–126. Springer.

Caussinus, H. and Ruiz-Gazen, A. (1994). Projection pursuit and generalized principal component analysis. *New Directions in Statistical Data Analysis and Robustness*, pages 35–46.

Caussinus, H. and Ruiz-Gazen, A. (1995). Metrics for finding typical structures by means of principal component analysis. *Data Science and its Applications*, pages 177–192.

Cerioli, A. (2010). Multivariate outlier detection with high-breakdown estimators. *Journal of the American Statistical Association*, 105(489):147–156.

Cerioli, A. and Farcomeni, A. (2011). Error rates for multivariate outlier detection. *Computational Statistics & Data Analysis*, 55(1):544–553.

Cerioli, A., Riani, M., and Atkinson, A. C. (2009). Controlling the size of multivariate outlier tests with the MCD estimator of scatter. *Statistics and Computing*, 19(3):341–353.

Chalmers, R. P. and Flora, D. B. (2015). faoutlier: An R package for detecting influential cases in exploratory and confirmatory factor analysis. *Applied Psychological Measurement*, 39(7):573–574.

Chandola, V., Banerjee, A., and Kumar, V. (2007). Outlier detection: A survey. Technical report, University of Minnesota.

Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys (CSUR)*, 41(3):15.

Chen, K. (2005). *Matrix preconditioning techniques and applications*. Cambridge University Press.

Cinar, A. and Undey, C. (1999). Statistical process and controller performance monitoring. a tutorial on current methods and future directions. In *Proceedings of the American Control Conference*, volume 4, pages 2625–2639. IEEE.

Croux, C., Filzmoser, P., and Fritz, H. (2013). Robust sparse principal component analysis. *Technometrics*, 55(2):202–214.

Croux, C., Filzmoser, P., and Oliveira, M. R. (2007). Algorithms for projection-pursuit robust principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 87(2):218–225.

Croux, C. and Haesbroeck, G. (1999). Influence function and efficiency of the minimum covariance determinant scatter matrix estimator. *Journal of Multivariate Analysis*, 71(2):161–190.

Croux, C. and Ruiz-Gazen, A. (2005). High breakdown estimators for principal components: the projection-pursuit approach revisited. *Journal of Multivariate Analysis*, 95(1):206–226.

Dang, X. H., Assent, I., Ng, R. T., Zimek, A., and Schubert, E. (2014). Discriminative features for identifying and interpreting outliers. In *IEEE 30th International Conference on Data Engineering (ICDE)*, pages 88–99. IEEE.

Darlington, R. B. (1970). Is kurtosis really "peakedness"? *The American Statistician*, 24(2):19–22.

Debruyne, M., Höppner, S., Serneels, S., and Verdonck, T. (2017). Outlyingness: why do outliers lie out? *arXiv preprint arXiv:1708.03761v1*.

Debruyne, M. and Hubert, M. (2009). The influence function of the Stahel-Donoho covariance estimator of smallest outlyingness. *Statistics & Probability Letters*, 79(3):275–282.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 2:1–38.

Devlin, S. J., Gnanadesikan, R., and Kettenring, J. R. (1981). Robust estimation of dispersion matrices and principal components. *Journal of the American Statistical Association*, 76(374):354–362.

DO Q, L. (2012). Numerically efficient methods for solving least squares problems.

Donoho, D. L. (1982). Breakdown properties of multivariate location estimators. Technical report, Ph.D. Qualifying paper, Dept. Statistics, Harvard University, Boston.

Donoho, D. L. and Gasko, M. (1992). Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *The Annals of Statistics*, 20(4):1803–1827.

Dray, S. (2008). On the number of principal components: A test of dimensionality based on measurements of similarity between matrices. *Computational Statistics & Data Analysis*, 52(4):2228–2237.

Droesbeke, J.-J., Saporta, G., and Thomas-Agnan, C. (2015). *Méthodes robustes en statistique*.

Duda, R. O., Hart, P. E., and Stork, D. G. (2012). *Pattern classification*. John Wiley & Sons.

Dutta, H., Giannella, C., Borne, K., and Kargupta, H. (2007). Distributed top-k outlier detection from astronomy catalogs using the demac system. In *Proceedings of the 2007 SIAM International Conference on Data Mining*, pages 473–478. SIAM.

Engelen, S., Hubert, M., and Branden, K. V. (2005). A comparison of three procedures for robust PCA in high dimensions. *Austrian Journal of Statistics*, 34(2):117–126.

Eo, S.-H. and Cho, H. (2014). *OutlierDC: Outlier Detection using quantile regression for Censored Data*. R package version 0.3-0.

Fan, C. (2015). *HighDimOut: Outlier Detection Algorithms for High-Dimensional Data*. R package version 1.0.0.

Farcomeni, A. and Greco, L. (2016). *Robust methods for data reduction*. CRC press.

Filzmoser, P., Garrett, R. G., and Reimann, C. (2005). Multivariate outlier detection in exploration geochemistry. *Computers & Geosciences*, 31(5):579–587.

Filzmoser, P. and Gschwandtner, M. (2015). *mvoutlier: Multivariate outlier detection based on robust methods*. R package version 2.0.6.

Filzmoser, P., Maronna, R., and Werner, M. (2008). Outlier identification in high dimensions. *Computational Statistics & Data Analysis*, 52(3):1694–1711.

Filzmoser, P. and Todorov, V. (2011). Review of robust multivariate statistical methods in high dimension. *Analytica Chimica Acta*, 705(1):2–14.

Filzmoser, P. and Todorov, V. (2013). Robust tools for the imperfect world. *Information Sciences*, 245:4–20.

Fischer, D., Berro, A., Nordhausen, K., and Ruiz-Gazen, A. (2016a). REPPlab: An R package for detecting clusters and outliers using exploratory projection pursuit. Technical report, arXiv:1612.06518v1.

Fischer, D., Berro, A., Nordhausen, K., and Ruiz-Gazen, A. (2016b). *REPPlab: R Interface to 'EPP-Lab', a Java Program for Exploratory Projection Pursuit*. R package version 0.9.4.

Fischer, D., Honkatukia, M., Tuiskula-Haavisto, M., Nordhausen, K., Cavero, D., Preisinger, R., and Vilkki, J. (2017). Subgroup detection in genotype data using invariant coordinate selection. *BMC Bioinformatics*, 18(1):173.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Human Genetics*, 7(2):179–188.

Friedman, J. H. and Tukey, J. W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers*, 100(9):881–890.

Fujimaki, R., Yairi, T., and Machida, K. (2005). An approach to spacecraft anomaly detection problem using kernel feature space. In *Proceedings of the eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, pages 401–410. ACM.

Gao, J. and Tan, P.-N. (2006). Converting output scores from outlier detection algorithms into probability estimates. In *ICDM'06 - Sixth International Conference on Data Mining*, pages 212–221. IEEE.

Genest, M., Masse, J.-C., and Plante, J.-F. (2012). *depth: Depth functions tools for multivariate analysis*. R package version 2.0-0.

Genz, A. and Bretz, F. (2009). *Computation of Multivariate Normal and t Probabilities*. Lecture Notes in Statistics. Springer-Verlag, Heidelberg.

Geun Kim, M. (2000). Multivariate outliers and decompositions of Mahalanobis distance. *Communications in Statistics-Theory and Methods*, 29(7):1511–1526.

Ghoting, A., Parthasarathy, S., and Otey, M. E. (2006). Fast mining of distance-based outliers in high-dimensional datasets. In *Proceedings of the 2006 SIAM International Conference on Data Mining*, pages 609–613. SIAM.

Gnanadesikan, R. and Kettenring, J. R. (1972). Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, 28(1):81–124.

Golub, G. H. and Van der Vorst, H. A. (2000). Eigenvalue computation in the 20th century. *Journal of Computational and Applied Mathematics*, 123(1):35–65.

Golub, G. H. and Van Loan, C. F. (1996). Matrix computations. 1996. *Johns Hopkins University, Press, Baltimore, MD, USA*, pages 374–426.

Green, C. G. and Martin, D. (2017a). *CerioliOutlierDetection: Outlier Detection Using the Iterated RMCD Method of Cerioli (2010)*. R package version 1.1.9.

Green, C. G. and Martin, R. D. (2017b). An extension of a method of Hardin and Rocke, with an application to multivariate outlier detection via the IRMCD method of Cerioli. Technical report, Working Paper, 2017. Available from http://christopherggreen.github.io/papers/hr05_extension.pdf.

Greene, W. H. (2012). Econometric analysis, 71e. *Stern School of Business, New York University*.

Grubbs, F. E. (1950). Sample criteria for testing outlying observations. *The Annals of Mathematical Statistics*, 21(1):27–58.

Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. *Technometrics*, 11(1):1–21.

Guillouet, B. (2012). Rapport de stage: Détection de pièces défaillantes pour haute fiabilité. Technical report, INSA Toulouse.

Hadi, A. S., Imon, A., and Werner, M. (2009). Detection of outliers. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1(1):57–70.

Hampel, F., Ronchetti, E., Rousseeuw, P., and Stahel, W. (1986). Robust statistics: the approach based on influence functions. *Wiley & Sons, New York*.

Hardin, J. and Rocke, D. M. (2005). The distribution of robust distances. *Journal of Computational and Graphical Statistics*, 14(4):928–946.

Harkat, M.-F., Mourot, G., and Ragot, J. (2002). Différentes méthodes de localisation de défauts basées sur les dernières composantes principales. In *Conférence Internationale Francophone d'Automatique (CIFA)*.

Hassan, A. H. (2014). *Détection multidimensionnelle au test paramétrique avec recherche automatique des causes*. PhD thesis, Université Grenoble.

Hasselman, B. and Lapack authors (2017). *geigen: Calculate Generalized Eigenvalues, the Generalized Schur Decomposition and the Generalized Singular Value Decomposition of a Matrix Pair with Lapack*. R package version 2.0.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA.

Hawkins, D. M. (1980). *Identification of outliers*, volume 11. Springer.

Hawkins, D. M., Bradu, D., and Kass, G. V. (1984). Location of several outliers in multiple-regression data using elemental sets. *Technometrics*, 26(3):197–208.

Higham, N. J. (2000). QR factorization with complete pivoting and accurate computation of the svd. *Linear Algebra and its Applications*, 309(1):153–174.

Higham, N. J. (2002). *Accuracy and stability of numerical algorithms*. Siam.

Hodge, V. J. and Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22(2):85–126.

Hogben, L. (2006). *Handbook of linear algebra*. CRC Press.

Hotelling, H. (1931). The generalization of Student's ratio. *The Annals of Mathematical Statistics*, 2(3):360–378.

Hotteling, H. (1947). Multivariate quality control illustrated by the air testing of sample bombsites. *Selected Techniques of Statistical Analysis*, page 111.

Howe, D. C. (2015). *kmodR: K-Means with Simultaneous Outlier Detection*. R package version 0.1.0.

Howland, P., Jeon, M., and Park, H. (2003). Structure preserving dimension reduction for clustered text data based on the generalized singular value decomposition. *SIAM Journal on Matrix Analysis and Applications*, 25(1):165–179.

Howland, P. and Park, H. (2004). Generalizing discriminant analysis using the generalized singular value decomposition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(8):995–1006.

Howland, P., Wang, J., and Park, H. (2006). Solving the small sample size problem in face recognition using generalized discriminant analysis. *Pattern Recognition*, 39(2):277–287.

Hu, Y., Murray, W., Shan, Y., and Australia (2015). *Rlof: R Parallel Implementation of Local Outlier Factor (LOF)*. R package version 1.1.1.

Huber, P. J. (1985). Projection pursuit. *The Annals of Statistics*, 13(2):435–475.

Hubert, M., Reynkens, T., Schmitt, E., and Verdonck, T. (2016). Sparse PCA for high-dimensional data with outliers. *Technometrics*, 58(4):424–434.

Hubert, M., Rousseeuw, P. J., and Vanden Branden, K. (2005). ROBPCA: a new approach to robust principal component analysis. *Technometrics*, 47(1):64–79.

Hubert, M., Rousseeuw, P. J., and Verboven, S. (2002). A fast method for robust principal components with applications to chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 60(1):101–111.

Hyvärinen, A., Karhunen, J., and Oja, E. (2004). *Independent component analysis*, volume 46. John Wiley & Sons.

Jain, A. K. and Dubes, R. C. (1988). *Algorithms for clustering data*. Prentice-Hall, Inc.

JEDEC (2009). Outlier identification and management system for electronic components. Technical report, JEDEC, Replaced by JESD50.

Jensen, W. A., Birch, J. B., and Woodall, W. H. (2007). High breakdown estimation methods for phase I multivariate control charts. *Quality and Reliability Engineering International*, 23(5):615–629.

Jimenez, J. (2015). *abodOutlier: Angle-Based Outlier Detection*. R package version 0.1.

Jobe, J. M. and Pokojovy, M. (2015). A cluster-based outlier detection scheme for multivariate data. *Journal of the American Statistical Association*, 110(512):1543–1551.

Johnson, R. A. and Wichern, D. W. (1998). *Applied Multivariate Statistical Analysis (6th Edition)*. Prentice Hall.

Jolliffe, I. (2002). *Principal component analysis*. Wiley Online Library.

Josse, J. and Sardy, S. (2016). Adaptive shrinkage of singular values. *Statistics and Computing*, 26(3):715–724.

Josse, J., Sardy, S., and Wager, S. (2016a). denoiseR: A package for low rank matrix estimation. *arXiv preprint arXiv:1602.01206*.

Josse, J., Sardy, S., and Wager, S. (2016b). *denoiseR: Regularized Low Rank Matrix Estimation*. R package version 1.0.

Keller, F., Muller, E., and Bohm, K. (2012). HiCS: High contrast subspaces for density-based outlier ranking. In *IEEE 28th International Conference on Data Engineering (ICDE)*, pages 1037–1048. IEEE.

Kim, H., Howland, P., and Park, H. (2005). Dimension reduction in text classification with support vector machines. *Journal of Machine Learning Research*, 6(Jan):37–53.

Knorr, E. M. and Ng, R. T. (1998). Algorithms for mining distancebased outliers in large datasets. In *Proceedings of the International Conference on Very Large Data Bases*, pages 392–403. Citeseer.

Knorr, E. M. and Ng, R. T. (1999). Finding intensional knowledge of distance-based outliers. In *VLDB*, volume 99, pages 211–222.

Komsta, L. (2011). *outliers: Tests for outliers*. R package version 0.14.

Komsta, L. and Novomestky, F. (2015). *moments: Moments, cumulants, skewness, kurtosis and related tests*. R package version 0.14.

Kriegel, H.-P., Kröger, P., Schubert, E., and Zimek, A. (2009). Outlier detection in axis-parallel subspaces of high dimensional data. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 831–838. Springer.

Kriegel, H.-P., Kroger, P., Schubert, E., and Zimek, A. (2011). Interpreting and unifying outlier scores. In *Proceedings of the 2011 SIAM International Conference on Data Mining*, pages 13–24. SIAM.

Kriegel, H.-P., Kroger, P., Schubert, E., and Zimek, A. (2012). Outlier detection in arbitrarily oriented subspaces. In *IEEE 12th International Conference on Data Mining (ICDM)*, pages 379–388. IEEE.

Kriegel, H.-P., Kröger, P., and Zimek, A. (2010). Outlier detection techniques. *Tutorial at KDD*, 10.

Kriegel, H.-P., Zimek, A., et al. (2008). Angle-based outlier detection in high-dimensional data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 444–452. ACM.

Kwitt, R. and Hofmann, U. (2006). Robust methods for unsupervised PCA-based anomaly detection. *Proc. of IEEE/IST WorNshop on Monitoring, AttacN Detection and Mitigation*, pages 1–3.

Lafaye de Micheaux, D. (2000). Prolonger la MSP par la "maîtrise globale du processus". *Qualité références*.

Lafaye de Micheaux, D., Cembrynski, T., Dalancon, T., and Demonsant, J. (2007). Réduction de la dispersion des caracteristiques produit, méthodologie GPC et application en carrosserie automobile. In *7ième édition du Congrès International Pluridisciplinaire Qualita 2007, Tanger (Maroc)*.

Lafaye de Micheaux, D. and Vieux, D. (Janvier 2005). MSP multidimensionnelle, détecter et identifier "l'invisible". *Qualité références*, pages 79–82.

Laurikkala, J., Juhola, M., Kentala, E., Lavrac, N., Miksch, S., and Kavsek, B. (2000). Informal identification of outliers in medical data. In *Fifth International Workshop on Intelligent Data Analysis in Medicine and Pharmacology*, volume 1, pages 20–24.

Lazarevic, A. and Kumar, V. (2005). Feature bagging for outlier detection. In *Proceedings of the eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, pages 157–166. ACM.

Lê, S., Josse, J., and Husson, F. (2008). FactoMineR: A package for multivariate analysis. *Journal of Statistical Software*, 25(1):1–18.

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411.

Ledoit, O. and Wolf, M. (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *The Annals of Statistics*, 40(2).

Lee, J.-M., Yoo, C., Choi, S. W., Vanrolleghem, P. A., and Lee, I.-B. (2004a). Nonlinear process monitoring using kernel principal component analysis. *Chemical Engineering Science*, 59(1):223–234.

Lee, J.-M., Yoo, C., and Lee, I.-B. (2004b). Statistical process monitoring with independent component analysis. *Journal of Process Control*, 14(5):467–485.

Liquet, B. and Saracco, J. (2012). A graphical tool for selecting the number of slices and the dimension of the model in SIR and SAVE approaches. *Computational Statistics*, 27(1):103–125.

Liski, E., Nordhausen, K., and Oja, H. (2014). Supervised invariant coordinate selection. *Statistics*, 48(4):711–731.

Livne, O. E. and Golub, G. H. (2004). Scaling by binormalization. *Numerical Algorithms*, 35(1):97–120.

Locantore, N., Marron, J., Simpson, D., Tripoli, N., Zhang, J., Cohen, K., Boente, G., Fraiman, R., Brumback, B., Croux, C., et al. (1999). Robust principal component analysis for functional data. *Test*, 8(1):1–73.

Luu, K., Blum, M. G., and Duforet-Frebourg, N. (2016). *pcadapt: Fast Principal Component Analysis for Outlier Detection*. R package version 3.0.2.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate analysis*. Academic press.

Markou, M. and Singh, S. (2003). Novelty detection: a review part 1: statistical approaches. *Signal Processing*, 83(12):2481–2497.

Maronna, R., Martin, R. D., and Yohai, V. (2006). *Robust statistics*. John Wiley & Sons, Chichester. ISBN.

Maronna, R. A. (1976). Robust M-estimators of multivariate location and scatter. *The Annals of Statistics*, 4(1):51–67.

Maronna, R. A., Stahel, W. A., and Yohai, V. J. (1992). Bias-robust estimators of multivariate scatter based on projections. *Journal of Multivariate Analysis*, 42(1):141–161.

Maronna, R. A. and Yohai, V. J. (1995). The behavior of the Stahel-Donoho robust multivariate estimator. *Journal of the American Statistical Association*, 90(429):330–341.

Maronna, R. A. and Zamar, R. H. (2002). Robust estimates of location and dispersion for high-dimensional datasets. *Technometrics*, 44(4).

May, J. (2017). Internship report: Development of an R Shiny application to implement the invariant coordinate selection method. Technical report, Toulouse School of Economics, TSE-R.

Mercier, S. and Bergeret, F. (2011). *Maîtrise Statistique des procédés - Principes et cas industriels*. Dunod/Usine Nouvelle.

Mnassri, B., Ananou, B., Ouladsine, M., Gasnier, F., et al. (2008). Détection et localisation de défauts des wafers par des approches statistiques multivariees et calcul des contributions. In *CIFA 2008, Conférence Internationale Francophone d'Automatique*.

Moler, C. B. and Stewart, G. W. (1973). An algorithm for generalized matrix eigenvalue problems. *SIAM Journal on Numerical Analysis*, 10(2):241–256.

Moreno-Lizaranzu, M. J. and Cuesta, F. (2013). Improving electronic sensor reliability by robust outlier screening. *Sensors*, 13(10):13521–13542.

Muller, E., Assent, I., Iglesias, P., Mulle, Y., and Bohm, K. (2012). Outlier ranking via subspace analysis in multiple views of the data. In *IEEE 12th International Conference on Data Mining (ICDM)*, pages 529–538. IEEE.

Muller, E., Assent, I., Steinhausen, U., and Seidl, T. (2008). OutRank: ranking outliers in high dimensional data. In *ICDEW 2008, IEEE 24th International Conference on Data Engineering Workshop*, pages 600–603. IEEE.

Müller, E., Schiffer, M., Gerwert, P., Hannen, M., Jansen, T., and Seidl, T. (2010a). SOREX: Subspace outlier ranking exploration toolkit. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 607–610. Springer.

Müller, E., Schiffer, M., and Seidl, T. (2010b). Adaptive outlierness for subspace outlier ranking. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pages 1629–1632. ACM.

Müller, E., Schiffer, M., and Seidl, T. (2011). Statistical selection of relevant subspace projections for outlier ranking. In *IEEE 27th International Conference on Data Engineering (ICDE)*, pages 434–445. IEEE.

Nguyen, H. V., Ang, H. H., and Gopalkrishnan, V. (2010). Mining outliers with ensemble of heterogeneous detectors on random subspaces. In *International Conference on Database Systems for Advanced Applications*, pages 368–383. Springer.

Nordhausen, K., Oja, H., and Tyler, D. E. (2008). Tools for exploring multivariate data: The package ICS. *Journal of Statistical Software*, 28(6):1–31.

Nordhausen, K., Oja, H., and Tyler, D. E. (2016). Asymptotic and bootstrap tests for subspace dimension. Technical report, arXiv:1611.04908v1.

Nordhausen, K., Oja, H., Tyler, D. E., and Virta, J. (2017). Asymptotic and Bootstrap Tests for the Dimension of the Non-Gaussian Subspace. *Signal Processing Letters*, 24:887–891.

Nordhausen, K. and Tyler, D. E. (2015). A cautionary note on robust covariance plug-in methods. *Biometrika*, 102(3):573–588.

Oja, H., Sirkiä, S., and Eriksson, J. (2006). Scatter matrices and independent component analysis. *Austrian Journal of Statistics*, 35(2&3):175–189.

Ollila, E. and Tyler, D. E. (2014). Regularized-estimators of scatter matrix. *IEEE Transactions on Signal Processing*, 62(22):6059–6070.

Paige, C. (1986). Computing the generalized singular value decomposition. *SIAM Journal on Scientific and Statistical Computing*, 7(4):1126–1146.

Paige, C. C. and Saunders, M. A. (1981). Towards a generalized singular value decomposition. *SIAM Journal on Numerical Analysis*, 18(3):398–405.

Parlett, B. N. (1998). *The symmetric eigenvalue problem*. SIAM.

Parra, L., Deco, G., and Miesbach, S. (1996). Statistical independence and novelty detection with information preserving nonlinear maps. *Neural Computation*, 8(2):260–269.

Pearson, E. S. and Sekar, C. C. (1936). The efficiency of statistical tools and a criterion for the rejection of outlying observations. *Biometrika*, 28(3/4):308–320.

Peirce, B. (1852). Criterion for the rejection of doubtful observations. *The Astronomical Journal*, 2:161–163.

Peña, D. and Prieto, F. J. (2001a). Cluster identification using projections. *Journal of the American Statistical Association*, 96(456).

Peña, D. and Prieto, F. J. (2001b). Multivariate outlier detection and robust covariance matrix estimation. *Technometrics*, 43(3).

Penny, K. I. and Jolliffe, I. T. (1999). Multivariate outlier detection applied to multiply imputed laboratory data. *Statistics in Medicine*, 18(14):1879–1895.

Peres-Neto, P. R., Jackson, D. A., and Somers, K. M. (2005). How many principal components? stopping rules for determining the number of non-trivial axes revisited. *Computational Statistics & Data Analysis*, 49(4):974–997.

Pham, N. and Pagh, R. (2012). A near-linear time approximation algorithm for angle-based outlier detection in high-dimensional data. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 877–885. ACM.

Pimentel, M. A., Clifton, D. A., Clifton, L., and Tarassenko, L. (2014). A review of novelty detection. *Signal Processing*, 99:215–249.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1996). *Numerical recipes in C*, volume 2. Cambridge university press Cambridge.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Radovanović, M., Nanopoulos, A., and Ivanović, M. (2010). On the existence of obstinate results in vector space models. In *Proceedings of the 33rd international ACM SIGIR conference on Research and Development in Information Retrieval*, pages 186–193. ACM.

Ramaswamy, S., Rastogi, R., and Shim, K. (2000). Efficient algorithms for mining outliers from large data sets. In *ACM Sigmod Record*, pages 427–438. ACM.

Rehage, A. and Kuhnt, S. (2016). *alphaOutlier: Obtain Alpha-Outlier Regions for Well-Known Probability Distributions*. R package version 1.2.0.

Reynkens, T., Hubert, M., Schmitt, E., and Verdonck, T. (2015). Sparse PCA for high-dimensional data with outliers. *Technometrics*.

Rider, P. R. (1933). *Criteria for rejection of observations*. Washington University Studies, St. Louis.

Ro, K., Zou, C., Wang, Z., Yin, G., et al. (2015). Outlier detection for high-dimensional data. *Biometrika*, 102(3):589–599.

Rocke, D. M. (1989). Robust control charts. *Technometrics*, 31(2):173–184.

Rocke, D. M. (1992). X_Q and R_Q charts: Robust control charts. *The Statistician*, 41(1):97–104.

Rocke, D. M. and Woodruff, D. L. (1996). Identification of outliers in multivariate data. *Journal of the American Statistical Association*, 91(435):1047–1061.

Rohlf, F. J. (1975). Generalization of the gap test for the detection of multivariate outliers. *Biometrics*, 31(1):93–101.

Rousseeuw, P., Croux, C., Todorov, V., Ruckstuhl, A., Salibian-Barrera, M., Verbeke, T., Koller, M., and Maechler, M. (2016). *robustbase: Basic Robust Statistics*. R package version 0.92-6.

Rousseeuw, P. and Hubert, M. (2013). High-breakdown estimators of multivariate location and scatter. In *Robustness and Complex Data Structures*, pages 49–66. Springer Berlin Heidelberg.

Rousseeuw, P. J. (1986). Multivariate estimation with high breakdown point. In Grossman, W., Pflug, G., Vincze, I., and Wertz, W., editors, *Mathematical Statistics and Applications*, pages 283–297. Reidel, Dordrecht.

Rousseeuw, P. J. and Kaufman, L. (1990). *Finding Groups in Data*. Wiley Online Library.

Rousseeuw, P. J. and Leroy, A. M. (2005). *Robust regression and outlier detection*, volume 589. John wiley & sons.

Rousseeuw, P. J. and Van den Bossche, W. (2017). Detecting deviating data cells. *Technometrics*, pages 1–11.

Rousseeuw, P. J. and Van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85(411):633–639.

Ruiz-Gazen, A., Marie-Sainte, S. L., and Berro, A. (2010). Detecting multivariate outliers using projection pursuit with particle swarm optimization. In *Proceedings of COMPSTAT'2010*, pages 89–98. Springer.

Ruts, I. and Rousseeuw, P. J. (1996). Computing depth contours of bivariate point clouds. *Computational Statistics & Data Analysis*, 23(1):153–168.

Safo, S. E., Ahn, J., Jeon, Y., and Jung, S. (2016). Sparse generalized eigenvalue problem with application to canonical correlation analysis for integrative analysis of methylation and gene expression data. *arXiv preprint arXiv:1611.01066v1*.

Saracco, J., Larramendy, I., and Aragon, Y. (1999). La regression inverse par tranches ou méthode SIR: presentation générale. *La revue de Modulad*, (22):21–39.

Schott, J. R. (2005). *Matrix analysis for statistics*. Wiley.

Schubert, E., Wojdanowski, R., Zimek, A., and Kriegel, H.-P. (2012). On evaluation of outlier rankings and outlier scores. In *Proceedings of the 2012 SIAM International Conference on Data Mining*, pages 1047–1058. SIAM.

Serfling, R. and Mazumder, S. (2013). Computationally easy outlier detection via projection pursuit with finitely many directions. *Journal of Nonparametric Statistics*, 25(2):447–461.

Serfling, R. J. (1980). *Approximation theorems of mathematical statistics*, volume 162. John Wiley & Sons.

She, Y., Li, S., and Wu, D. (2016). Robust orthogonal complement principal component analysis. *Journal of the American Statistical Association*, 111(514):763–771.

Shen, H. and Huang, J. Z. (2008). Sparse principal component analysis via regularized low rank matrix approximation. *Journal of Multivariate Analysis*, 99(6):1015–1034.

Shyu, M.-L., Chen, S.-C., Sarinnapakorn, K., and Chang, L. (2003). A novel anomaly detection scheme based on principal component classifier. Technical report, DTIC Document.

Singh, K. and Upadhyaya, S. (2012). Outlier detection: applications and techniques. *International Journal of Computer Science Issues*, 9(1):307–323.

Smith, R., Bivens, A., Embrechts, M., Palagiri, C., and Szymanski, B. (2002). Clustering approaches for anomaly based intrusion detection. *Proceedings of Intelligent Engineering Systems Through Artificial Neural Networks*, pages 579–584.

Stahel, W., Maechler, M., and potentially others (2013). *robustX: eXperimental Functionality for Robust Statistics*. R package version 1.1-4.

Stahel, W. A. (1981). *Breakdown of covariance estimators*. Fachgruppe für Statistik, Eidgenössische Techn. Hochsch.

Stahel, W. A. and Mächler, M. (2009). Comment on "invariant co-ordinate selection". *Journal of the Royal Statistical Society B*, 71(584–586).

Stewart, G. W. (1998). Perturbation theory for the singular value decomposition. Technical report.

Sullivan, J. H. and Woodall, W. H. (1996). A comparison of multivariate control charts for individual observations. *Journal of Quality Technology*, 28(4):398–408.

Tang, J., Chen, Z., Fu, A. W.-C., and Cheung, D. W. (2002). Enhancing effectiveness of outlier detections for low density patterns. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 535–548. Springer.

Tao, Y., Xiao, X., and Zhou, S. (2006). Mining distance-based outliers from large databases in any metric space. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 394–403. ACM.

Taouali, O., Jaffel, I., Lahdhiri, H., Harkat, M. F., and Messaoud, H. (2016). New fault detection method based on reduced kernel principal component analysis (RKPCA). *The International Journal of Advanced Manufacturing Technology*, 85(5-8):1547–1552.

Tarr, G., Müller, S., and Weber, N. C. (2016). Robust estimation of precision matrices under cellwise contamination. *Computational Statistics & Data Analysis*, 93:404–420.

Tatum, L. G. (1997). Robust estimation of the process standard deviation for control charts. *Technometrics*, 39(2):127–141.

Tebbens, J. D. and Schlesinger, P. (2007). Improving implementation of linear discriminant analysis for the high dimension/small sample size problem. *Computational Statistics & Data Analysis*, 52(1):423–437.

Tellaroli, P. and Donato, M. (2016). *A Partial Clustering Algorithm with Automatic Estimation of the Number of Clusters and Identification of Outliers*. R package version 3.0.

Todorov, V. (2016). *rrcovHD: Robust Multivariate Methods for High Dimensional Data*. R package version 0.2-4.

Todorov, V. and Filzmoser, P. (2009). An object-oriented framework for robust multivariate analysis. *Journal of Statistical Software*, 32(3):1–47.

Torgo, L. (2010). *Data Mining with R, learning with case studies*. Chapman and Hall/CRC.

Torgo, L. (2016). *Data Mining with R, learning with case studies, 2nd edition*. Chapman and Hall/CRC.

Trefethen, L. N. and Bau III, D. (1997). *Numerical linear algebra*, volume 50. SIAM.

Tyler, D. E. (1994). Finite sample breakdown points of projection based multivariate location and scatter statistics. *The Annals of Statistics*, 22(2):1024–1044.

Tyler, D. E. (2010). A note on multivariate location and scatter statistics for sparse data sets. *Statistics & Probability Letters*, 80(17):1409–1413.

Tyler, D. E., Critchley, F., Dümbgen, L., and Oja, H. (2009). Invariant coordinate selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):549–592.

Ueberhuber, C. W. (2012a). *Numerical computation 1: methods, software, and analysis*. Springer Science & Business Media.

Ueberhuber, C. W. (2012b). *Numerical computation 2: methods, software, and analysis*. Springer Science & Business Media.

Vaissie, P., Monge, A., and Husson, F. (2016). *Factoshiny: Perform Factorial Analysis from 'FactoMineR' with a Shiny Application*. R package version 1.0.5.

van der Loo, M. (2010). *extremevalues, an R package for outlier detection in univariate data*. R package version 2.3.

Van der Sluis, A. (1969). Condition numbers and equilibration of matrices. *Numerische Mathematik*, 14(1):14–23.

Vargas N., J. A. (2003). Robust estimation in multivariate control charts for individual observations. *Journal of Quality Technology*, 35(4):367–376.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0.

Venkatasubramanian, V., Rengaswamy, R., Yin, K., and Kavuri, S. N. (2003). A review of process fault detection and diagnosis: Part I: Quantitative model-based methods. *Computers & Chemical Engineering*, 27(3):293–311.

Verbanck, M., Josse, J., and Husson, F. (2015). Regularised PCA to denoise and visualise data. *Statistics and Computing*, 25(2):471–486.

Virta, J. (2014). Some tools for linear dimension reduction. Master's thesis, University of Turku.

Wilkinson, L. (2016). *HDoutliers: Leland Wilkinson's Algorithm for Detecting Multidimensional Outliers*. R package version 0.8.

Wilks, S. (1962). *Mathematical Statistics*. John Wiley & Sons.

Willems, G., Joe, H., and Zamar, R. (2009). Diagnosing multivariate outliers detected by robust estimators. *Journal of Computational and Graphical Statistics*, 18(1):73–91.

Williams, K. (2016). *ldbod: Local Density-Based Outlier Detection*. R package version 0.1.0.

Witten, D. M. and Tibshirani, R. (2011). Penalized classification using Fisher's linear discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5):753–772.

Wu, M. and Jermaine, C. (2006). Outlier detection by sampling with accuracy guarantees. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 767–772. ACM.

Xiong, L., Chen, X., and Schneider, J. (2011). Direct robust matrix factorizatoin for anomaly detection. In *IEEE 11th International Conference on Data Mining (ICDM)*, pages 844–853. IEEE.

Yazici, B. and Yolacan, S. (2007). A comparison of various tests of normality. *Journal of Statistical Computation and Simulation*, 77(2):175–183.

Zhang, J., Lou, M., Ling, T. W., and Wang, H. (2004). Hos-miner: a system for detecting outlyting subspaces of high-dimensional data. In *Proceedings of the Thirtieth International Conference on Very Large Data Bases*, volume 30, pages 1265–1268. VLDB Endowment.

Zimek, A., Campello, R. J., and Sander, J. (2014a). Data perturbation for outlier detection ensembles. In *Proceedings of the 26th International Conference on Scientific and Statistical Database Management*, page 13. ACM.

Zimek, A., Campello, R. J., and Sander, J. (2014b). Ensembles for unsupervised outlier detection: challenges and research questions a position paper. volume 15, pages 11–22. ACM.

Zimek, A., Gaudet, M., Campello, R. J., and Sander, J. (2013). Subsampling for efficient and effective unsupervised outlier detection ensembles. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 428–436. ACM.

Zimek, A., Schubert, E., and Kriegel, H.-P. (2012). A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 5(5):363–387.

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2):265–286.

**Résumé**

La détection d'observations atypiques de manière non-supervisée est un enjeu crucial dans la pratique de la statistique. Dans le domaine de la détection de défauts industriels, cette tâche est d'une importance capitale pour assurer une production de haute qualité. Avec l'accroissement exponentiel du nombre de mesures effectuées sur les composants électroniques, la problématique de la grande dimension se pose lors de la recherche d'anomalies. Pour relever ce challenge, l'entreprise ippon innovation, spécialiste en statistique industrielle et détection d'anomalies, s'est associée au laboratoire de recherche TSE-R en finançant ce travail de thèse. Le premier chapitre commence par présenter le contexte du contrôle de qualité et les différentes procédures déjà mises en place, principalement dans les entreprises de semi-conducteurs pour l'automobile. Comme ces pratiques ne répondent pas aux nouvelles attentes requises par le traitement de données en grande dimension, d'autres solutions doivent être envisagées. La suite du chapitre résume l'ensemble des méthodes multivariées et non supervisées de détection d'observations atypiques existantes, en insistant tout particulièrement sur celles qui gèrent des données en grande dimension. Le Chapitre 2 montre théoriquement que la très connue distance de Mahalanobis n'est pas adaptée à la détection d'anomalies si celles-ci sont contenues dans un sous-espace de petite dimension alors que le nombre de variables est grand. Dans ce contexte, la méthode Invariant Coordinate Selection (ICS) est alors introduite comme une alternative intéressante à la mise en évidence de la structure des données atypiques. Une méthodologie pour sélectionner seulement les composantes d'intérêt est proposée et ses performances sont comparées aux standards habituels sur des simulations ainsi que sur des exemples réels industriels. Cette nouvelle procédure a été mise en œuvre dans un package R, ICSOutlier, présenté dans le Chapitre 3 ainsi que dans une application R shiny (package ICSShiny) qui rend son utilisation plus simple et plus attractive. Une des conséquences directes de l'augmentation du nombre de dimensions est la singularité des estimateurs de dispersion multivariés, dès que certaines variables sont colinéaires ou que leur nombre excède le nombre d'individus. Or, la définition d'ICS par Tyler et al. (2009) se base sur des estimateurs de dispersion définis positifs. Le Chapitre 4 envisage différentes pistes pour adapter le critère d'ICS et investigue de manière théorique les propriétés de chacune des propositions présentées. La question de l'affine invariance de la méthode est en particulier étudiée. Enfin le dernier chapitre, se consacre à l'algorithme développé pour l'entreprise. Bien que cet algorithme soit confidentiel, le chapitre donne les idées générales et précise les challenges relevés, notamment numériques.

**Mots-clés** : détection d'anomalies, ICS, distance de Mahalanobis, analyse multivariée, faible taille d'échantillon, haute fiabilité.

**Abstract**

The unsupervised outlier detection is a crucial issue in statistics. More specifically, in the industrial context of fault detection, this task is of great importance for ensuring a high quality production. With the exponential increase in the number of measurements on electronic components, the concern of high dimensional data arises in the identification of outlying observations. The ippon innovation company, an expert in industrial statistics and anomaly detection, wanted to deal with this new situation. So, it collaborated with the TSE-R research laboratory by financing this thesis work. The first chapter presents the quality control context and the different procedures mainly used in the automotive industry of semiconductors. However, these practices do not meet the new expectations required in dealing with high dimensional data, so other solutions need to be considered. The remainder of the chapter summarizes unsupervised multivariate methods for outlier detection, with a particular emphasis on those dealing with high dimensional data. Chapter 2 demonstrates that the well-known Mahalanobis distance presents some difficulties to detect the outlying observations that lie in a smaller subspace while the number of variables is large. In this context, the Invariant Coordinate Selection (ICS) method is introduced as an interesting alternative for highlighting the structure of outlierness. A methodology for selecting only the relevant components is proposed. A simulation study provides a comparison with benchmark methods. The performance of our proposal is also evaluated on real industrial data sets. This new procedure has been implemented in an R package, ICSOutlier, presented in Chapter 3, and in an R shiny application (package ICSShiny) that makes it more user-friendly. When the number of dimensions increases, the multivariate scatter matrices turn out to be singular as soon as some variables are collinear or if their number exceeds the number of individuals. However, in the presentation of ICS by Tyler et al. (2009), the scatter estimators are defined as positive definite matrices. Chapter 4 proposes three different ways for adapting the ICS method to singular scatter matrices and theoretically investigates their properties. The question of affine invariance is analyzed in particular. Finally, the last chapter is dedicated to the algorithm developed for the company. Although the algorithm is confidential, the chapter presents the main ideas and the challenges, mostly numerical, encountered during its development.

**Keywords**: anomaly detection, ICS, Mahalanobis distance, multivariate analysis, high reliability, low sample size.