Linear Inverse Problems in Structural Econometrics Estimation based on spectral decomposition and regularization*

Marine Carrasco
University of Rochester

Jean-Pierre Florens
Université de Toulouse (GREMAQ and IDEI)

Eric Renault
Université de Montréal (CIRANO and CIREQ)

September 2003

Abstract

This chapter provides an introduction to the estimation of the solution of inverse problems. The beginning of this chapter focuses on integral equations of the first kind. Solving these equations is particularly challenging as the solution does not necessarily exist, may not be unique, and is not continuous. As a result, a regularized (or smoothed) solution needs to be implemented. We review different regularization methods and study the properties of the estimator. Integral equations of the first kind appear, among others, in the generalized method of moments when the number of moment conditions is infinite and in the nonparametric estimation of instrumental variable regressions. At the end of this chapter, we investigate integral equations of the second kind, which solution may not be unique but is continuous. Such equations appear when estimating nonparametrically additive models and measurement error models.

Keywords: Fredholm equation, Generalized Method of Moments, Hilbert Schmidt operator, Instrumental variable, Tikhonov regularization.

JEL: C13, C14, C20.

*This chapter is a working draft for eventual publication in the *Handbook of Econometrics*, Vol 6. Carrasco gratefully aknowledges financial support from the National Science Foundation, grant # SES-0211418.

1. Introduction

1.1. Structural models and functional estimation

The objective of this chapter is to analyze functional estimation in structural econometric models. There exist different approaches to structural inference in econometrics and our presentation may be viewed as a non parametric extension of the basic example of structural models, namely the static linear simultaneous equations model (SEM). Let us consider Y a vector of random endogenous variables and Z a vector of exogenous random variables. A SEM is characterized by a system

$$B_\theta Y + C_\theta Z = U \tag{1.1}$$

where B_θ and C_θ are matrices that are functions of an unknown "structural" parameter θ and $E[U|Z] = 0$. The reduced form is a multivariate regression model

$$Y = \Pi Z + V \tag{1.2}$$

where Π is the matrix of ordinary regression coefficients. The relation between reduced and structural form is, in absence of higher moments restrictions, characterized by:

$$B_\theta \Pi + C_\theta = 0. \tag{1.3}$$

The two essential issues of structural modeling, the identification and the overidentification problems, follow from the consideration of Equation (1.3). The uniqueness of the solution in θ for given Π defines the identification problem. The existence of a solution (or restrictions imposed to Π to guarantee the existence) defines the overidentification question. The reduced form parameter Π can be estimated by OLS and if a unique solution in θ exists for any Π, it provides the Indirect Least Square estimate of θ. If the solution does not exist for any Π, θ can be estimated by a suitable minimization of $B_\theta \hat{\Pi} + C_\theta$ where $\hat{\Pi}$ is an estimator of Π.

We address in this chapter the issue of functional extension of this construction. The data generating process (DGP) is described by a stationary ergodic stochastic process which generates a sequence of observed realizations of a random vector X.

The structural econometric models considered in this chapter are about the stationary distribution of X. This distribution is characterized by its cumulative distribution function (c.d.f.) F while the functional parameter of interest is an element φ of some infinite dimensional Hilbert space. The structural econometric model defines the connection between φ and F under the form of a functional equation:

$$A(\varphi, F) = 0. \tag{1.4}$$

This equation extends Equation (1.3) and the definitions of identification (uniqueness of this solution) and of overidentification (constraints on F such that a solution exists) are analogous to the SEM case.

Estimation is also performed in the same line : F can be estimated by the empirical distribution of the sample or by a more sophisticated estimator (like kernel smoothing) belonging to the domain of A and φ is estimated by solving (1.4) or, in the presence of overidentification by a minimization of a suitable norm of $A(\varphi, F)$ after plugging in the estimator of F.

This framework may be clarified by some remarks.

1. All the variables are treated as random in our model and this construction seems to differ from the basic econometric models which are based on a distinction between exogenous or conditioning variables and endogenous variables. Actually this distinction may be used in our framework. Let X be decomposed into Y and Z and F into $F_Y(\ |Z = z)$ the conditional c.d.f. of Y given $Z = z$, and the marginal c.d.f. of Z, F_Z. Then, the exogeneity of Z is tantamount to the conjunction of two conditions.

   Firstly, the solution φ of (1.4) only depends on $F_Y(\ |Z = z)$ and φ is identified by the conditional model only. Secondly if $F_Y(\ |Z = z)$ and F_Z are "variations free" in a given statistical model defined by a family of sampling distributions (intuitively no restrictions link $F_Y(\ |Z = z)$ and F_Z), no information on $F_Y(\ |Z = z)$ (and then on φ) is lost by neglecting the estimation of F_Z. This definition fully encompasses the usual definition of exogeneity in terms of cuts (see Engle, Hendry and Richard (1983), Florens and Mouchart (1985)). Extension of that approach to sequential models and then to sequential or weak exogeneity is straightforward.

2. Our construction does not explicitly involve residuals or other unobservable variables. As it will be illustrated in the examples below, most of the structural econometric models are formalized by a relationship between observable and unobservable random elements. A first step in the analysis of these models is to express the relationship between the functional parameters of interest and the DGP, or, in our terminology, to specify the relation $A(\varphi, F) = 0$. We start our presentation at the second step of this approach and our analysis is devoted to the study of this equation and to its use for estimation.

3. The overidentification is handled by extending the definition of the parameter in order to estimate overidentified models. Even if $A(\varphi, F) = 0$ does not have a solution for a given F, the parameter φ is still defined as the minimum of a norm of $A(\varphi, F)$. Then φ can be estimated from an estimation of F which does not satisfy the overidentification constraints. This approach extends the original Generalized Method of Moments (GMM) treatment of overidentification. Another way to take into account overidentification constraints consists in estimating F under these constraints (the estimator of F is the nearest distribution to the empirical distribution for which there exists a solution, φ, of $A(\varphi, F) = 0$). This method extends the new approach to GMM called the empirical likelihood analysis (see Owen (2001) and references therein). In this chapter, we remain true to the first approach: actually if the equation $A(\varphi, F) = 0$ has no solution it will be replaced by the first order

condition of the minimization of a norm of $A(\varphi, F)$. In that case, this first order condition defines a functional equation usually still denoted $A(\varphi, F) = 0$.

1.2. Notation

In this chapter, X is a random element of a finite or infinite dimensional space $\mathcal{X}$. In most of the examples, $\mathcal{X}$ is a finite dimensional euclidean space ($\mathcal{X} \subset \mathbb{R}^m$) and the distribution on X, denoted F is assumed to belong to a set $\mathcal{F}$. If F is absolutely continuous, its density is denoted by f. Usually, X is decomposed into several components, $X = (Y, Z, W) \in \mathbb{R}^p \times \mathbb{R}^q \times \mathbb{R}^r (p + q + r = m)$ and marginal c.d.f. or probability density function (p.d.f.) are denoted by F_Y, F_Z, F_W. and f_Y, f_X, f_W respectively. Conditional c.d.f. are denoted by $F_Y(.|Z = z)$ or $F_Y(.|z)$ and conditional density by $f_Y(.|Z = z)$ or $f_Y(.|z)$. The sample may be an i.i.d. sample of X (denoted in that case $(x_i)_{i=1,...,n}$) or weakly dependent time series sample denoted $(x_t)_{t=1,...,T}$ in the dynamic case.

The paper focuses on the estimation of an infinite dimensional parameter denoted by φ, which is an element of a Hilbert space $\mathcal{H}$ (mathematical concepts are recalled in Section 2). In some particular cases, finite dimensional parameters are considered and this feature is underlined by the notation $\theta \in \Theta \subset \mathbb{R}^d$ for this particular case.

The structural model is expressed by an operator A from $\mathcal{H} \times \mathcal{F}$ into an Hilbert space $\mathcal{E}$ and defines the equation $A(\varphi, F) = 0$. The (possibly local) solution of this equation is denoted by:

$$\varphi = \Psi(F). \tag{1.5}$$

For statistical discussions, a specific notation for the true value is helpful and F_0 will denote the true c.d.f. (associated with the density f_0 and with true parameter φ_0 (or θ_0)). The estimators of the c.d.f. will be denoted by F_n in an i.i.d. setting or F_T in a dynamic environment.

The operator A may take various forms. Particular cases are linear operators with respect to F or to φ. The first case will be illustrated in the GMM example but most of the paper will be devoted to the study of linear operator relatively to φ. In that case, equation $A(\varphi, F) = 0$ can be rewritten :

$$A(\varphi, F) = K\varphi - r = 0 \tag{1.6}$$

where K is a linear operator from $\mathcal{H}$ to $\mathcal{E}$ depending on F and r is an element of $\mathcal{E}$ and is also a function of F. The properties of K are essential and we will present different examples of integral or differential operators. More generally, A may be non linear either with respect to F or to φ but, as usual in functional analysis, most of the analysis of non linear operators may be done locally (around the true value typically) and reduces to the linear case. Game theoretic model or surplus estimation give examples of non linear models.

The problem of solving Equation (1.4) enters in the class of inverse problems. An inverse problem consists into the resolution of an equation where the elements of the

equations are imperfectly known. In the linear case, the equation is $K\varphi = r$ and F is not exactly known but only estimated. Then, r is also imperfectly known. The econometric situation is more complex than most of the inverse problems studied in the statistical literature because K is also only imperfectly known. According to the classification proposed by Vapnik (1998), the stochastic inverse problems of interest in this chapter are more often than not characterized by equations with both the operator and the right-hand side approximately defined. Inverse problems are said to be well posed if a unique solution exists and depends continuously of the imperfectly known elements of the equation. In our notation, this means that Ψ in (1.5) exists as a function of F and is continuous. Then, if F is replaced by F_n, the solution φ_n of $A(\varphi_n, F_n) = 0$ exists and the convergence of F_n to F_0 implies by continuity the convergence of φ_n to φ_0. Unfortunately a large class of inverse problems relevant to econometric applications are not well posed (they are then said to be ill-posed in the Hadamard sense, see e.g. Kress (1999), Vapnik (1998)).

1.3. Examples

This section presents various examples of inverse problems motivated by structural econometric models. We will start by the GMM example, which is the most familiar to econometricians. Subsequently, we present several examples of linear (w.r.t. φ) inverse problems. The last two examples are devoted to non linear inverse problems.

1.3.1. Generalized Method of Moments (GMM)

Let us assume that X is m dimensional and the parameter of interest θ is also finite dimensional ($\theta \in \Theta \subset \mathbb{R}^d$). We consider a function

$$h : \mathbb{R}^m \times \Theta \to \mathcal{E} \tag{1.7}$$

and the equation connecting θ and F is defined by:

$$A(\theta, F) = E^F(h(X, \theta)) = 0 \tag{1.8}$$

A particular case is given by $h(X, \theta) = \mu(X) - \theta$ where θ is exactly the expectation of a transformation μ of the data. More generally, θ may be replaced by an infinite dimensional parameter φ but we do not consider here this extension.

The GMM method was introduced by Hansen (1982) and has received numerous extensions (see Ai and Chen (1999) for the case of an infinite dimensional parameter). GMM consists in estimating θ by solving an inverse problem linear in F but non linear in θ. It is usually assumed that θ is identified i.e. that θ is uniquely characterized by Equation (1.8). Econometric specifications are generally overidentified and a solution to (1.8) only exists for some particular F, including the true DGP F_0, under the hypothesis of correct specification of the model. The c.d.f F is estimated by the empirical distribution and the equation (1.8) becomes:

$$\frac{1}{n}\sum_{i=1}^{n} h(x_i, \theta) = 0, \tag{1.9}$$

which has no solution in general. Overidentification is treated by an extension of the definition of θ as follows:

$$\theta = \arg\min_{\theta} \|BE^F(h)\|^2 \tag{1.10}$$

where B is a linear operator in $\mathcal{E}$ and $\|\|$ denotes the norm in $\mathcal{E}$. This definition coincides with (1.8) if F satisfies the overidentification constraints. Following Equation (1.10), the estimator is:

$$\hat{\theta}_n = \arg\min_{\theta} \|B_n \left(\frac{1}{n} \sum_{i=1}^{n} h(x_i, \theta) \right) \|^2 \tag{1.11}$$

where B_n is a sequence of operators converging to B. If the number of moment conditions is finite, B_n and B are square matrices.

As θ is finite dimensional, the inverse problem generated by the first order conditions of (1.10) or (1.11) is well posed and consistency of the estimators follows from standard regularity conditions. As it will be illustrated in Section 6, an ill-posed inverse problem arises if the number of moment conditions is infinite and if optimal GMM is used. In finite dimension, optimal GMM is obtained for a specific weighting matrix, $B = \Sigma^{-\frac{1}{2}}$, where Σ is the asymptotic variance of $\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} h(x_i, \theta)\right)$ ($\Sigma = Var(h)$ in i.i.d. sampling). In the general case, optimal GMM requires the minimization of $\|g\|^2$ where

$$\Sigma^{\frac{1}{2}} g = E^F(h) \tag{1.12}$$

The function g is then the solution of a linear inverse problem. If the dimension of h is not finite, Equation (1.12) defines an ill-posed inverse problem, which requires a regularization scheme (see Section 3)

1.3.2. Instrumental variables

Instrumental regression is a possible strategy to perform non parametric estimation when explanatory variables are endogenous. Let us decompose X into (Y, X, W) where $Y \in \mathbb{R}$, $Z \in \mathbb{R}^q$, $W \in \mathbb{R}^r$. The subvectors Z and W may have elements in common. The econometrician starts with a relation

$$Y = \varphi(Z) + U \tag{1.13}$$

where U is a random term which does not satisfy $E(U|Z) = 0$. This assumption is replaced by the more general hypothesis

$$E(U|W) = 0 \tag{1.14}$$

and W is called the set of instrumental variables. Condition (1.14) defines φ as the solution of an integral equation. In terms of density, (1.14) means that

$$A(\varphi, F) = \int \varphi(z) f_Z(z|W = w) dz - \int y f_Y(y|W = w) dy = 0 \tag{1.15}$$

Using previous notation, the first part of (1.15) is denoted $K\varphi$ and the second part is equal to r.

This expression is actually linear in φ and in F (after multiplication by $f_W(w)$) but the linearity with respect to the distribution does not play any role. As we will see later, this problem is essentially nonlinear in F because, even if the denominator can be eliminated in (1.15), the treatment of overidentification and of regularization will necessarily reintroduce the denominator.

Instrumental regression introduced in (1.15) can be generalized to local instrumental regression and to generalized local instrumental regression. These extensions are relevant in more complex models than (1.13) where in particular the additive form is not preserved (see for such a treatment, Florens, Heckman, Meghir, and Vytlacil (2002)). For example, consider the equation

$$Y = \varphi(Z) + Z\varepsilon + U \tag{1.16}$$

where Z is scalar and ε is a random unobservable heterogeneity component. It can be proved that, under a set of identification assumptions, φ satisfies the equations :

$$A_j(\varphi, F) = E^F\left(\frac{\partial\varphi(Z)}{\partial Z}|W = w\right) - \frac{\dfrac{\partial}{\partial W_j}E(Y|W = w)}{\dfrac{\partial}{\partial W_j}E(Z|W = w)} = 0 \tag{1.17}$$

for any $j = 1, ..., r$. This equation, linear with respect to φ, combines integral and differential operators.

Instrumental variable estimation and its local extension define ill-posed inverse problems as it will be seen in Section 5.

1.3.3. Deconvolution

Another classical example of ill-posed inverse problem is given by the deconvolution problem. Let us assume that X, Y, Z be three scalar random elements such that

$$Y = X + Z \tag{1.18}$$

Only Y is observable. The two components X and Z are independent. The density of Z (the error term) is known and denoted g. The parameter of interest is the density φ of X. Then φ is solution of:

$$\begin{aligned} A(\varphi, F) &= \int \varphi(y)g(x-y)dy - f(x) = 0 \\ &= K\varphi - r \end{aligned} \tag{1.19}$$

This example is comparable to the instrumental variables case but only the r.h.s. $r = f$ is unknown whereas the operator K is given.

1.3.4. Regression with many regressors

This example constitutes also a case of linear ill-posed inverse problems. Let us consider a regression model where the regressors are indexed by τ belonging to an infinite index set provided with a measure Π. The model says:

$$Y = \int Z(\tau)\varphi(\tau)\Pi(d\tau) + U \tag{1.20}$$

where $E(U|(Z(\tau))_\tau) = 0$ and φ is the parameter of interest and is infinite dimensional. Examples of regression with many regressors are now common in macroeconomics (see Stock and Watson (2002) or Forni and Reichlin (1998) for two presentations of this topic).

Let us assume that Y and $(Z(\tau))_\tau$ are observable. Various treatments of (1.20) can be done and we just consider the following analysis. The conditional moment equation $E(U|(Z(\tau))_\tau) = 0$ implies an infinite number of conditions for any τ which implies:

$$E(Z(\tau)U) = 0, \qquad \forall\tau$$

or equivalently

$$\int E^F(Z(\tau)Z(\rho))\varphi(\rho)\Pi(d\rho) - E^F(YZ(\tau)) = 0, \quad \forall\tau \tag{1.21}$$

This equation generalizes to an infinite number of regressors the usual normal equations of the linear regression. The inverse problem defined in (1.21) is linear in both F and φ but it is ill posed. An intuitive argument to illustrate this issue is to consider the estimation using a finite number of observations of the second moment operator $E^F(Z(\tau)Z(\rho))$ which is infinite dimensional. The resulting multicolinearity problem is solved by a ridge regression. The "infinite matrix" $E^F(Z(.)Z(.))$ is replaced by $\alpha I + E^F(Z(.)Z(.))$ where I is the identity and α a positive number or by a reduction of the set of regressors to the first principal components. These two solutions are particular examples of regularization methods (namely the Tikhonov and the spectral cut-off regularizations), which will be introduced in Section 3.

1.3.5. Additive models

The nature of the integral equations generated by this example and by the next one is very different from that of the three previous examples. We consider an additive regression model:

$$Y = \varphi(Z) + \psi(W) + U \tag{1.22}$$

where $E(U|Z,W) = 0$ and $X = (Y,Z,W)$ is the observable element. The parameters of interest are the two functions φ and ψ. The approach we propose here is the backfitting

approach (see Hastie and Tibshirani (1990)). Other treatments of additive models have been considered in the literature (see Pagan and Ullah (1999)). Equation (1.22) implies

$$\begin{aligned} E^F(Y|Z=z) &= \varphi(z) + E^F(\psi(W)|Z=z) \\ E^F(Y|W=w) &= E^F(\varphi(Z)|W=w) + \psi(w) \end{aligned} \tag{1.23}$$

and by substitution

$$\begin{aligned} &\varphi(z) - E^F(E^F(\varphi(Z)|W)|Z=z) \\ &\quad = E^F(Y|Z=z) - E^F(E^F(Y|W)|Z=z) \end{aligned} \tag{1.24}$$

or, in our notations:

$$(I-K)\,\varphi = r$$

where $K = E^F(E^F(\,.\,|W)|Z)$.

An analogous equation characterizes ψ. Actually even if (1.22) is not well specified, these equations provide the best approximation of the regression of Y given Z and W by an additive form. Equation (1.24) is a linear integral equation and even if this inverse problem is ill-posed because K is not one-to-one (φ is only determined up to a constant term), the solution is still continuous and therefore the difficulty is not as important as that of the previous examples.

1.3.6. Measurement-error models or non parametric analysis of panel data

We denote by η an unobservable random variable for which two measurements are available Y_1 and Y_2. These measurements are affected by a bias dependent of observable variables Z_1 and Z_2. More formally:

$$\begin{cases} Y_1 = \eta + \varphi(Z_1) + U_1 & E(U_1|\eta, Z_1, Z_2) = 0 \\ Y_2 = \eta + \varphi(Z_2) + U_2 & E(U_2|\eta, Z_1, Z_2) = 0 \end{cases} \tag{1.25}$$

An i.i.d. sample $(y_{1i}, y_{2i}, \eta_i, z_{1i}, z_{2i})$ is drawn but the η_i are unobservable. Equivalently this model may be seen as a two period panel data with individual effects η_i.

The parameter of interest is the "bias function" φ, identical for the two observations. In the measurement context, it is natural to assume that the distribution of the observable is independent of the order of the observations, or, equivalently (Y_1, Z_1, Y_2, Z_2) is distributed as (Y_2, Z_2, Y_1, Z_1). This assumption is not relevant in a dynamic context.

The model is transformed in order to eliminate the unobservable variable by difference:

$$Y = \varphi(Z_2) - \varphi(Z_1) + U \tag{1.26}$$

where $Y = Y_2 - Y_1$, $\quad U = U_2 - U_1$, and $E(U|Z_1, Z_2) = 0$.

This model is similar to an additive model except for the symmetry between the variables and the fact that, with the notation of (1.22), φ and ψ are identical. An application of this model may be found in Gaspar and Florens (1998) where y_{1i} and y_{2i} are two measurements of the level of the ocean in location i by a satellite radar altimeter, η_i is the true level and φ is the "sea state bias" depending on the waves' height and the wind speed (Z_{1i} and Z_{2i} are both two dimensional).

The model is treated through the relation:

$$E(Y|Z_2 = z_2) = \varphi(z_2) - E(\varphi(Z_1)|Z_2 = z_2) \tag{1.27}$$

which defines an integral equation $K\varphi = r$. The exchangeable property between the variables implies that conditioning on Z_1 gives the same equation.

1.3.7. Game theoretic model

This example and the next one present economic models formalized by non linear inverse problems. The analysis of non linear functional equations raises numerous questions: uniqueness and existence of the solution, asymptotic properties of the estimator, implementation of the estimation procedure and numerical computation of the solution.

Most of these questions are usually solved locally by a linear approximation of the non linear problem deduced from a suitable concept of derivative. A strong concept of derivation (typically Frechet derivative) is needed to deal with the implicit form of the model which requires the use of the Implicit Function theorem.

The first example of nonlinear inverse problems follows from the strategic behavior of the players in a game. Let us assume that for each game, each player receives a random signal or type denoted by ξ and plays an action X. The signal is generated by a probability described by its c.d.f. φ and the players all adopt a strategy σ dependent on φ which associates X with ξ ,i.e.

$$X = \sigma_\varphi(\xi) \tag{1.28}$$

The strategy σ_φ is determined as an equilibrium of the game (e.g. Nash equilibrium) or by an approximation of the equilibrium (bounded rationality behavior). The signal ξ is a private knowledge for the player but is unobservable for the econometrician and the c.d.f. φ is common knowledge for the players but is unknown for the statistician. The strategy σ_φ is determined from the rule of the game and by the assumptions on the behavior of the players. The essential feature of the game theoretic model from a statistical viewpoint is that the relation between the unobservable and the observable variables depends on the distribution of the unobservable component. The parameter of interest is the c.d.f. φ of the signals.

Let us restrict our attention to cases where ξ and X are scalar and where σ_φ is strictly increasing. Then the c.d.f. F of the observable X is connected with φ by:

$$A(\varphi, F) = F \circ \sigma_\varphi - \varphi = 0 \tag{1.29}$$

If the signals are i.i.d. across the different players and different games, F can be estimated by a smooth transformation of the empirical distribution and Equation (1.29) is solved in φ. The complexity of this relation can be illustrated by the auction model. In the private value first price auction model, ξ is the value of the object and X the bid. If the number of bidders is $N+1$ the strategy function is equal to:

$$X = \xi - \frac{\int_{\underline{\xi}}^{\xi} \varphi^N(u)du}{\varphi^N(\xi)} \tag{1.30}$$

where $[\underline{\xi}, \bar{\xi}]$ is the support of ξ and $\varphi^N(u) = [\varphi(u)]^N$ is the c.d.f. of the maximum private value among N players.

Model (1.29) may be extended to a non iid setting (depending on exogenous variables) or to the case where σ_φ is partially unknown. The analysis of this model has been done by Guerre, Perrigne and Vuong (2000) in a non parametric context. The framework of inverse problem is used by Florens, Protopopescu and Richard (1997).

1.3.8. Solution of a differential equation

In several models like the analysis of the consumer surplus, the function of interest is solution of a differential equation depending on the data generating process.

Consider for example a class of problem where $X = (Y, Z, W) \in \mathbb{R}^3$ is i.i.d., F is the c.d.f. of X and the parameter φ verifies:

$$\frac{d}{dz}\varphi(z) = m_F(z, \varphi(z)) \tag{1.31}$$

when m_F is a regular function depending on F. A first example is

$$m_F(z, w) = E^F(Y|Z = z, W = w)) \tag{1.32}$$

but more complex examples may be constructed in order to take into account the endogeneity of one or two variables. For example Z may be endogenous and m_F may be defined by:

$$E(Y|W_1 = w_1, W_2 = w_2) = E(m_F(Z, W_1)|W_1 = w_1, W_2 = w_2) \tag{1.33}$$

Economic applications can be found in Hausman (1981, 1985) and Hausman and Newey (1995) and a theoretical treatment of these two problems is given by Vanhems (2000) and Loubes and Vanhems (2001).

1.4. Organization of the chapter

Section 2 reviews the basic definitions and properties of operators in Hilbert spaces. The focus is on compact operators that have the advantage to have a discrete spectrum. We recall some laws of large numbers and central limit theorems for Hilbert valued random

elements. Finally, we discuss how to estimate the spectrum of a compact operator and how to estimate the operators themselves.

Section 3 is devoted to solving integral equation of the first kind. As these equations are ill-posed, the solution needs to be regularized (or smoothed). We investigate the properties of the regularized solutions for different types of regularizations.

In Section 4, we show under suitable assumptions, the consistency and asymptotic normality of regularized solutions.

Section 5 detail three examples: the infinite number of regressors, the deconvolution and the instrumental variables estimation.

Section 6 has two parts. First, it recalls the main results relative to reproducing kernels. Reproducing kernel theory is closely related to that of the integral equations of the first kind. Second, we explain the extension of GMM to a continuum of moment conditions and how the GMM objective function reduces to the norm of the moment functions in a specific reproducing kernel Hilbert space. Several examples are provided.

Section 7 tackles the problem of solving integral equation of the second kind. A typical example of such a problem is the additive model introduced earlier.

2. Spaces and Operators

The purpose of this section is to introduce terminology and to state the main properties of operators in Hilbert spaces which are used in our econometric applications. Most of these results can be found in Debnath and Mikusinsky (1999) and Kress (1999).

2.1. Hilbert spaces

We start by recalling some of the basic concepts of analysis. In the sequel, $\mathbf{C}$ denotes the set of complex numbers. A vector space equipped by a norm is called a normed space. A sequence (φ_n) of elements in a normed space is called a Cauchy sequence if for every $\varepsilon > 0$ there exists an integer $N(\varepsilon)$ such that

$$\|\varphi_n - \varphi_m\| < \varepsilon$$

for all $n,\ m \geq N(\varepsilon)$, i.e, if $\lim_{n,m\to\infty} \|\varphi_n - \varphi_m\| = 0$. A space S is complete if every Cauchy sequence converges to an element in S. A complete normed vector space is called a Banach space.

Let $(E, \mathcal{E}, \Pi)$ be a probability space and

$$L_C^p(E, \mathcal{E}, \Pi) = \left\{ f : E \to \mathbf{C} \text{ measurable s.t. } \|f\| \equiv \left(\int |f|^p \, d\Pi \right)^{1/p} < \infty \right\}, p \geq 1$$

$L_C^p(E, \mathcal{E}, \Pi)$ is a Banach space. If we only consider functions valued in $\mathbb{R}$ this space is still a Banach space and is denoted in that case by L^p (we drop the subscript C). In the sequel, we also use the following notation. If E is a subset of $\mathbf{R}^p$, then the $\sigma-$field $\mathcal{E}$ will always be the Borel $\sigma-$field and will be omitted in the notation $L^p(\mathbf{R}^p, \Pi)$. If Π

has a density π with respect to Lebesgue measure, Π will be replaced by π. If the pdf is uniform, it will be omitted in the notation.

Definition 2.1 (Inner product). *Let H be a complex vector space. A mapping $\langle , \rangle : H \times H \to \mathbf{C}$ is called an inner product in H if for any $\varphi, \psi, \xi \in H$ and $\alpha, \beta \in \mathbf{C}$ the following conditions are satisfied:*
(a) $\langle \varphi, \psi \rangle = \overline{\langle \psi, \varphi \rangle}$ (the bar denotes the complex conjugate),
(b) $\langle \alpha\varphi + \beta\psi, \xi \rangle = \alpha \langle \varphi, \xi \rangle + \beta \langle \psi, \xi \rangle$,
(c) $\langle \varphi, \varphi \rangle \geq 0$ and $\langle \varphi, \varphi \rangle = 0 \Longleftrightarrow \varphi = 0$.
A vector space equipped by an inner product is called an inner product space.

Example. The space $\mathbf{C}^N$ of ordered N-tuples $x = (x_1, ..., x_N)$ of complex numbers, with the inner product defined by

$$\langle x, y \rangle = \sum_{l=1}^{N} x_l \overline{y_l}$$

is an inner product space

Example. The space l^2 of all sequences $(x_1, x_2, ...)$ of complex numbers such that $\sum_{j=1}^{\infty} |x_j|^2 < \infty$ with the inner product defined by $\langle x, y \rangle = \sum_{j=1}^{\infty} x_j \overline{y_j}$ for $x = (x_1, x_2, ...)$ and $y = (y_1, y_2, ...)$ is an infinite dimensional inner product space.

Example. The space $L_C^2(E, \mathcal{E}, \Pi)$ associated with the inner product defined by

$$\langle \varphi, \psi \rangle = \int \varphi \overline{\psi} d\Pi$$

is an inner product space. On the other hand, $L_C^p(E, \mathcal{E}, \Pi)$ is not a inner product space if $p \neq 2$.

An inner product satisfies the Cauchy-Schwartz inequality, that is,

$$|\langle \varphi, \psi \rangle|^2 \leq \langle \varphi, \varphi \rangle \langle \psi, \psi \rangle$$

for all φ, $\psi \in H$. Remark that $\langle \varphi, \varphi \rangle$ is real because $\langle \varphi, \varphi \rangle = \overline{\langle \varphi, \varphi \rangle}$. It actually defines a norm $\|\varphi\| = \langle \varphi, \varphi \rangle^{1/2}$ (this is the norm induced by the inner product $\langle , \rangle$).

Definition 2.2 (Hilbert space). *If an inner product space is complete in the induced norm, it is called a Hilbert space.*

A standard theorem in functional analysis guarantees that every inner product space H can be completed to form a Hilbert space $\mathcal{H}$. Such a Hilbert space is said to be the completion of H.

Example. $\mathbf{C}^N$, l^2 and $L^2(\mathbb{R}, \Pi)$ are Hilbert spaces.

Example. (Sobolev space) Let $\Omega = [a, b]$ be an interval of $\mathbb{R}$. Denote by $\tilde{H}^m(\Omega)$, $m = 1, 2, ...$, the space of all complex-valued functions $\varphi \in \mathcal{C}^m$ such that for all $|l| \leq m$, $\varphi^{(l)} = \partial^l \varphi(\tau) / \partial \tau^l \in L^2(\Omega)$. The inner product on $\tilde{H}^m(\Omega)$ is

$$\langle \varphi, \psi \rangle = \int_a^b \sum_{l=0}^{m} \varphi^{(l)}(\tau) \overline{\psi^{(l)}(\tau)} d\tau.$$

$\tilde{H}^m(\Omega)$ is an inner product space but it is not a Hilbert space because it is not complete. The completion of $\tilde{H}^m(\Omega)$, denoted $H^m(\Omega)$, is a Hilbert space.

Definition 2.3 (Convergence). *A sequence (φ_n) of vectors in an inner product space H is called strongly convergent to a vector $\varphi \in H$ if $\|\varphi_n - \varphi\| \to 0$ as $n \to \infty$.*

Remark that if (φ_n) converges strongly to φ in H then $\langle \varphi_n, \psi \rangle \to \langle \varphi, \psi \rangle$ as $n \to \infty$, for every $\psi \in H$. The converse is false.

Definition 2.4. *Let H be an inner product space. A sequence (φ_n) of nonzero vectors in H is called an orthogonal sequence if $\langle \varphi_m, \varphi_n \rangle = 0$ for $n \neq m$. If in addition $\|\varphi_n\| = 1$ for all n, it is called orthonormal sequence.*

Example. Let $\pi(x)$ be the pdf of a normal with mean μ and variance σ^2. Denote by ϕ_j the Hermite polynomials of degree j:

$$\phi_j(x) = (-1)^j \frac{\frac{d^j \pi}{dx^j}}{\pi}. \tag{2.1}$$

The functions $\phi_j(x)$ form an orthogonal system in $L^2(\mathbb{R}, \pi)$.

Any sequence of vectors (ψ_j) in an inner product space that is linearly independent, i.e.,

$$\sum_{j=1}^{\infty} \alpha_j \psi_j = 0 \Rightarrow \alpha_j = 0 \; \forall j = 1, 2, ...$$

can be transformed into an orthonormal sequence by the method called Gram-Schmidt orthonormalization process. This process consists in the following steps. Given (ψ_j), define a sequence (φ_j) inductively as

$$\begin{aligned}
\varphi_1 &= \frac{\psi_1}{\|\psi_1\|}, \\
\varphi_2 &= \frac{\psi_2 - \langle \psi_2, \varphi_1 \rangle \varphi_1}{\|\psi_2 - \langle \psi_2, \varphi_1 \rangle \varphi_1\|} \\
&\vdots \\
\varphi_n &= \frac{\psi_n - \sum_{l=1}^{n-1} \langle \psi_n, \varphi_l \rangle \varphi_l}{\left\|\psi_n - \sum_{l=1}^{n-1} \langle \psi_n, \varphi_l \rangle \varphi_l\right\|}.
\end{aligned}$$

As a result, (φ_j) is orthonormal and any linear combinations of vectors $\varphi_1, ..., \varphi_n$ is also a linear combinations of $\psi_1, ..., \psi_n$ and vice versa.

Theorem 2.5 (Pythagorean formula). *If $\varphi_1, ..., \varphi_n$ are orthogonal vectors in an inner product space, then*

$$\left\| \sum_{j=1}^{n} \varphi_j \right\|^2 = \sum_{j=1}^{n} \left\| \varphi_j \right\|^2 .$$

From the Pythagorean formula, it can be seen that the α_l that minimize

$$\left\| \varphi - \sum_{j=1}^{n} \alpha_j \varphi_j \right\|$$

are such that $\alpha_j = \left\langle \varphi, \varphi_j \right\rangle$. Moreover

$$\sum_{j=1}^{n} \left| \left\langle \varphi, \varphi_j \right\rangle \right|^2 \leq \left\| \varphi \right\|^2 . \tag{2.2}$$

Hence the series $\sum_{j=1}^{\infty} \left| \left\langle \varphi, \varphi_j \right\rangle \right|^2$ converges for every $\varphi \in H$. The expansion

$$\varphi = \sum_{j=1}^{\infty} \left\langle \varphi, \varphi_j \right\rangle \varphi_j \tag{2.3}$$

is called a generalized Fourier series of φ. In general, we do not know whether the series in (2.3) is convergent. Below we give a sufficient condition for convergence.

Definition 2.6 (Complete orthonormal sequence). *An orthonormal sequence $\left(\varphi_j\right)$ in an inner product space H is said to be complete if for every $\varphi \in H$, we have*

$$\varphi = \sum_{j=1}^{\infty} \left\langle \varphi, \varphi_j \right\rangle \varphi_j$$

where the equality means

$$\lim_{n \to \infty} \left\| \varphi - \sum_{j=1}^{n} \left\langle \varphi, \varphi_j \right\rangle \varphi_j \right\| = 0$$

where $\|.\|$ is the norm in H.

A complete orthonormal sequence $\left(\varphi_j\right)$ in an inner product space H is an orthonormal basis in H, that is every $\varphi \in H$ has a unique representation $\varphi = \sum_{j=1}^{\infty} \alpha_j \varphi_j$ where $\alpha_l \in \mathbf{C}$. If $\left(\varphi_j\right)$ is a complete orthonormal sequence in an inner product space H then the set

$$\operatorname{span} \{\varphi_1, \varphi_2, ...\} = \left\{ \sum_{j=1}^{n} \alpha_j \varphi_j : \forall n \in \mathbf{N},\ \forall \alpha_1, ..., \alpha_n \in \mathbf{C} \right\}$$

is dense in H.

Theorem 2.7. *An orthonormal sequence* $\left(\varphi_j\right)$ *in a Hilbert space* $\mathcal{H}$ *is complete if and only if* $\left\langle \varphi, \varphi_j \right\rangle = 0$ *for all* $j = 1, 2, ...$ *implies* $\varphi = 0$.

Theorem 2.8 (Parseval's formula). *An orthonormal sequence* $\left(\varphi_j\right)$ *in a Hilbert space* $\mathcal{H}$ *is complete if and only if*

$$\|\varphi\|^2 = \sum_{j=1}^{\infty} \left|\left\langle \varphi, \varphi_j \right\rangle\right|^2 \tag{2.4}$$

for every $\varphi \in \mathcal{H}$.

Definition 2.9 (Separable space). *A Hilbert space is called separable if it contains a complete orthonormal sequence.*

Example. A complete orthonormal sequence in $L^2\left([-\pi, \pi]\right)$ is given by

$$\phi_j\left(x\right) = \frac{e^{ijx}}{\sqrt{2\pi}},\ j = ..., -1, 0, 1, ...$$

Hence, the space $L^2\left([-\pi, \pi]\right)$ is separable.

Theorem 2.10. *Every separable Hilbert space contains a countable dense subset.*

2.2. Definitions and basic properties of operators

In the sequel, we denote $K : \mathcal{H} \to \mathcal{E}$ the operator that maps a Hilbert space $\mathcal{H}$ (with norm $\|.\|_{\mathcal{H}}$) into a Hilbert space $\mathcal{E}$ (with norm $\|.\|_{\mathcal{E}}$).

Definition 2.11. *An operator* $K : \mathcal{H} \to \mathcal{E}$ *is called linear if*

$$K\left(\alpha\varphi + \beta\psi\right) = \alpha K\varphi + \beta K\psi$$

for all φ, $\psi \in \mathcal{H}$ *and all* $\alpha, \beta \in \mathbf{C}$.

Definition 2.12. *(i) The null space of* $K : \mathcal{H} \to \mathcal{E}$ *is the set* $\mathcal{N}\left(K\right) = \left\{\varphi \in \mathcal{H} : K\varphi = 0\right\}$.
(ii) The range of $K : \mathcal{H} \to \mathcal{E}$ *is the set* $\mathcal{R}(K) = \left\{\psi \in \mathcal{E}\ : \psi = K\varphi \text{ for some } \varphi \in \mathcal{H}\right\}$.
(iii) The domain of $K : \mathcal{H} \to \mathcal{E}$ *is the subset of* $\mathcal{H}$ *denoted* $\mathcal{D}(K)$ *on which* K *is defined.*
(iv) An operator is called finite dimensional if its range is of finite dimension.

Theorem 2.13. *A linear operator is continuous if it is continuous at one element.*

Definition 2.14. *A linear operator* $K : \mathcal{H} \to \mathcal{E}$ *is called bounded if there exists a positive number* C *such that*

$$\|K\varphi\|_{\mathcal{E}} \leq C\,\|\varphi\|_{\mathcal{H}}$$

for all $\varphi \in \mathcal{H}$.

Definition 2.15. *The norm of a bounded operator K is defined as*

$$\|K\| \equiv \sup_{\|\varphi\| \leq 1} \|K\varphi\|_{\mathcal{E}}$$

Theorem 2.16. *A linear operator is continuous if and only if it is bounded.*

Example. The identity operator defined by $\mathcal{I}\varphi = \varphi$ for all $\varphi \in \mathcal{H}$ is bounded with $\|\mathcal{I}\| = 1$.

Example. Consider the differential operator:

$$(D\varphi)(x) = \frac{d\varphi(\tau)}{d\tau} = \varphi'(\tau)$$

defined on the space $E_1 = \{\varphi \in L^2([-\pi, \pi]) : \varphi' \in L^2([-\pi, \pi])\}$ with norm $\|\varphi\| = \sqrt{\int_{-\pi}^{\pi} |f(\tau)|^2 d\tau}$. For $\varphi_j(\tau) = \sin j\tau$, $j = 1, 2, ...$, we have $\|\varphi_j\| = \sqrt{\int_{-\pi}^{\pi} |\sin(j\tau)|^2 d\tau} = \sqrt{\pi}$ and $\|D\varphi_j\| = \sqrt{\int_{-\pi}^{\pi} |j\cos(j\tau)|^2 d\tau} = j\sqrt{\pi}$. Therefore $\|D\varphi_j\| = j\|\varphi_j\|$ proving that the differential operator is not bounded.

Theorem 2.17. *Each linear operator K from a finite dimensional normed space $\mathcal{H}$ into a normed space $\mathcal{E}$ is bounded.*

An important class of linear operators are valued in $\mathbf{C}$ and they are characterized by Riesz theorem. From (2.2), we know that for any fixed vector g in an inner product space H, the formula $G(\varphi) = \langle \varphi, g \rangle$ defines a bounded linear functional on H. It turns out that if H is a Hilbert space, then every bounded linear functional is of this form.

Theorem 2.18 (Riesz). *Let $\mathcal{H}$ be a Hilbert space. Then for each bounded linear function $G : \mathcal{H} \to \mathbf{C}$ there exists a unique element $g \in \mathcal{H}$ such that*

$$G(\varphi) = \langle \varphi, g \rangle$$

for all $\varphi \in \mathcal{H}$. The norms of the element g and the linear function F coincide

$$\|g\|_{\mathcal{H}} = \|G\|$$

where $\|.\|_{\mathcal{H}}$ is the norm in $\mathcal{H}$ and $\|.\|$ is the operator norm.

Definition 2.19 (Hilbert space isomorphism). *A Hilbert space $\mathcal{H}_1$ is said to be isometrically isomorphic (congruent) to a Hilbert space $\mathcal{H}_2$ if there exists a one-to-one linear mapping J from $\mathcal{H}_1$ to $\mathcal{H}_2$ such that*

$$\langle J(\varphi), J(\psi) \rangle_{\mathcal{H}_2} = \langle \varphi, \psi \rangle_{\mathcal{H}_1}$$

for all $\varphi, \psi \in \mathcal{H}_1$. Such a mapping J is called a Hilbert space isomorphism (or congruence) from $\mathcal{H}_1$ to $\mathcal{H}_2$.

The terminology "congruence" is used by Parzen (1959, 1970).

Theorem 2.20. *Let $\mathcal{H}$ be a separable Hilbert space.*
(a) If $\mathcal{H}$ is infinite dimensional, then it is isometrically isomorphic to l^2.
(b) If $\mathcal{H}$ has a dimension N, then it is isometrically isomorphic to $\mathbf{C}^N$.

A consequence of Theorem 2.20 is that two separable Hilbert spaces of same dimensions (finite or infinite) are isometrically isomorphic.

Theorem 2.21. *Let $\mathcal{H}$ and $\mathcal{E}$ be Hilbert spaces and let $K : \mathcal{H} \to \mathcal{E}$ be a bounded operator. Then there exists a uniquely determined linear operator $K^* : \mathcal{E} \to \mathcal{H}$ with the property*

$$\langle K\varphi, \psi \rangle_{\mathcal{E}} = \langle \varphi, K^*\psi \rangle_{\mathcal{H}}$$

for all $\varphi \in \mathcal{H}$ and $\psi \in \mathcal{E}$. Moreover the operator K^* is bounded and $\|K\| = \|K^*\|$. K^* is called the adjoint operator of K.

Riesz Theorem 2.18 implies that, in Hilbert spaces, the adjoint of a bounded operator always exists.

Example. An important kind of operator is the integral operator. Let $\mathcal{H} = L_C^2(\mathbb{R}^q, \pi)$ and $\mathcal{E} = L_C^2(\mathbb{R}^r, \rho)$ where π and ρ are pdf. The integral operator $K : \mathcal{H} \to \mathcal{E}$ is defined as

$$K\varphi(\tau) = \int k(\tau, s)\, \varphi(s)\, \pi(s)\, ds. \tag{2.5}$$

The function k is called kernel of the operator. If k satisfies

$$\int\int |k(\tau, s)|^2\, \pi(s)\, \rho(\tau)\, ds d\tau < \infty \tag{2.6}$$

(k is said to be a L^2-kernel) then K is a bounded operator and

$$\|K\| \leq \sqrt{\int\int |k(\tau, s)|^2\, \pi(s)\, \rho(\tau)\, ds d\tau}.$$

Indeed for any $\varphi \in \mathcal{H}$, we have

$$\begin{aligned} \|K\varphi\|_{\mathcal{E}}^2 &= \int \left| \int k(\tau, s)\, \varphi(s)\, \pi(s)\, ds \right|^2 \rho(\tau)\, d\tau \\ &= \int |\langle k(\tau, .), \varphi(.) \rangle_{\mathcal{H}}|^2\, \rho(\tau)\, d\tau \\ &\leq \int \|k(\tau, .)\|_{\mathcal{H}}^2\, \|\varphi\|_{\mathcal{H}}^2\, \rho(\tau)\, d\tau \end{aligned}$$

by Cauchy-Schwarz inequality. Hence we have

$$\begin{aligned}\|K\varphi\|_{\mathcal{E}}^2 &\leq \|\varphi\|_{\mathcal{H}}^2 \int \|k(\tau,.)\|_{\mathcal{H}}^2 \rho(\tau)\, d\tau \\ &= \|\varphi\|_{\mathcal{H}}^2 \int\int |k(\tau,s)|^2 \pi(s)\rho(\tau)\, ds d\tau.\end{aligned}$$

The upperbound for $\|K\|$ follows.

The adjoint K^* of the operator K is also an integral operator

$$K^*\psi(s) = \int k^*(s,\tau)\psi(\tau)\rho(\tau)\, d\tau$$

with $k^*(s,\tau) = \overline{k(\tau,s)}$. Indeed, we have

$$\begin{aligned}\langle K\varphi,\psi\rangle_{\mathcal{E}} &= \int (K\varphi)(\tau)\overline{\psi(\tau)}\rho(\tau)\, d\tau \\ &= \int \left(\int k(\tau,s)\varphi(s)\pi(s)\, ds\right)\overline{\psi(\tau)}\rho(\tau)\, d\tau \\ &= \int \varphi(s)\left(\int k(\tau,s)\overline{\psi(\tau)}\rho(\tau)\right)\pi(s)\, ds \\ &= \int \varphi(s)\overline{\left(\int k^*(s,\tau)\psi(\tau)\rho(\tau)\right)}\pi(s)\, ds \\ &= \langle\varphi, K^*\psi\rangle_{\mathcal{H}}.\end{aligned}$$

Definition 2.22 (Self-adjoint). *If $K = K^*$ then K is called self-adjoint (or Hermitian).*

Remark that if K is a self-adjoint integral operator then $k(s,\tau) = \overline{k(\tau,s)}$.

Theorem 2.23. *Let $K : \mathcal{H} \to \mathcal{H}$ be a self-adjoint operator then*

$$\|K\| = \sup_{\|\varphi\|=1} \left|\langle K\varphi,\varphi\rangle_{\mathcal{H}}\right|.$$

Definition 2.24 (Positive operator). *An operator $K : \mathcal{H} \to \mathcal{H}$ is called positive if it is self-adjoint and $\langle K\varphi,\varphi\rangle_{\mathcal{H}} \geq 0$.*

Definition 2.25. *A sequence (K_n) of operators $K_n : \mathcal{H} \to \mathcal{E}$ is called pointwise convergent if for every $\varphi \in \mathcal{H}$, the sequence $K_n\varphi$ converges in ε. A sequence (K_n) of bounded operators converges in norm to a bounded operator K if $\|K_n - K\| \to 0$ as $n \to \infty$.*

Definition 2.26 (Compact operator). *A linear operator $K : \mathcal{H} \to \mathcal{E}$ is called a compact operator if for every bounded sequence (φ_n) in $\mathcal{H}$, the sequence $(K\varphi_n)$ contains a convergent subsequence in $\mathcal{E}$.*

Theorem 2.27. *Compact linear operators are bounded.*

Not every bounded operator is compact. An example is given by the identity operator on an infinite dimensional space $\mathcal{H}$. Consider an orthonormal sequence (e_n) in $\mathcal{H}$. Then the sequence $\mathcal{I}e_n = e_n$ does not contain a convergent subsequence.

Theorem 2.28. *Finite dimensional operators are compact.*

Theorem 2.29. *Let the sequence $K_n : \mathcal{H} \to \mathcal{E}$ of compact linear operators that are norm convergent to a linear operator $K : \mathcal{H} \to \mathcal{E}$, i.e., $\|K_n - K\| \to 0$ as $n \to \infty$ then K is compact. Moreover, every compact operator is the limit of a sequence of operators with finite dimensional range.*

Hilbert Schmidt operators are discussed in Dunford and Schwartz (1988, p. 1009), Dautray and Lyons (1984, Vol 5, p.41, chapter VIII).

Definition 2.30 (Hilbert-Schmidt operator). *Let $\{\varphi_j, j = 1, 2, ...\}$ be a complete orthonormal set in a Hilbert space $\mathcal{H}$. An operator $K : \mathcal{H} \to \mathcal{E}$ is said to be a Hilbert-Schmidt operator if the quantity $\|.\|_{HS}$ defined by*

$$\|K\|_{HS} = \left\{ \sum_{j=1}^{\infty} \left\| K\varphi_j \right\|_{\mathcal{E}}^2 \right\}^{1/2}$$

is finite. The number $\|K\|_{HS}$ is called the Hilbert-Schmidt norm of K. Moreover

$$\|K\| \leq \|K\|_{HS} \tag{2.7}$$

and hence K is bounded.

¿From (2.7), it follows that HS norm convergence implies (operator) norm convergence.

Theorem 2.31. *The Hilbert-Schmidt norm is independent of the orthonormal basis used in its definition.*

Theorem 2.32. *Every Hilbert-Schmidt operator is compact.*

Theorem 2.33. *The adjoint of an Hilbert-Schmidt operator is itself a Hilbert-Schmidt operator and $\|K\|_{HS} = \|K^*\|_{HS}$.*

Theorem 2.32 implies that Hilbert-Schmidt (HS) operators can be approached by a sequence of finite dimensional operators.

Example. Let K be the integral operator defined by (2.5) and (2.6), then K is a Hilbert-Schmidt (HS) operator and its adjoint is also a HS operator. Actually, all Hilbert-Schmidt operators of $L^2(\mathbb{R}^q, \pi)$ in $L^2(\mathbb{R}^r, \rho)$ are integral operators. The following theorem is proved in Dautray and Lions (Vol. 5, p. 45).

Theorem 2.34. *An operator of $L^2(\mathbb{R}^q, \pi)$ in $L^2(\mathbb{R}^r, \rho)$ is Hilbert-Schmidt if and only if it admits a kernel representation (2.5) conformable to (2.6). In this case, the kernel k is unique.*

Example. For illustration, we consider the effect of restricting K on a subset of $L_C^2(\mathbb{R}^q, \pi)$. Consider $\widetilde{K}$ the operator defined by

$$\begin{aligned} \widetilde{K} &: L_C^2(\mathbb{R}^q, \widetilde{\pi}) \to L_C^2(\mathbb{R}^r, \widetilde{\rho}) \\ \widetilde{K}\varphi &= K\varphi \end{aligned}$$

for every $\varphi \in L_C^2(\mathbb{R}^q, \widetilde{\pi})$, where $L_C^2(\mathbb{R}^q, \widetilde{\pi}) \subset L_C^2(\mathbb{R}^q, \pi)$ and $L_C^2(\mathbb{R}^r, \widetilde{\rho}) \supset L_C^2(\mathbb{R}^r, \rho)$. Assume that K is a HS operator defined by (2.5). Under which conditions is $\widetilde{K}$ an HS operator? Let

$$\begin{aligned} \widetilde{K}\varphi(s) &= \int k(\tau, s)\varphi(s)\pi(s)\,ds \\ &= \int k(\tau, s)\frac{\pi(s)}{\widetilde{\pi}(s)}\varphi(s)\widetilde{\pi}(s)\,ds \\ &\equiv \int \widetilde{k}(\tau, s)\varphi(s)\widetilde{\pi}(s)\,ds. \end{aligned}$$

Note that

$$\begin{aligned} & \int \left|\widetilde{k}(\tau, s)\right|^2 \widetilde{\pi}(s)\widetilde{\rho}(\tau)\,ds d\tau \\ = & \int |k(\tau, s)|^2 \frac{\pi(s)}{\widetilde{\pi}(s)}\frac{\widetilde{\rho}(\tau)}{\rho(\tau)}\pi(s)\rho(\tau)\,ds d\tau \\ < & \sup_s \left|\frac{\pi(s)}{\widetilde{\pi}(s)}\right| \sup_\tau \left|\frac{\widetilde{\rho}(\tau)}{\rho(\tau)}\right| \int |k(\tau, s)|^2 \pi(s)\rho(\tau)\,ds d\tau. \end{aligned}$$

Hence the HS property is preserved if (a) there is a constant $c > 0$ such that $\pi(s) \leq c\widetilde{\pi}(s)$ for all $s \in \mathbb{R}^q$ and (b) there is a constant d such that $\widetilde{\rho}(\tau) \leq d\rho(\tau)$ for all $\tau \in \mathbb{R}^r$.

2.3. Spectral decomposition of compact operators

For compact operators, spectral analysis reduces to the analysis of eigenvalues and eigenfunctions. Let $K : \mathcal{H} \to \mathcal{H}$ be a compact linear operator.

Definition 2.35. *λ is an eigenvalue of K if there is a nonzero vector $\phi \in \mathcal{H}$ such that $K\phi = \lambda\phi$. ϕ is called eigenfunction of K corresponding to λ.*

Theorem 2.36. *All eigenvalues of a self-adjoint operator are real and eigenfunctions corresponding to different eigenvalues are orthogonal.*

Theorem 2.37. *All eigenvalues of a positive operator are nonnegative.*

Theorem 2.38. *For every eigenvalue λ of a bounded operator K, we have $|\lambda| \leq \|K\|$.*

Theorem 2.39. *Let K be a self-adjoint compact operator, the set of its eigenvalues (λ_j) is countable and its eigenvectors (ϕ_j) can be orthonormalized. Its largest eigenvalue (in absolute value) satisfies $|\lambda_1| = \|K\|$. If K has infinitely many eigenvalues $|\lambda_1| \geq |\lambda_2| \geq ...$, then $lim_{j\to\infty}\lambda_j = 0$.*

Let $K : \mathcal{H} \to \mathcal{E}$, K^*K and KK^* are self-adjoint positive operators on $\mathcal{H}$ and $\mathcal{E}$ respectively. Hence their eigenvalues are nonnegative by Theorem 2.37.

Definition 2.40. *Let $\mathcal{H}$ and $\mathcal{E}$ be Hilbert spaces, $K : \mathcal{H} \to \mathcal{E}$ be a compact linear operator and $K^* : \mathcal{E} \to \mathcal{H}$ be its adjoint. The square roots of the eigenvalues of the nonnegative self-adjoint compact operator $K^*K : \mathcal{H} \to \mathcal{H}$ are called singular values of K.*

The following results (Kress, 1999, Theorem 15.16) apply to operators that are not necessarily self-adjoint.

Theorem 2.41. *Let (λ_j) denote the sequence of the nonzero singular values of the compact linear operator K repeated according to their multiplicity. Then there exist orthonormal sequences ϕ_j of $\mathcal{H}$ and ψ_j of $\mathcal{E}$ such that*

$$K\phi_j = \lambda_j\psi_j,\ K^*\psi_j = \lambda_j\phi_j \tag{2.8}$$

for all $j \in N$. For each $\varphi \in \mathcal{H}$ we have the singular value decomposition

$$\varphi = \sum_{j=1}^{\infty} \langle \varphi, \phi_j \rangle \phi_j + Q\varphi \tag{2.9}$$

with the orthogonal projection operator $Q : \mathcal{H} \to \mathcal{N}(K)$ and

$$K\varphi = \sum_{j=1}^{\infty} \lambda_j \langle \varphi, \phi_j \rangle \psi_j. \tag{2.10}$$

$\{\lambda_j, \phi_j, \psi_j\}$ is called singular system of K. Note that λ_j^2 are the eigenvalues of KK^* and K^*K associated with the eigenfunctions ψ_j and ϕ_j respectively.

Theorem 2.42. *Let K be the integral operator defined by (2.5) and assume Condition (2.6) holds. Let $\{\lambda_j, \phi_j, \psi_j\}$ be as in (2.8). Then:*

(i) The Hilbert Schmidt norm of K can be written as

$$\|K\|_{HS} = \left\{ \sum_{j\in N} |\lambda_j|^2 \right\}^{1/2} = \left\{ \int\int |k(\tau, s)|^2 \pi(s) \rho(\tau) ds d\tau \right\}^{1/2}$$

where each λ_j is repeated according to its multiplicity.

(ii) (Mercer's formula) $k(\tau, s) = \sum_{j=1}^{\infty} \lambda_j \psi_j(\tau) \overline{\phi_j(s)}$.

Example (degenerate operator). Consider an integral operator defined on $L^2([a,b])$ such that

$$\begin{aligned} Kf(\tau) &= \int_a^b k(\tau,s) f(s)\, ds \\ k(\tau,s) &= \sum_{l=1}^n a_l(\tau) b_l(s). \end{aligned}$$

Assume that a_l and b_l belong to $L^2([a,b])$ for all l. By (2.6), it follows that K is bounded. As moreover K is finite dimensional, we have K compact by Theorem 2.28. Assume that the set of functions (a_l) is linearly independent. Pose $K\phi = \lambda\phi$, we obtain

$$\sum_{l=1}^n a_l(\tau) \int b_l(s)\phi(s)\, ds = \lambda\phi(\tau)$$

hence $\phi(\tau)$ is necessarily of the form $\sum_{l=1}^n c_l a_l(\tau)$. The dimension of the range of K is therefore n, there are at most n nonzero eigenvalues.

Example. Let $\mathcal{H} = L^2([0,1])$ and the integral operator $Kf(\tau) = \int_0^1 (\tau \wedge s) f(s)\, ds$ where $\tau \wedge s = \min(\tau, s)$. It is possible to compute explicitly the eigenvalues and eigenfunctions of K by solving $K\phi = \lambda\phi \iff \int_0^\tau s\phi(s)\, ds + \tau\int_\tau^1 \phi(s)\, ds = \lambda\phi(\tau)$. Using two successive differentiations with respect to τ, we obtain a differential equation $\phi(\tau) = -\lambda\phi''(\tau)$ with boundary conditions $\phi(0) = 0$ and $\phi'(1) = 0$. Hence the set of orthonormal eigenvectors is $\phi_j(\tau) = \sqrt{2}\sin((\pi j\tau)/2)$ associated with the eigenvalues $\lambda_j = 4/(\pi^2 j^2)$, $j = 1,3,5,....$We can see that the eigenvalues converge to zero at an arithmetic rate.

Example. Let π be the pdf of the standard normal distribution and $\mathcal{H} = L^2(\mathbb{R}, \pi)$. Define K be the integral operator with kernel

$$k(\tau,s) = \frac{l(\tau,s)}{\pi(\tau)\pi(s)}$$

where $l(\tau,s)$ is the joint pdf of the bivariate normal $\mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$. Then K is a self-adjoint operator with eigenvalues $\lambda_j = \rho^j$ and eigenfunctions the Hermite polynomials ϕ_j, $j = 1,2,...$ defined in (2.1). This is an example where the eigenvalues decay exponentially fast.

2.4. Random element in Hilbert spaces

2.4.1. Definitions

Let $\mathcal{H}$ be a real separable Hilbert space with norm $\|\|$ induced by the inner product $\langle , \rangle$. Let $(\Omega, \mathcal{F}, P)$ be a complete probability space. Let $X : \Omega \to \mathcal{H}$ be a Hilbert space-valued random element (an $\mathcal{H}$-r.e.). X is integrable or has finite expectation $E(X)$ if $E(\|X\|) = \int_\Omega \|X\|\, dP < \infty$, in that case $E(X)$ satisfies $E(X) \in \mathcal{H}$ and $E[\langle X, \varphi\rangle] = \langle E(X), \varphi\rangle$ for all $\varphi \in \mathcal{H}$. An $\mathcal{H}$-r.e. X is weakly second order if $E\left[\langle X, \varphi\rangle^2\right] < \infty$ for all $\varphi \in \mathcal{H}$. For a

weakly second order $\mathcal{H}$-r.e. X with expectation $E(X)$, we define the covariance operator K as

$$\begin{aligned} K &: \mathcal{H} \to \mathcal{H} \\ K\varphi &= E\left[\langle X - E(X), \varphi\rangle (X - E(X))\right] \end{aligned}$$

for all $\varphi \in \mathcal{H}$. Note that $var\langle X, \varphi\rangle = \langle K\varphi, \varphi\rangle$.

Example. Let $\mathcal{H} = L^2([0,1])$ with $\|g\| = \left[\int_0^1 g(\tau)^2 d\tau\right]^{1/2}$ and $X = h(\tau, Y)$ where Y is a random variable and $h(.,Y) \in L^2([0,1])$ with probability one. Assume $E(h(\tau,Y)) = 0$, then the covariance operator takes the form:

$$\begin{aligned} K\varphi(\tau) &= E\left[\langle h(.,Y), \varphi\rangle h(\tau,Y)\right] \\ &= E\left[\left(\int h(s,Y)\varphi(s)ds\right) h(\tau,Y)\right] \\ &= \int E\left[h(\tau,Y)h(s,Y)\right]\varphi(s)ds \\ &\equiv \int k(\tau,s)\varphi(s)ds. \end{aligned}$$

If moreover, $h(\tau,Y) = I\{Y \leq \tau\} - F(\tau)$ then $k(\tau,s) = F(\tau \wedge s) - F(\tau)F(s)$.

Definition 2.43. *An $\mathcal{H}$-r.e. Y has a Gaussian distribution on $\mathcal{H}$ if for all $\varphi \in \mathcal{H}$ the real-valued r.v. $\langle \varphi, Y\rangle$ has a Gaussian distribution on $\mathbb{R}$.*

Definition 2.44 (strong mixing). *Let $\{X_{i,n}, i = ..., -1, 0, 1, ...; n \geq 1\}$ be an array of $\mathcal{H}$-r.e., defined on the probability space $(\Omega, \mathcal{F}, P)$ and define $\mathcal{A}_{n,a}^{n,b} = \sigma(X_{i,n}, a \leq i \leq b)$ for all $-\infty \leq a \leq b \leq +\infty$, and $n \geq 1$. The array $\{X_{i,n}\}$ is called a strong or $\alpha-$mixing array of $\mathcal{H}$-r.e. if $\lim_{j\to\infty}\alpha(j) = 0$ where*

$$\alpha(j) = \sup_{n\geq 1}\sup_{l}\sup_{A,B}\left[|P(A\cap B) - P(A)P(B)| : A \in \mathcal{A}_{n,-\infty}^{n,l},\ B \in \mathcal{A}_{n,l+j}^{n,+\infty}\right].$$

2.4.2. Central limit theorem for mixing processes

We want to study the asymptotic properties of $Z_n = n^{-1/2}\sum_{i=1}^n X_{i,n}$ where $\{X_{i,n} : 1 \leq 1 \leq n\}$ is an array of $\mathcal{H}$-r.e.. Weak and strong laws of large numbers for near epoch dependent (NED) processes can be found in Chen and White (1996). Here we provide sufficient conditions for the weak convergence of processes to be denoted $\Rightarrow$ (see Davidson, 1994, for a definition). Weak convergence is stronger than the standard central limit theorem (CLT) as illustrated by a simple example. Let (X_i) an iid sequence of zero mean weakly second order elements of $\mathcal{H}$. Then for any Z in $\mathcal{H}$, $\langle X_i, Z\rangle$ is an iid zero mean sequence of $\mathbf{C}$ with finite variance $\langle KZ, Z\rangle$. Then standard CLT implies the asymptotic normality of $\frac{1}{\sqrt{n}}\sum_{i=1}^n \langle X_i, Z\rangle$. The weak convergence of $\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i$ to a Gaussian process $\mathcal{N}(0,K)$ in $\mathcal{H}$ requires an extra assumption, namely $E\|X_1\|^2 < \infty$. Weak convergence theorems

for NED processes that might have trending mean (hence are not covariance stationary) are provided by Chen and White (1998). Here, we report results for mixing processes proved by Politis and Romano (1994). See also van der Vaart and Wellner (1996) for iid sequences.

Theorem 2.45. *Let $\{X_{i,n} : 1 \leq 1 \leq n\}$ be a double array of stationary mixing $\mathcal{H}$-r.e. with zero mean such that, for all n, $\|X_{i,n}\| < B$ with probability one, and $\sum_{j=1}^{m} j^2 \alpha(j) \leq Km^r$ for all $1 \leq m \leq n$ and n, and some $r < 3/2$. Assume, for any integer $l \geq 1$, that $(X_{1,n}, ..., X_{l,n})$, regarded as a r.e. of $\mathcal{H}^l$, converges in distribution to $(X_1, ..., X_l)$, say. Moreover, assume $E[\langle X_{1,n}, X_{l,n}\rangle] \to E[\langle X_1, X_l\rangle]$ as $n \to \infty$ and*

$$\lim_{n\to\infty} \sum_{l=1}^{n} E[\langle X_{1,n}, X_{l,n}\rangle] = \sum_{l=1}^{\infty} E[\langle X_1, X_l\rangle] < \infty.$$

Let $Z_n = n^{-1/2} \sum_{i=1}^{n} X_{i,n}$. For any $\varphi \in \mathcal{H}$, let $\sigma^2_{\varphi,n}$ denote the variance of $\langle Z_n, \varphi\rangle$. Assume

$$\sigma^2_{\varphi,n} \underset{n\to\infty}{\to} \sigma^2_{\varphi} \equiv Var(\langle X_1, \varphi\rangle) + 2\sum_{i=1}^{\infty} cov(\langle X_1, \varphi\rangle, \langle X_{1+i}, \varphi\rangle). \tag{2.11}$$

Then Z_n converges weakly to a Gaussian process $\mathcal{N}(0, K)$ in $\mathcal{H}$, with zero mean and covariance operator K satisfying $\langle K\varphi, \varphi\rangle = \sigma^2_{\varphi}$ for each $\varphi \in \mathcal{H}$.

In the special case when the $X_{i,n} = X_i$ form a stationary sequence, the conditions simplify considerably:

Theorem 2.46. *Assume $X_1, X_2, ...$is a stationary sequence of $\mathcal{H}$-r.e. with mean μ and mixing coefficient α. Let $Z_n = n^{-1/2} \sum_{i=1}^{n} (X_i - \mu)$.*

(i)If $E\left(\|X_1\|^{2+\delta}\right) < \infty$ for some $\delta > 0$, and $\sum_j [\alpha(j)]^{\delta/(2+\delta)} < \infty$

(ii) or if $X_1, X_2, ...$is iid and $E\|X_1\|^2 < \infty$

Then Z_n converges weakly to a Gaussian process $G \sim \mathcal{N}(0, K)$ in $\mathcal{H}$. The distribution of G is determined by the distribution of its marginals $\langle G, \varphi\rangle$ which are $\mathcal{N}\left(0, \sigma^2_{\varphi}\right)$ distributed for every $\varphi \in \mathcal{H}$ where σ^2_{φ} is defined in (2.11).

Let $\{e_l\}$ be a complete orthonormal basis of $\mathcal{H}$ then $\|X_1\|^2 = \sum_{l=1}^{\infty} \langle X_1, e_l\rangle^2$ hence, in the iid case, it suffices to check that $E\|X_1\|^2 = \sum_{l=1}^{\infty} E\left[\langle X_1, e_l\rangle^2\right] < \infty$.

The following theorem is stated in more general terms in Chen and White (1992).

Theorem 2.47. *Let A_n be a random bounded linear operator from $\mathcal{H}$ to $\mathcal{H}$ and $A \neq 0$ be a nonrandom bounded linear operator from $\mathcal{H}$ to $\mathcal{H}$. If $\|A_n - A\| \to 0$ in probability as $n \to \infty$ and $Y_n \Rightarrow Y \sim \mathcal{N}(0, K)$ in $\mathcal{H}$. Then $A_n Y_n \Rightarrow AY \sim \mathcal{N}(0, AKA^*)$.*

In Theorem 2.47, the boundedness of A is crucial. In most of our applications, A will not be bounded and we will not be able to apply Theorem 2.47. Instead we will have to check the Liapunov condition (Davidson 1994) "by hand".

Theorem 2.48. *Let the array* $\{X_{i,n}\}$ *be independent with zero mean and variance sequence* $\left\{\sigma_{i,n}^2\right\}$ *satisfying* $\sum_{t=1}^{n}\sigma_{i,n}^2 = 1$. *Then* $\sum_{i=1}^{n} X_{i,n} \xrightarrow{d} \mathcal{N}(0,1)$ *if*

$$\lim_{n\to\infty}\sum_{i=1}^{n} E\left[\left|X_{i,n}\right|^{2+\delta}\right] = 0 \qquad \text{(Liapunov condition)}$$

for some $\delta > 0$.

2.5. Estimation of an operator and its adjoint

2.5.1. The estimator of the adjoint of an operator and the adjoint of the estimator of an operator

Let $K : \mathcal{H} \to \mathcal{E}$ and $\hat{K}_n$ be an estimator of K. In general the adjoint of $\hat{K}_n$, $\left(\hat{K}_n\right)^*$, differs from the estimator of the adjoint $\widehat{(K^*)}_n$ for the spaces $\mathcal{H}$ and $\mathcal{E}$. That is, we do not have

$$\left\langle \hat{K}_n\varphi, \psi\right\rangle_{\mathcal{E}} = \left\langle \varphi, \widehat{(K^*)}_n\psi\right\rangle_{\mathcal{H}}. \tag{2.12}$$

First we will discuss cases where Equality (2.12) holds. Then, we will turn to cases where it does not hold. In the latter case, we will show that we can define two Hilbert spaces $\mathcal{E}_n$ and $\mathcal{H}_n$ for which

$$\left\langle \hat{K}_n\varphi, \psi\right\rangle_{\mathcal{E}_n} = \left\langle \varphi, \widehat{(K^*)}_n\psi\right\rangle_{\mathcal{H}_n}. \tag{2.13}$$

Assume $\mathcal{H} = L^2(\mathbb{R}^q, \pi)$ and $\mathcal{E} = L^2(\mathbb{R}^r, \rho)$ where π and ρ are two pdf given a priori. Define

$$K\varphi(w) = \int k(w,z)\,\varphi(z)\,\pi(z)\,dz \tag{2.14}$$

for some function $k(w,z)$. This definition of K is quite general as we do not impose any continuity of $k(w,z)$. We have

$$\begin{aligned}\langle K\varphi, \psi\rangle_{\mathcal{E}} &= \int\int k(w,z)\,\varphi(z)\,\pi(z)\,dz\psi(w)\,\rho(w)\,dw \\ &= \int \varphi(z)\left(\int k(w,z)\,\psi(w)\,\rho(w)\,dw\right)\pi(z)\,dz.\end{aligned}$$

Hence the kernel of K^* is $k^*(z,w) = k(w,z)$.

$$K^*\psi(z) = \int k(w,z)\,\psi(w)\,\rho(w)\,dw. \tag{2.15}$$

Assume that K and K^* are estimated $\hat{K}_n$ and $\widehat{(K^*)}_n$ obtained by replacing k by k_n in (2.14) and (2.15) respectively. Then it is easy to check that $\widehat{(K^*)}_n = \left(\hat{K}_n\right)^*$ for the spaces of reference $\mathcal{H}$ and $\mathcal{E}$.

There are cases where it is natural to define the spaces of reference as a function of unknown pdfs. This typically happens when K is a conditional expectation operator. Let $(Z, W) \in \mathbb{R}^q \times \mathbb{R}^r$ be a r.v. with distribution $F_{Z,W}$, let F_Z, and F_W be the marginal distributions of Z and W respectively. The corresponding pdfs are denoted $f_{Z,W}$, f_Z, and f_W. Define

$$\begin{aligned} \mathcal{H} &= L^2\left(\mathbb{R}^q, f_Z\right) \equiv L^2_Z, \\ \mathcal{E} &= L^2\left(\mathbb{R}^r, f_W\right) \equiv L^2_W. \end{aligned}$$

Let K be the conditional expectation operator:

$$\begin{aligned} K &: L^2_Z \to L^2_W \\ \varphi &\to E\left[\varphi(Z)|W\right]. \end{aligned} \tag{2.16}$$

Its adjoint is also a conditional expectation operator:

$$\begin{aligned} K^* &: L^2_W \to L^2_Z \\ \psi &\to E\left[\psi(W)|Z\right]. \end{aligned}$$

Indeed, we have

$$\langle \varphi, K^*\psi \rangle_Z = \langle K\varphi, \psi \rangle_W.$$

Using the notation of (2.15), K is an integral operator with kernel

$$k(w,z) = \frac{f_{Z,W}(z,w)}{f_Z(z) f_W(w)}$$

and $\pi = f_Z$. K^* has for kernel $k^*(z,w) = k(w,z)$ but is defined in a different space ($\rho = f_W$). By Theorem 2.34, a sufficient condition for K and K^* be compact is

$$\int\int \left[\frac{f_{Z,W}(z,w)}{f_Z(z) f_W(w)}\right]^2 f_Z(z) f_W(w)\, dzdw < \infty.$$

Let $\hat{f}_{Z,W}$, $\hat{f}_Z(z)$, and $\hat{f}_W(w)$ be nonparametric estimators of $f_{Z,W}$, $f_Z(z)$, and $f_W(w)$ obtained either by kernel or sieves estimators. Assume that K and K^* are estimated by replacing the unknown pdfs by their estimators, that is:

$$\begin{aligned} \hat{K}_n\varphi(w) &= \int \frac{\hat{f}_{Z,W}(z,w)}{\hat{f}_Z(z)}\varphi(z)\, dz, \\ \widehat{(K^*)}_n\psi(z) &= \int \frac{\hat{f}_{Z,W}(z,w)}{\hat{f}_W(w)}\psi(w)\, dw. \end{aligned}$$

Then we have $\widehat{(K^*)}_n \neq \left(\hat{K}_n\right)^*$ for $\mathcal{H} = L^2_Z$ and $\mathcal{E} = L^2_W$. However, we have $\widehat{(K^*)}_n = \left(\hat{K}_n\right)^*$ for $\mathcal{H}_n$ and $\mathcal{E}_n$ defined by $\mathcal{H}_n = L^2\left(\mathbb{R}^q, \hat{f}_Z\right)$ and $\mathcal{E}_n = L^2\left(\mathbb{R}^r, \hat{f}_W\right)$.

It is important for our estimation procedure to make sense that (2.13) holds. To achieve this, we need also to estimate the spaces $\mathcal{H}$, $\mathcal{E}$, that is, to replace the inner products on these spaces by estimated inner products. The new spaces $\mathcal{H}_n$ and $\mathcal{E}_n$ depend on the sample size and on the estimation procedure. Another approach consists in defining $\mathcal{H} = L^2(\mathbb{R}^q, \pi)$ and $\mathcal{E} = L^2(\mathbb{R}^r, \rho)$ where π and ρ are known and satisfy: There exist $c, c' > 0$ such that $f_Z(z) \leq c\pi(z)$ and $f_W(w) \geq c'\rho(w)$. Then

$$\begin{aligned} K^*\psi(z) &= \int \frac{f_{Z,W}(z,w)}{f_W(w)} \frac{\rho(w)}{\pi(z)} \psi(w)\, dw \\ &\neq E[\psi(W)|Z = z]. \end{aligned}$$

In that case, $\widehat{(K^*)}_n = \left(\hat{K}_n\right)^*$ for $\mathcal{H}$ and $\mathcal{E}$ but the choice of π and ρ require some knowledge on the support and the tails of the distributions of W and Z.

An alternative solution to estimating K and K^* is to estimate the spectrum of K and to apply Mercer's formula. Let $\mathcal{H} = L^2_Z$ and $\mathcal{E} = L^2_W$. The singular system $\{\lambda_j, \phi_j, \psi_j\}$ of K satisfies

$$\lambda_j = \sup_{\phi_j, \psi_j} E\left[\phi_j(Z)\psi_j(W)\right], \; j = 1, 2... \tag{2.17}$$

subject to $\|\phi_j\|_{\mathcal{H}} = 1, \langle \phi_j, \phi_l \rangle_{\mathcal{H}} = 0$, $l = 1, 2, ..., j-1$, $\|\psi_j\|_{\mathcal{E}} = 1, \langle \psi_j, \psi_l \rangle_{\mathcal{E}} = 0$, $l = 1, 2, ..., j-1$. Assume the econometrician observes a sample $\{w_i, z_i : i = 1, ..., n\}$. To estimate $\{\lambda_j, \phi_j, \psi_j\}$, one can either estimate (2.17) by replacing the expectation by the sample mean or by replacing the joint pdf by a nonparametric estimator.

The first approach was adopted by Darolles, Florens, and Renault (1998). Let

$$\begin{aligned} \mathcal{H}_n &= \left\{\varphi : \mathbb{R}^q \to \mathbb{R}, \int \varphi(z)^2 d\widehat{F}_Z(z) < \infty\right\}, \\ \mathcal{E}_n &= \left\{\psi : \mathbb{R}^r \to \mathbb{R}, \int \psi(w)^2 d\widehat{F}_W(w) < \infty\right\} \end{aligned}$$

where $\widehat{F}_Z$ and $\widehat{F}_W$ are the empirical distributions of Z and W that is $\|\varphi\|^2_{\mathcal{H}_n} = \frac{1}{n}\sum_{i=1}^n \varphi(z_i)^2$ and $\|\psi\|^2_{\mathcal{E}_n} = \frac{1}{n}\sum_{i=1}^n \psi(w_i)^2$. Darolles, Florens, and Renault (1998) propose to estimate $\{\lambda_j, \phi_j, \psi_j\}$ by solving

$$\hat{\lambda}_j = \sup_{\hat{\phi}_j, \hat{\psi}_j} \frac{1}{n}\sum_{i=1}^n \left[\hat{\phi}_j(z_i)\hat{\psi}_j(w_i)\right], \; j = 1, 2... \tag{2.18}$$

subject to $\left\|\hat{\phi}_j\right\|_{\mathcal{H}_n} = 1, \left\langle \hat{\phi}_j, \hat{\phi}_l \right\rangle_{\mathcal{H}_n} = 0$, $l = 1, 2, ..., j-1$, $\left\|\hat{\psi}_j\right\|_{\mathcal{E}_n} = 1, \left\langle \hat{\psi}_j, \hat{\psi}_l \right\rangle_{\mathcal{E}_n} = 0$,

$l = 1, 2, ..., j-1$ where $\hat{\phi}_j$ and $\hat{\psi}_j$ are elements of increasing dimensional spaces

$$\begin{aligned} \hat{\phi}_j(z) &= \sum_{j=1}^{J} \alpha_j a_j(z), \\ \hat{\psi}_j(w) &= \sum_{j=1}^{J} \beta_j b_j(w) \end{aligned}$$

for some basis $\{a_j\}$ and $\{b_j\}$. By Mercer's formula (2.10), K can be estimated by

$$\begin{aligned} \hat{K}_n\varphi(w) &= \sum \hat{\lambda}_j \left(\int \hat{\phi}_j(z) \varphi(z) d\widehat{F}_Z \right) \hat{\psi}_j(w) \\ \widehat{(K^*)}_n \psi(z) &= \sum \hat{\lambda}_j \left(\int \hat{\psi}_j(w) \psi(w) d\widehat{F}_W \right) \hat{\phi}_j(z). \end{aligned}$$

Hence $\widehat{(K^*)}_n = \left(\hat{K}_n\right)^*$ for $\mathcal{H}_n$ and $\mathcal{E}_n$.

The second approach consists in replacing $f_{Z,W}$ by a nonparametric estimator $\hat{f}_{Z,W}$. Darolles, Florens, and Gourieroux (2000) use a kernel estimator, whereas Chen, Hansen and Scheinkman (1998) use B-spline wavelets. Let $\mathcal{H}_n = L^2\left(\mathbb{R}^q, \hat{f}_Z\right)$ and $\mathcal{E}_n = L^2\left(\mathbb{R}^r, \hat{f}_W\right)$ where $\hat{f}_Z$ and $\hat{f}_W$ are the marginals of $\hat{f}_{Z,W}$. (2.17) can be replaced

$$\hat{\lambda}_j = \sup_{\phi_j, \psi_j} \int \phi_j(z) \psi_j(w) \hat{f}_{Z,W}(z,w) dz dw, \; j = 1, 2... \tag{2.19}$$

subject to $\|\phi_j\|_{\mathcal{H}_n} = 1, \langle \phi_j, \phi_l \rangle_{\mathcal{H}_n} = 0$, $l = 1, 2, ..., j-1$, $\|\psi_j\|_{\mathcal{E}_n} = 1, \langle \psi_j, \psi_l \rangle_{\mathcal{E}_n} = 0$, $l = 1, 2, ..., j-1$. Denote $\left\{\hat{\lambda}_j, \hat{\phi}_j, \hat{\psi}_j\right\}$ the resulting estimators of $\{\lambda_j, \phi_j, \psi_j\}$. By Mercer's formula, K can be approached by

$$\begin{aligned} \hat{K}_n\varphi(w) &= \sum \hat{\lambda}_j \left(\int \hat{\phi}_j(z) \varphi(z) \hat{f}_Z(z) dz \right) \hat{\psi}_j(w) \\ \widehat{(K^*)}_n \psi(z) &= \sum \hat{\lambda}_j \left(\int \hat{\psi}_j(w) \psi(w) \hat{f}_W(w) dw \right) \hat{\phi}_j(z). \end{aligned}$$

Hence $\widehat{(K^*)}_n = \left(\hat{K}_n\right)^*$ for $\mathcal{H}_n$ and $\mathcal{E}_n$. Note that in the three articles mentioned above, $Z = X_{t+1}$ and $W = X_t$ where $\{X_t\}$ is a Markov process. These papers are mainly concerned with estimation. When the data are the discrete observations of a diffusion process, the nonparametric estimations of a single eigenvalue-eigenfunction pair and of the marginal distribution are enough to recover a nonparametric estimate of the diffusion coefficient. The techniques described here can also be used for testing the reversibility of the process $\{X_t\}$, see Darolles, Florens, and Gourieroux (2000).

2.5.2. Computation of eigenvalues and eigenfunctions of finite dimensional operators

We take a different perspective from before. Here, we assume that we have some estimators of K and K^*, denoted $\hat{K}_n$ and $\hat{K}_n^*$ such that $\widehat{(K^*)}_n = \left(\hat{K}_n\right)^* \equiv \hat{K}_n^*$. The aim is to calculate the singular values of $\hat{K}_n$. Assume that $\hat{K}_n$ and $\hat{K}_n^*$ have finite range and satisfy

$$\hat{K}_n\varphi = \sum_{l=1}^{L_n} a_l(\varphi)\,\varepsilon_l \tag{2.20}$$

$$\hat{K}_n^*\psi = \sum_{l=1}^{L_n} b_l(\psi)\,\eta_l \tag{2.21}$$

where $\varepsilon_l \in \mathcal{E}$, $\eta_l \in \mathcal{H}$, $a_l(\varphi)$ is linear in φ and $b_l(\psi)$ is linear in ψ. Moreover the $\{\varepsilon_l\}$ and $\{\eta_l\}$ are assumed to be linearly independent. We have

$$\begin{aligned}\hat{K}_n^*\hat{K}_n\varphi &= \sum_{l=1}^{L_n} b_l\left(\sum_{l'=1}^{L_n} a_{l'}(\varphi)\,\varepsilon_{l'}\right)\eta_l \\ &= \sum_{l,l'=1}^{L_n} a_{l'}(\varphi)\,b_l(\varepsilon_{l'})\,\eta_l. \end{aligned}\tag{2.22}$$

Examples:

1 - Covariance operator

$$K\varphi(\tau_1) = \int E\left[h(\tau_1, X)\,h(\tau_2, X)\right]\varphi(\tau_2)\,d\tau_2$$

$$\begin{aligned}\hat{K}_n\varphi(\tau_1) &= \int\left(\frac{1}{n}\sum_{i=1}^{n} h(\tau_1, x_i)\,h(\tau_2, x_i)\right)\varphi(\tau_2)\,d\tau_2 \\ &= \sum_{i=1}^{n} a_i(\varphi)\,\varepsilon_i\end{aligned}$$

with

$$a_i(\varphi) = \frac{1}{n}\int h(\tau_2, x_i)\,\varphi(\tau_2)\,d\tau_2 \text{ and } \varepsilon_i = h(\tau_1, x_i).$$

Note that in this case, the rate of convergence of $\hat{K}_n^*\hat{K}_n$ is parametric: $\left\|\hat{K}_n^*\hat{K}_n - K^*K\right\| = O\left(1/\sqrt{n}\right)$.

2 - Conditional expectation

$$K\varphi(w) = E\left[\varphi(Z)\,|W = w\right].$$

The kernel estimator with kernel ω and bandwidth c_n is given by

$$\begin{aligned}\hat{K}_n\varphi(w) &= \frac{\sum_{i=1}^n \varphi(z_i)\,\omega\left(\frac{w-w_i}{c_n}\right)}{\sum_{i=1}^n \omega\left(\frac{w-w_i}{c_n}\right)} \\ &= \sum_{i=1}^n a_i(\varphi)\,\varepsilon_i\end{aligned}$$

where

$$a_i(\varphi) = \varphi(z_i) \text{ and } \varepsilon_i = \left[\frac{\omega\left(\frac{w-w_i}{c_n}\right)}{\sum_{i=1}^n \omega\left(\frac{w-w_i}{c_n}\right)}\right].$$

In this case, the rate of convergence of $\hat{K}_n^*\hat{K}_n$ is nonparametric. See Subsection 4.1.

Now we calculate the eigenvalues and eigenfunctions of $\hat{K}_n^*\hat{K}_n$ by solving

$$\hat{K}_n^*\hat{K}_n\phi = \lambda^2\phi.$$

Hence ϕ is necessarily of the form: $\phi = \sum_l \beta_l\eta_l$. Replacing in (2.22), we have

$$\lambda^2\beta_l = \sum_{l',j=1}^{L_n} \beta_j a_{l'}\left(\eta_j\right) b_l\left(\varepsilon_{l'}\right). \tag{2.23}$$

Denote $\underline{\hat{\beta}} = \left[\beta_1, ..., \beta_{L_n}\right]$ the solution of (2.23). Solving (2.23) is equivalent to finding the L_n nonzero eigenvalues $\hat{\lambda}_1^2, ..., \hat{\lambda}_{L_n}^2$ and eigenvectors $\underline{\hat{\beta}}^{\,1}, ..., \underline{\hat{\beta}}^{L_n}$ of an $L_n \times L_n-$matrix C with principle element

$$c_{l,j} = \sum_{l'=1}^{L_n} a_{l'}\left(\eta_j\right) b_l\left(\varepsilon_{l'}\right).$$

The eigenfunctions of $\hat{K}_n^*\hat{K}_n$ are

$$\hat{\phi}_j = \sum_{l=1}^{L_n} \hat{\beta}_l^j \eta_{l,}\ j = 1, ...L_n$$

associated with $\hat{\lambda}_1^2, ..., \hat{\lambda}_{L_n}^2$. $\left\{\hat{\phi}_j : j = 1, .., L_n\right\}$ need to be orthonormalized. The estimators of the singular values are $\hat{\lambda}_j = \sqrt{\hat{\lambda}_j^2}$.

3. Regularized solutions of integral equations of the first kind

This section discusses the property of the integral equations (also called Fredholm equations) of the first kind that is $K\varphi = r$ where K is an integral compact operator. Solving in φ such an equation is ill-posed because (a) the solution might not exist, (b) when it exists, it may not be unique, (c) the solution is not continuous in r. For these reasons, a regularized solution (that will be continuous in r) needs to be implemented.

3.1. Ill-posed problems

Definition 3.1. *Let $K : \mathcal{H} \to \mathcal{E}$ be an operator from a Hilbert space $\mathcal{H}$ into a Hilbert space $\mathcal{E}$. The equation*

$$K\varphi = r,\ \varphi \in \mathcal{H} \tag{3.1}$$

is said to be well-posed if
(i) $\mathcal{N}(K) = \{0\}$ (uniqueness), $r \in K(\mathcal{H})$ (existence)
(ii) $K^{-1} : K(\mathcal{H}) \to \mathcal{H}$ is continuous.

The condition $\mathcal{N}(K) = \{0\}$ is necessary and sufficient to guarantee that, for $r \in K(\mathcal{U})$, the equation $K\varphi = r$ has a unique solution φ. In the case of compact operators, this identification condition is characterized by the positivity of all the singular values.

Proposition 3.2. *(Criterion for identification) Let $K : \mathcal{H} \to \mathcal{E}$ be a compact operator. It is injective, that is $\mathcal{N}(K) = \{0\}$, if and only if all its singular values are nonzero.*

Proof. (i) $\mathcal{N}(K^*K) = \{0\} \Rightarrow \mathcal{N}(K) = \{0\}$ because $\mathcal{N}(K) \subset \mathcal{N}(K^*K)$. (ii) Now assume $\mathcal{N}(K) = \{0\}$. If one singular value λ_j were zero, ϕ_j would belong to $\mathcal{N}(K)$ because $K\phi_j = 0\psi_j = 0$. ■

The singular value decompositions (2.8) to (2.10) give some insight about the solvability issue. A necessary condition to get solvability in $\mathcal{H}$ (existence of the solution) that is $r \in \mathcal{R}(K) = K(\mathcal{H})$, is

$$\sum_{j=1}^{\infty} \frac{\langle r, \psi_j \rangle^2}{\lambda_j^2} < \infty \tag{3.2}$$

because

$$\begin{aligned} \sum_{j=1}^{\infty} \frac{\langle r, \psi_j \rangle^2}{\lambda_j^2} &= \sum_{j=1}^{\infty} \frac{\langle K\varphi, \psi_j \rangle^2}{\lambda_j^2} \\ &= \sum_{j=1}^{\infty} \langle \varphi, \phi_j \rangle^2 \\ &\leq \|\varphi\|^2 . \end{aligned}$$

More generally, we define:

Definition 3.3. *Let* $\{\lambda_j, \phi_j, \psi_j\}$ *be the singular system of the compact operator* K*. For all* $\beta \geq 0$*, we denote*

$$\Phi_\beta = \left\{ \varphi \in \mathcal{H} \text{ such that } \sum_{j=1}^{\infty} \frac{\langle \varphi, \phi_j \rangle^2}{\lambda_j^{2\beta}} < \infty \right\} \tag{3.3}$$

and symmetrically:

$$\Psi_\beta = \left\{ \psi \in \mathcal{E} \text{ such that } \sum_{j=1}^{\infty} \frac{\langle \psi, \psi_j \rangle^2}{\lambda_j^{2\beta}} < \infty \right\}. \tag{3.4}$$

Φ_β *and* Ψ_β *are called* $\beta-$*regularity spaces of the operators* K *and* K^* *respectively.*

Then $\beta \leq \beta' \Rightarrow \Psi_\beta \supset \Psi_{\beta'}$ and $\Phi_\beta \supset \Phi_{\beta'}$ and a necessary condition for solvability is $r \in \Psi_1$. Note however that the maintained assumption of injectivity of K implies that all the spaces Φ_β are dense in $\mathcal{H}$. To see this, let us denote by $\tilde{\Phi}$ the vectorial space spanned by $\{\phi_j\}$ and $\overline{\tilde{\Phi}}$ its closure. Since (i) $\tilde{\Phi} \subset \Phi_\beta \subset \mathcal{H}$ for all β, (ii) $\overline{\tilde{\Phi}} = \mathcal{H}$ (by $\mathcal{N}(K) = \{0\}$), we can conclude that $\overline{\Phi}_\beta = \mathcal{H}$.

Proposition 3.4. $\Phi_1 = \mathcal{R}(K^*)$

Proof. (i) $\Phi_1 \supset \mathcal{R}(K^*)$ since

$$\sum_{j=1}^{\infty} \frac{\langle K^*\psi, \phi_j \rangle^2}{\lambda_j^2} = \sum_{j=1}^{\infty} \frac{\langle \psi, K\phi_j \rangle^2}{\lambda_j^2} = \sum_{j=1}^{\infty} \langle \psi, \phi_j \rangle^2 = \|\psi\|^2 < \infty.$$

(ii) We want to show that $\Phi_1 \subset \mathcal{R}(K^*)$. Let $\varphi \in \Phi_1$. Let $\psi = \sum_{j=1}^{\infty} \frac{\langle \varphi, \phi_j \rangle}{\lambda_j} \psi_j$. Then $\psi \in \mathcal{E}$ and $K^*\psi = \varphi$ by (2.9) and $\mathcal{N}(K) = \{0\}$. Hence $\varphi \in \mathcal{R}(K^*)$. ■

More generally it can be shown that

$$\Phi_\beta = \mathcal{R}\left[(K^*K)^{\frac{\beta}{2}}\right]$$

(See Loubes and Vanhems (2003)).

The relation between the Φ_β spaces and the reproducing kernel Hilbert space may be deduced from this property and is studied in Section 6.Loubes and Vanhems (2003) also characterize the relation between Φ_β spaces and smoothness properties.

Consider now $r \in \Psi_\beta$. If $\beta \geq 1$, the function $\varphi = \sum_{j=1}^{\infty} \frac{\langle r, \psi_j \rangle}{\lambda_j} \phi_j$ is a well-defined element of $\Phi_{\beta-1}$ and

$$K\varphi = \sum_{j=1}^{\infty} \langle r, \psi_j \rangle \psi_j$$

coincides with r if and only if $r \in (\mathcal{N}(K^*))^{\perp}$. Indeed for all $\psi \in \mathcal{E}$, we can write by analogy with (2.9) :

$$\psi = \sum_{j=1}^{\infty} \langle \psi, \psi_j \rangle \psi_j + R\psi \tag{3.5}$$

where R is the orthogonal projector $\mathcal{E} \to \mathcal{N}(K^*)$. Note also that because $(\mathcal{N}(K^*))^{\perp}$ is the closure of $\mathcal{R}(K)$, the condition $r \in (\mathcal{N}(K^*))^{\perp}$ is necessary for solvability. We have then shown:

Proposition 3.5 (Criterion for solvability). *Let $K : \mathcal{H} \to \mathcal{E}$ be a compact operator with singular system $\{\lambda_j, \phi_j, \psi_j\}$. The equation*

$$K\varphi = r,\ \varphi \in \mathcal{H} \tag{3.6}$$

is solvable if and only if r belongs to $(\mathcal{N}(K^*))^{\perp} \cap \Psi_\beta$ for some $\beta \geq 1$. In this case, a solution is given by

$$\varphi = \sum_{j=1}^{\infty} \frac{\langle r, \psi_j \rangle}{\lambda_j} \phi_j \tag{3.7}$$

and belongs to $\Phi_{\beta-1}$.

In the particular case $\beta = 1$, Proposition 3.5 is known as Picard's theorem.

Proposition 3.5 clearly demonstrates the ill-posed nature of the equation $K\varphi = r$. If we perturb the right-hand side r by $r^\delta = r + \delta\psi_j$, we obtain the solution $\varphi^\delta = \varphi + \delta\phi_j/\lambda_j$. Hence, the ratio $\|\varphi^\delta - \varphi\| / \|r^\delta - r\| = 1/\lambda_j$ can be made arbitrary large due to the fact that the singular values tend to zero. Since the influence of measurement errors in r is controlled by the rate of this convergence, Kress (1999, p. 280) says that the equation is "mildly ill-posed" if the singular values decay slowly to zero and that it is "severely ill-posed" if they decay rapidly. Actually, the critical property is the relative decay rate of the sequence $\langle r, \psi_j \rangle$ with respect to the decay of the sequence λ_j. To see this, note that the solution φ has to be determined from its Fourier coefficients by solving the equations

$$\lambda_j \langle \varphi, \phi_j \rangle = \langle r, \psi_j \rangle, \text{ for all } j.$$

Then, we may expect high instability of the solution φ if λ_j goes to zero faster than $\langle \varphi, \phi_j \rangle$, that is if $\varphi \notin \Phi_1$ (or equivalently $r \notin \Psi_2$). More generally, the higher the coefficient β such that $r \in \Psi_\beta$, the better the estimation. Note that it is not hopeless that $r \in \Psi_2$ as $\mathcal{R}(KK^*) \subset \Psi_2$ (since $\mathcal{R}(K^*) \subset \Phi_1$ and $\mathcal{R}(K) \subset \Psi_1$). Indeed, we have

$$\begin{aligned} \varphi &\in \mathcal{R}(K^*) \Longleftrightarrow r = K\varphi \in \mathcal{R}(KK^*) \\ &\Rightarrow r \in \Psi_2. \end{aligned}$$

The condition $r \in \Psi_2$ (and $\varphi \in \Phi_1$) is fulfilled as soon as $\varphi \in \mathcal{R}(K^*)$.

Generally speaking, rather than trying to control the estimation error $\left\|\varphi^{\delta}-\varphi\right\|$ resulting from a measurement error $\left\|r^{\delta}-r\right\|$ by restricting a priori the set of solutions of interest, we will focus on regularization schemes to control the effect of $1/\lambda_j$ in the solution formula (3.7). In this framework, the set of solutions of interest will be tightly related to the decreasing family of subsets Φ_β, $\beta \geq 0$ that characterizes the relevant rate of instability. Moreover as we will consider measurement errors originating not only from r but also from the operator K itself, we must reinforce the common stability properties of regularization schemes. This is the reason why we will refer to these schemes as second order regularization.

3.2. Second order regularization schemes

According to Kress (1999, Theorem 15.21), a regularized solution for Equation (3.1) with K injective compact operator ($\mathcal{N}(K)=\{0\}$) is a family of operators $R_\alpha : \mathcal{E} \to \mathcal{H}$, $\alpha > 0$, defined by

$$R_\alpha r = \sum_{j=1}^{\infty} \frac{1}{\lambda_j} q(\alpha, \lambda_j) \left\langle r, \psi_j \right\rangle \phi_j \tag{3.8}$$

where q is a real function defined on $\mathbb{R}_*^+ \times (0, \|K\|)$ such that there exists $c(\alpha) > 0$, which satisfies for all $\lambda \in (0, \|K\|)$:

$$\begin{aligned} |q(\alpha,\lambda)| &\leq c(\alpha)\lambda, \text{ and} && (3.9)\\ \lim_{\alpha\to 0} q(\alpha,\lambda) &= 1. && (3.10) \end{aligned}$$

Condition (3.9) insures that the operator R_α is bounded and satisfies $\|R_\alpha\| \leq c(\alpha)$. Let $\varphi_\alpha = R_\alpha r$ and φ be defined in (3.7), the regularization bias $\varphi - \varphi_\alpha$ is

$$\varphi - \varphi_\alpha = \sum_{j=1}^{\infty} [1 - q(\alpha, \lambda_j)] \left\langle \varphi, \phi_j \right\rangle \phi_j \tag{3.11}$$

Condition (3.10) implies that the regularization bias vanishes asymptotically. R_α has the advantage over K^{-1} to be bounded because $q(\alpha, \lambda_j)$ converges to zero at least as fast as λ_j. We will reinforce the stability condition by assuming that it goes to zero even faster.

Definition 3.6. *The family of operators R_α defined by (3.8) is a second order regularization if q is a real function defined on $\mathbb{R}_*^+ \times (0, \|K\|)$ such that there exists $d(\alpha) > 0$ which satisfies for all $\lambda \in (0, \|K\|)$:*

$$\begin{aligned} |q(\alpha,\lambda)| &\leq d(\alpha)\lambda^2 && (3.12)\\ \lim_{\alpha\to 0} q(\alpha,\lambda) &= 1. \end{aligned}$$

In the following, we will always normalize the exponent of the regularization parameter α such that $\alpha d(\alpha)$ has a positive finite limit when α goes to zero. As shown in Subsection 3.5 below, second order regularization will be well-suited to deal with the case where a measurement error affects K also (and not just r). It allows to define a bounded operator A_α such that

$$R_\alpha = A_\alpha K^*. \tag{3.13}$$

Note that (3.13) leaves unconstrained the values of A_α on the space $\mathcal{R}(K^*)^\perp = \mathcal{N}(K)$. However, since following Kress (1999, Theorem 15.21), we maintain in this subsection the identification assumption $\mathcal{N}(K) = \{0\}$, A_α is uniquely defined as

$$A_\alpha \varphi = \sum_{j=1}^{\infty} \frac{1}{\lambda_j^2} q(\alpha, \lambda_j) \langle \varphi, \varphi_j \rangle \varphi_j \tag{3.14}$$

for all $\varphi \in \mathcal{H}$. Note that as q is real, A_α is self-adjoint. Then by (3.12), A_α is a bounded operator from $\mathcal{H}$ into $\mathcal{H}$ with

$$\|A_\alpha\| \leq d(\alpha). \tag{3.15}$$

It is one-to-one in the particular case where the regularization weights $q(\alpha, \lambda_j)$ are nonzero for all $j \geq 1$.

It is worthwhile to notice that our notion of second order regularization is conformable to another approach of regularization in terms of penalization. The idea (see e.g. Vapnik (1998)) is to find a solution for (3.6) as an element φ minimizing a certain functional:

$$R(\varphi) = D^2(r, K\varphi) + \alpha W(\varphi) \tag{3.16}$$

where D is some metric in the space $\mathcal{E}$ and W is a functional defined on $\mathcal{H}$ such that the sets

$$\xi_c = \{\varphi : W(\varphi) \leq c\}, \ c \geq 0,$$

are all compact. Typically, the role of α is to penalize large values of φ in order to enforce the stability of the solution. With $\alpha = 0$, the problem of minimization of (3.16) would be equivalent to the solution of equation (3.6) and therefore would also be ill-posed.

Then, when using an euclidean metric D, the minimizer φ of (3.16) will be of the form (3.13). This is also closely related to the concept of generalized inverse as

$$\lim_{\alpha \to 0} A_\alpha = (K^* K)^{-1}$$

and

$$\lim_{\alpha \to 0} R_\alpha = (K^* K)^{-1} K^*$$

is the Moore Penrose generalized inverse. However, $(K^*K)^{-1}$ is also nonstable since its eigenvalues $1/\lambda_j^2$ diverge. Therefore the notion of generalized inverse, or equivalently the problem (3.13) without penalization, cannot be used directly for estimation purpose. It will be better suited for the application of reproducing kernels, see Section 6.

Below, we show that the most common regularization schemes fulfill not only the conditions (3.9), and (3.10), but also the second order requirement (3.12). We characterize these schemes through the definitions of the weights $q(\alpha, \lambda_j)$ in the regularization (3.8). The first regularization method was proposed by A.N. Tikhonov in 1963 and corresponds to the minimization of (3.16) with canonical norms.

Example (Tikhonov regularization).

$$q(\alpha, \lambda) = \frac{\lambda^2}{\alpha + \lambda^2}.$$

Then:

$$|q(\alpha, \lambda)| = \frac{\lambda^2}{\alpha + \lambda^2} \leq \frac{\lambda^2}{\alpha}$$

and $d(\alpha) = 1/\alpha$.

Example (Landweber-Fridman).

$$q(\alpha, \lambda) = 1 - \left(1 - c\lambda^2\right)^{1/\alpha}$$

for some c chosen in the interval $\left(0, 1/\|K\|^2\right)$. It is easy to check that for $\alpha \leq 1$:

$$0 \leq q(\alpha, \lambda) \leq \frac{c\lambda^2}{\alpha}$$

that is $d(\alpha) = c/\alpha$.

Example (Spectral cut-off).

$$q(\alpha, \lambda) = I\left\{\lambda \geq \sqrt{\alpha}\right\} = \begin{cases} 1 \text{ if } \lambda \geq \sqrt{\alpha} \\ 0 \text{ otherwise} \end{cases}.$$

The advantage of spectral cut-off over Tikhonov regularization is that there is no bias for the largest eigenvalues. However, its bias is larger than that obtained with Tikhonov for the smallest eigenvalues. This regularization is not second order. Therefore, we introduce a new regularization scheme closed to spectral cut-off and that is second order.

Example (Extended Spectral cut-off).

$$q(\alpha, \lambda) = I\left\{\lambda \geq \sqrt{\alpha}\right\} + \frac{\lambda^2}{\alpha} I\left\{\lambda < \sqrt{\alpha}\right\} = \begin{cases} \frac{\lambda^2}{\alpha} \text{ if } \lambda < \sqrt{\alpha} \\ 1 \text{ otherwise} \end{cases}.$$

This new regularization scheme maintain the advantage of the spectral cut-off (zero bias for large eigenvalues) while getting a smaller bias for the smallest eigenvalues. Actually this bias, equal to $\left(\alpha - \lambda^2\right)/\alpha$, is similar to Tikhonov bias $\alpha/\left(\alpha + \lambda^2\right)$. Note that

$$0 \leq q(\alpha, \lambda) \leq \frac{\lambda^2}{\alpha}$$

and $d(\alpha) = 1/\alpha$.

Remark that for all the examples but the spectral cut-off, we have

$$0 < q(\alpha, \lambda) \leq 1 \text{ for all } (\alpha, \lambda),$$

which implies that A_α is one-to-one for all α. This is the reason why we will consider in the remaining of the section the extended spectral cut-off rather than the common one. Actually, even though this assumption could be relaxed, it will be better suited to consider only regularizations schemes that satisfies $R_\alpha = A_\alpha K^*$ with A_α bounded and one-to-one.

The practical implementation of these various regularization schemes is discussed below.

3.3. Implementation

Assume $\hat{K}_n$ and $\hat{K}_n^*$ are finite dimensional and denote $\left\{\hat{\lambda}_j, \hat{\phi}_j, \hat{\psi}_j : j = 1, ..., L_n\right\}$ the singular values of $\hat{K}_n$ obtained by the method discussed in Section 2.5.2. We investigate various regularized solutions of the equation:

$$K\varphi = r$$

with $K : \mathcal{H} \to \mathcal{E}$, $\varphi \in \mathcal{H}$, $r \in \mathcal{E}$.

The Tikhonov regularization is based on

$$\begin{aligned} (\alpha_n I + K^* K)\, \varphi_{\alpha_n} &= K^* r \Leftrightarrow \\ \varphi_{\alpha_n} &= \sum_{j=1}^{\infty} \frac{\lambda_j}{\lambda_j^2 + \alpha_n} \langle r, \psi_j \rangle \phi_j \end{aligned}$$

for a penalization term α_n and $\lambda_j = \sqrt{\lambda_j^2}$. Replacing the singular values by their estimates, we obtain

$$\hat{\varphi}_n = \sum_{j=1}^{L_n} \frac{\hat{\lambda}_j}{\hat{\lambda}_j^2 + \alpha_n} \left\langle r, \hat{\psi}_j \right\rangle \hat{\phi}_j$$

When $\hat{K}_n$ and $\hat{K}_n^*$ can be written as in (2.20) and (2.21), one can avoid the estimation of the singular values by solving the equation

$$\begin{aligned} \left(\alpha_n I + \hat{K}_n^* \hat{K}_n\right) \varphi &= \hat{K}_n^* r \Leftrightarrow \\ \alpha_n \varphi + \sum_{l,l'=1}^{L_n} a_{l'}(\varphi)\, b_l(\varepsilon_{l'})\, \eta_l &= \sum_{l=1}^{L_n} b_l(r)\, \eta_l \end{aligned} \tag{3.17}$$

1) First we compute $a_l(\varphi)$:

Apply a_j to (3.17):

$$\alpha_n a_j(\varphi) + \sum_{l,l'=1}^{L_n} a_{l'}(\varphi)\, b_l(\varepsilon_{l'})\, a_j(\eta_l) = \sum_{l=1}^{L_n} b_l(r)\, a_j(\eta_l) \tag{3.18}$$

(3.18) can be rewritten as

$$(\alpha_n I + A)\,\underline{a} = \underline{b}$$

where $\underline{a} = \left[\begin{array}{cccc} a_1(\varphi) & a_2(\varphi) & \cdots & a_{L_n}(\varphi) \end{array}\right]'$, A is the $L_n \times L_n-$matrix with principal element

$$A_{j,l'} = \sum_{l=1}^{L_n} b_l(\varepsilon_{l'})\, a_j(\eta_l)$$

and

$$\underline{b} = \left[\begin{array}{c} \sum_l b_l(r)\, a_1(\eta_l) \\ \vdots \\ \sum_l b_l(r)\, a_{L_n}(\eta_l) \end{array}\right].$$

2) From (3.17), we have

$$\hat{\varphi}_n = \frac{1}{\alpha_n}\left[\sum_{l=1}^{L_n} b_l(r)\,\eta_l - \sum_{l,l'=1}^{L_n} a_{l'}(\varphi)\, b_l(\varepsilon_{l'})\,\eta_l\right].$$

Landweber-Fridman regularization

Let c be a constant so that $0 < c < 1/\|K\|^2$ and α such that $\dfrac{1}{\alpha}$ is integer.

$$\varphi_\alpha = \sum_{j=1}^{\infty} \frac{1}{\lambda_j} q(\alpha, \lambda_j) \left\langle r, \psi_j \right\rangle \phi_j$$

with

$$\begin{aligned} q(\alpha, \lambda_j) &= 1 - \left(1 - c\lambda_j^2\right)^{1/\alpha} \\ &= c\lambda_j^2 \sum_{l=0}^{1/\alpha - 1} \left(1 - c\lambda_j^2\right)^l. \end{aligned}$$

Thus,

$$\begin{aligned}\varphi_\alpha &= c\sum_{l=0}^{1/\alpha-1}\sum_{j=1}^{\infty}\lambda_j\left(1-c\lambda_j^2\right)^l\left\langle r,\psi_j\right\rangle\phi_j \\ &= c\sum_{l=0}^{1/\alpha-1}\sum_{j=1}^{\infty}\lambda_j^2\left(1-c\lambda_j^2\right)^l\left\langle \varphi,\phi_j\right\rangle\phi_j \\ &= c\sum_{l=0}^{1/\alpha-1}\left(I-cK^*K\right)^l K^*K\varphi.\end{aligned}$$

The implementation of this regularization requires to estimate K first and to select c second. In some cases, $\|K\|$ is known a priori. For example, if K is the conditional expectation operator (see (2.16)), $\|K\|=1$.

For a given c and regularization parameter α_n, the estimator of φ is

$$\hat{\varphi}_n = c\sum_{l=0}^{1/\alpha_n-1}\left(I-c\hat{K}_n^*\hat{K}_n\right)^l\hat{K}_n^*\hat{r}_n.$$

$\hat{\varphi}_n$ can be computed recursively by

$$\hat{\varphi}_{l,n} = \left(I-c\hat{K}_n^*\hat{K}_n\right)\hat{\varphi}_{l-1,n}+c\hat{K}_n^*\hat{r}_n,\ l=1,2,...,1/\alpha_n-1.$$

starting with $\hat{\varphi}_{0,n}=c\hat{K}_n^*\hat{r}_n$. This scheme is known as the Landweber-Fridman iteration (see Kress, 1999, p. 287). Whereas Tikhonov requires the inversion of a $L_n\times L_n-$matrix, Landweber-Fridman is an iterative method.

The extended spectral cut-off regularization is given by

$$\varphi_{\alpha_n} = \sum_{j/\lambda_j\geq\sqrt{\alpha_n}}\frac{1}{\lambda_j}\left\langle r,\psi_j\right\rangle\phi_j+\sum_{j/\lambda_j<\sqrt{\alpha_n}}\frac{\lambda_j}{\alpha_n}\left\langle r,\psi_j\right\rangle\phi_j$$

for some threshold α_n. The estimator is

$$\hat{\varphi}_n = \sum_{j/\hat{\lambda}_j\geq\sqrt{\alpha_n}}\frac{1}{\hat{\lambda}_j}\left\langle \hat{r}_n,\hat{\psi}_j\right\rangle\hat{\phi}_j+\sum_{j/\hat{\lambda}_j<\sqrt{\alpha_n}}\frac{\hat{\lambda}_j}{\alpha_n}\left\langle \hat{r}_n,\hat{\psi}_j\right\rangle\hat{\phi}_j.$$

3.4. Regularization bias

In this subsection, we focus on the control of the bias associated with the regularized solution $\varphi_\alpha = R_\alpha r = R_\alpha K\varphi$:

$$\|\varphi_\alpha-\varphi\| = \|\left(R_\alpha K-I\right)\varphi\|.$$

More precisely, we would like to characterize the decay rate of $\|\varphi_\alpha - \varphi\|$ when α goes to zero. First, it is important to realize that the norm of bounded operators does not help in this respect. Indeed, we have

$$\|(R_\alpha K - I)\,\varphi\| \leq \|R_\alpha K - I\|\,\|\varphi\|$$

but $\|R_\alpha K - I\|$ does not converge toward zero when α goes to zero. To see this (see also Kress, 1999, Theorem 15.6), suppose for a moment that for some $\bar{\alpha} > 0$:

$$\|R_{\bar{\alpha}} K - I\| \leq \frac{1}{2}.$$

Then

$$\begin{aligned} \left\|K^{-1}r\right\| &\leq \left\|K^{-1}r - R_{\bar{\alpha}} K K^{-1} r\right\| + \|R_{\bar{\alpha}} r\| \\ &\leq \frac{1}{2}\left\|K^{-1}r\right\| + \|R_{\bar{\alpha}}\|\,\|r\| \\ &\Rightarrow \left\|K^{-1}r\right\| \leq 2\,\|R_{\bar{\alpha}}\|\,\|r\| \end{aligned}$$

which would imply that the operator K^{-1} is bounded with $\|K^{-1}\| \leq 2\,\|R_{\bar{\alpha}}\|$. Since we know that it is not the case, there is no hope to get some convergence results about $\|\varphi_\alpha - \varphi\|$ which would be uniform on the unit sphere. Therefore we must consider some specific sets $\mathcal{U}$ of functions φ to solve the equation:

$$K\varphi = r,\ \varphi \in \mathcal{U}.$$

As already announced, we do not consider the case of regularization by compacity and we prefer to focus on specific sets $\mathcal{U} = \Phi_\beta$, $\beta > 0$ of functions φ characterized through the rate of convergence of their Fourier coefficients. From (3.11), we know that

$$\|\varphi - \varphi_\alpha\|^2 = \sum_{j=1}^{\infty} \left[1 - q\left(\alpha, \lambda_j\right)\right]^2 \left\langle \varphi, \phi_j \right\rangle^2. \tag{3.19}$$

Since for $\varphi \in \Phi_\beta$,

$$\sum_{j=1}^{\infty} \frac{\left\langle \varphi, \phi_j \right\rangle^2}{\lambda_j^{2\beta}} < \infty,$$

the bias (3.19) will be controlled on Φ_β if the function

$$\lambda \to \left[1 - q\left(\alpha, \lambda\right)\right]^2 \lambda^{2\beta} \tag{3.20}$$

is upper bounded uniformly on $[0, \|K\|]$ by a function of α going to zero with α. Note that, since

$$\begin{aligned} \lim_{\lambda \to 0} \left[1 - q\left(\alpha, \lambda\right)\right]^2 \lambda^{2\beta} &= 0, \\ \lim_{\alpha \to 0} \left[1 - q\left(\alpha, \lambda\right)\right]^2 \lambda^{2\beta} &= 0, \end{aligned}$$

we may expect that the maximum of (3.20) is reached for a positive value $\lambda_{\alpha,\beta}$ which goes to zero with α. We confine our attention to the regularization schemes defined below.

Definition 3.7. *A regularization scheme $q(\alpha, \lambda)$ is geometrically unbiased if, for all $\beta \in (0, 2)$ and α positive in a neighborhood of zero,*

$$\lambda_{\alpha,\beta} \equiv \arg \max_{\lambda \in [0, \|K\|]} \left[1 - q(\alpha, \lambda)\right]^2 \lambda^{2\beta} \in \mathbb{R}_*^+$$

satisfies

$$\lim_{\alpha \to 0} \frac{\lambda_{\alpha,\beta}^2}{\alpha} \in \mathbb{R}_*^+, \tag{3.21}$$

$$\lim_{\alpha \to 0} \left[1 - q(\alpha, \lambda_{\alpha,\beta})\right]^2 \in \mathbb{R}_*^+. \tag{3.22}$$

Remark that if $\lim_{\alpha \to 0} \lambda_{\alpha,\beta}^2 / \alpha = \gamma > 0$, one may expect that

$$\lim_{\alpha \to 0} q(\alpha, \lambda_{\alpha,\beta}) \in (0, 1), \tag{3.23}$$

that is (3.22), since $\lim_{\alpha \to 0} q(\alpha, \lambda) = 1$ for all λ and $\lim_{\lambda \to 0} q(\alpha, \lambda) = 0$ for all α. Actually, it is easy to check that the regularization schemes considered previously are all geometrically unbiased except for the spectral cut-off.

Example (Tikhonov regularization - continued)

$$\lambda_{\alpha,\beta} = \arg \max_{\lambda} \frac{\alpha^2}{\left(\alpha + \lambda^2\right)^2} \lambda^{2\beta}$$

gives for all $\beta \in (0, 2)$:

$$\lambda_{\alpha,\beta}^2 = \frac{\alpha\beta}{2 - \beta}.$$

Then

$$\frac{\lambda_{\alpha,\beta}^2}{\alpha} = \frac{\beta}{2 - \beta} \in \mathbb{R}_*^+$$

and

$$1 - q(\alpha, \lambda_{\alpha,\beta}) = \frac{\alpha}{\alpha + \frac{\alpha\beta}{2-\beta}} = 1 - \frac{\beta}{2} \in \mathbb{R}_*^+.$$

In other words, the functions of α defined in (3.21) and (3.22) are positive constant and then coincide identically with their limits when α goes to zero.

Example (Landweber-Fridman - continued)

$$\lambda_{\alpha,\beta} = \arg \max_{\lambda} \left[1 - c\lambda^2\right]^{2/\alpha} \lambda^{2\beta}$$

gives

$$\lambda_{\alpha,\beta} = \frac{\beta}{c}\left[\beta + \frac{2}{\alpha}\right]^{-1}.$$

Then, we have

$$\lim_{\alpha\to 0}\frac{\lambda_{\alpha,\beta}^2}{\alpha} = \lim_{\alpha\to 0}\frac{\beta}{c}\left[\alpha\beta + 2\right]^{-1} = \frac{\beta}{2c} \in \mathbb{R}_*^+$$

and

$$\begin{aligned}\lim_{\alpha\to 0}\left[1 - q\left(\alpha, \lambda_{\alpha,\beta}\right)\right]^2 &= \lim_{\alpha\to 0}\left[1 - \frac{\beta}{\beta + \frac{2}{\alpha}}\right]^{2/\alpha} \\ &= \lim_{\alpha\to 0}\exp\left[\frac{2}{\alpha}\ln\left(1 - \frac{\beta}{\beta + \frac{2}{\alpha}}\right)\right] \\ &= \exp\left(-\beta\right) \in \mathbb{R}_*^+.\end{aligned}$$

Example (Extended Spectral cut-off - continued)

$$\lambda_{\alpha,\beta} = \arg\max_{\lambda}\left[1 - \frac{\lambda^2}{\alpha}\right]^2 I\left\{\lambda < \sqrt{\alpha}\right\}\lambda^{2\beta}$$

gives

$$\lambda_{\alpha,\beta}^2 = \frac{\alpha\beta}{\beta + 2}.$$

Then,

$$\frac{\lambda_{\alpha,\beta}^2}{\alpha} = \frac{\beta}{\beta + 2} \in \mathbb{R}_*^+$$

and

$$1 - q\left(\alpha, \lambda_{\alpha,\beta}\right) = 1 - \frac{\beta}{\beta + 2} = \frac{2}{\beta + 2} \in \mathbb{R}_*^+$$

for all $\beta \in (0, 2)$. In this case too, the functions of α defined in (3.21) and (3.22) are positive constant.

Geometrically unbiased regularization schemes allow one to get geometric rates of convergence of the regularization:

Proposition 3.8. *Let $K : \mathcal{H} \to \mathcal{E}$ be an injective compact operator. Let us assume that the solution φ of $K\varphi = r$ is in the $\beta-$regularity space of operator K, for some $\beta \in (0, 2)$. Then, if φ_α is defined by a second order geometrically unbiased regularization scheme, we have*

$$\left\|\varphi_\alpha - \varphi\right\|^2 = O\left(\alpha^\beta\right).$$

Proof. To see this, just note that

$$\frac{\|\varphi - \varphi_\alpha\|^2}{\alpha^\beta} = \sum_{j=1}^{\infty} \frac{[1 - q(\alpha, \lambda_j)]^2 \langle \varphi, \phi_j \rangle^2}{\alpha^\beta}$$
$$\leq \sum_{j=1}^{\infty} [1 - q(\alpha, \lambda_{\alpha,\beta})]^2 \frac{\lambda_{\alpha,\beta}^{2\beta}}{\alpha^\beta} \frac{\langle \varphi, \phi_j \rangle^2}{\lambda_j^{2\beta}}$$

But, by Assumptions (3.19) and (3.20), the function

$$\alpha \to [1 - q(\alpha, \lambda_{\alpha,\beta})]^2 \frac{\lambda_{\alpha,\beta}^{2\beta}}{\alpha^\beta}$$

admits a finite upper bound c.

Then

$$\frac{\|\varphi - \varphi_\alpha\|^2}{\alpha^\beta} \leq c \sum_{j=1}^{\infty} \frac{\langle \varphi, \phi_j \rangle^2}{\lambda_j^{2\beta}} < +\infty$$

for $\varphi \in \Phi_\beta$. ■

In other words, while the solvability criterion of $K\varphi = r$ imposes $r \in \Psi_\gamma$ for some $\gamma \geq 1$, we will get a geometric rate of decay of the regularization bias as soon as $r \in \Psi_\gamma$ for some $\gamma > 1$. Note that the upper bound $\beta = 2$ of the rate of decay is just a matter of normalization of the parameter α of regularization. By (3.19), $\|\varphi_\alpha - \varphi\|^2$ cannot go to zero faster than $[1 - q(\alpha, \lambda)]^2$ (for a given λ), that is α^2.

Finally, note that Vanhems and Loubes (2003) give conditions on regularization schemes such that the reciprocal of (3.8) is verified. They show in particular that, for Tikhonov regularization Φ_β is exactly the set of function φ such that

$$\frac{\|\varphi - \varphi_\alpha\|^2}{\alpha^\beta} \leq c \sum_{j=1}^{\infty} \frac{\langle \varphi, \phi_j \rangle^2}{\lambda_j^{2\beta}} < +\infty.$$

3.5. Estimation bias

Regularization schemes have precisely been introduced because the right hand side r of the inverse problem $K\varphi = r$ is generally unknown and replaced by an estimator. Let us denote by $\hat{r}_n$ an estimator computed from an observed sample of size n. As announced in the introduction, a number of relevant inverse problems in econometrics are even more complicated since the operator K itself is unknown.

Actually, in order to apply a regularization scheme, we may need not only an estimator of K but also of its adjoint K^* and of its singular system $\{\lambda_j, \phi_j, \psi_j : j = 1, 2, ...\}$. In this subsection, we consider such estimators $\hat{K}_n$, $\hat{K}_n^*$ and $\left\{\hat{\lambda}_j, \hat{\phi}_j, \hat{\psi}_j : j = 1, ..., L_n\right\}$ as given. We also maintain the identification assumption, so that the equation $K\varphi = r$ defines without ambiguity a true unknown value φ_0.

If $\varphi_\alpha = A_\alpha K^* r$ is the chosen regularized solution, the proposed estimator $\hat{\varphi}_n$ of φ_0 is defined by

$$\hat{\varphi}_n = \hat{A}_{\alpha_n} \hat{K}_n^* \hat{r}_n. \tag{3.24}$$

Note that the definition of this estimator involves two decisions: First, we need to select a sequence (α_n) of regularization parameters so that $\lim_{n\to\infty} \alpha_n = 0$ (possibly in a stochastic sense in the case of a data-driven regularization) in order to get a consistent estimator of φ_0. Second, for a given α_n, we estimate the second order regularization scheme $A_{\alpha_n} K^*$ by $\hat{A}_{\alpha_n} \hat{K}_n^*$. Generally speaking, $\hat{A}_{\alpha_n}$ is defined from (3.14) where the singular values are replaced by their estimators and the inner products $\langle \varphi, \phi_j \rangle$ are replaced by some empirical counterpart (see Subsection 2.5.1). Yet, we will show later (in Subsection 3.3) that in some cases, the estimation of the regularized solution does not involve the estimators $\hat{\lambda}_j$ but only the estimators $\hat{K}_n$ and $\hat{K}_n^*$.

In any case, the resulting estimator bias $\hat{\varphi}_n - \varphi_0$ has two components:

$$\hat{\varphi}_n - \varphi_0 = \hat{\varphi}_n - \varphi_{\alpha_n} + \varphi_{\alpha_n} - \varphi_0. \tag{3.25}$$

While the second component $\varphi_{\alpha_n} - \varphi_0$ defines the regularization bias characterized in the previous subsection 3.4, the first component $\hat{\varphi}_n - \varphi_{\alpha_n}$ is the bias corresponding to the estimation of the regularized solution of φ_{α_n}. The goal of this subsection is to point out a set of statistical assumptions about the estimators $\hat{K}_n$, $\hat{K}_n^*$, and $\hat{r}_n$ that allow to upper bound (asymptotically) the specific estimation bias magnitude $\left\|\hat{\varphi}_n - \varphi_{\alpha_n}\right\|$ when the regularization bias $\left\|\varphi_{\alpha_n} - \varphi_0\right\|$ is controlled.

Proposition 3.9 (Estimation bias). *If $\varphi_\alpha = A_\alpha K^* r$ is the regularized solution where A_α is a second order regularization scheme conformable to (3.14)-(3.15) and $\hat{\varphi}_n = \hat{A}_{\alpha_n} \hat{K}_n^* \hat{r}_n$, then*

$$\begin{aligned} & \left\|\hat{\varphi}_n - \varphi_{\alpha_n}\right\| \\ \leq \;& d(\alpha_n) \left\|\hat{K}_n^* \hat{r}_n - \hat{K}_n^* \hat{K}_n \varphi_0\right\| + \left\|\left(\hat{A}_{\alpha_n} \hat{K}_n^* \hat{K}_n - A_{\alpha_n} K^* K\right) \varphi_0\right\| \end{aligned} \tag{3.26}$$

If, in addition, a regularization scheme is said to be smooth when

$$\begin{aligned} & \left\|\left(\hat{A}_{\alpha_n} \hat{K}_n^* \hat{K}_n - A_{\alpha_n} K^* K\right) \varphi_0\right\| \\ \leq \;& d(\alpha_n) \left\|\hat{K}_n^* \hat{K}_n - K^* K\right\| \left\|\varphi_{\alpha_n} - \varphi_0\right\| (1 + \varepsilon_n) \end{aligned} \tag{3.27}$$

with $\varepsilon_n = O\left(\left\|\hat{K}_n^* \hat{K}_n - K^* K\right\|\right)$, then both the Tikhonov and Landweber-Fridman regularization schemes are smooth. In the Tikhonov case, $\varepsilon_n = 0$ identically.

Proof.

$$\begin{aligned} \hat{\varphi}_n - \varphi_{\alpha_n} &= \hat{A}_{\alpha_n} \hat{K}_n^* \hat{r}_n - A_{\alpha_n} K^* r \\ &= \hat{A}_{\alpha_n} \hat{K}_n^* \left(\hat{r}_n - \hat{K}_n \varphi_0\right) + \hat{A}_{\alpha_n} \hat{K}_n^* \hat{K}_n \varphi_0 - A_{\alpha_n} K^* K \varphi_0 \end{aligned}$$

Thus,

$$\left\|\hat{\varphi}_n - \varphi_{\alpha_n}\right\| \leq d\left(\alpha_n\right)\left\|\hat{K}_n^*\hat{r}_n - \hat{K}_n^*\hat{K}_n\varphi_0\right\| + \left\|\hat{A}_{\alpha_n}\hat{K}_n^*\hat{K}_n\varphi_0 - A_{\alpha_n}K^*K\varphi_0\right\|.$$

- Case of Tikhonov regularization:

$$\begin{aligned} & \hat{A}_{\alpha_n}\hat{K}_n^*\hat{K}_n\varphi_0 - A_{\alpha_n}K^*K\varphi_0 \\ = \;& \hat{A}_{\alpha_n}\left(\hat{K}_n^*\hat{K}_n - K^*K\right)\varphi_0 + \left(\hat{A}_{\alpha_n} - A_{\alpha_n}\right)K^*K\varphi_0. \end{aligned} \tag{3.28}$$

  Since, in this case,

$$A_\alpha = \left(\alpha I + K^*K\right)^{-1},$$

  the identity

$$B^{-1} - C^{-1} = B^{-1}(C - B)C^{-1}$$

  gives

$$\hat{A}_{\alpha_n} - A_{\alpha_n} = \hat{A}_{\alpha_n}\left(K^*K - \hat{K}_n^*\hat{K}_n\right)A_{\alpha_n}$$

  and thus,

$$\begin{aligned} \left(\hat{A}_{\alpha_n} - A_{\alpha_n}\right)K^*K\varphi_0 &= \hat{A}_{\alpha_n}\left(K^*K - \hat{K}_n^*K_n\right)A_{\alpha_n}K^*K\varphi_0 \\ &= \hat{A}_{\alpha_n}\left(K^*K - \hat{K}_n^*\hat{K}_n\right)\varphi_{\alpha_n}. \end{aligned} \tag{3.29}$$

  (3.28) and (3.29) together give

$$\begin{aligned} & \hat{A}_{\alpha_n}\hat{K}_n^*\hat{K}_n\varphi_0 - A_{\alpha_n}K^*K\varphi_0 \\ = \;& \hat{A}_{\alpha_n}\left(\hat{K}_n^*\hat{K}_n - K^*K\right)\left(\varphi_0 - \varphi_{\alpha_n}\right), \end{aligned}$$

  which shows that Tikhonov regularization is smooth with $\varepsilon_n = 0$.

- Case of Landweber-Fridman regularization:
  In this case,

$$\begin{aligned} \varphi_\alpha &= \sum_{j=1}^{\infty}\left[1 - \left(1 - c\lambda_j^2\right)^{1/\alpha}\right] < \varphi_0, \varphi_j > \varphi_j \\ &= \left[I - \left(I - cK^*K\right)^{1/\alpha}\right]\varphi_0. \end{aligned}$$

Thus,

$$\begin{aligned}
& \hat{A}_{\alpha_n}\hat{K}_n^*\hat{K}_n\varphi_0 - A_{\alpha_n}K^*K\varphi_0 \\
= & \left[(I-cK^*K)^{1/\alpha_n} - \left(I-c\hat{K}_n^*\hat{K}_n\right)^{1/\alpha_n}\right]\varphi_0 \\
& \left[I-\left(I-c\hat{K}_n^*\hat{K}_n\right)^{1/\alpha_n}(I-cK^*K)^{-1/\alpha_n}\right](I-cK^*K)^{1/\alpha_n}\varphi_0 \\
& \left[I-\left(I-c\hat{K}_n^*\hat{K}_n\right)^{1/\alpha_n}(I-cK^*K)^{-1/\alpha_n}\right]\left(\varphi_0-\varphi_{\alpha_n}\right).
\end{aligned}$$

Then, a Taylor expansion gives:

$$\begin{aligned}
& \left\|I-\left(I-c\hat{K}_n^*\hat{K}_n\right)^{1/\alpha_n}(I-cK^*K)^{-1/\alpha_n}\right\| \\
= & \left\|\frac{c}{\alpha_n}\left(\hat{K}_n^*\hat{K}_n-K^*K\right)\right\|(1+\varepsilon_n)
\end{aligned}$$

with $\varepsilon_n = O\left(\left\|\hat{K}_n^*\hat{K}_n - K^*K\right\|\right)$.

The result follows with $d(\alpha) = c/\alpha$. ■

Note that a similar result has not been derived for a spectral cut off regularization. In that case, the threshold introduces a lack of smoothness which precludes a similar Taylor expansion based argument.

The result of Proposition 3.9 jointly with (3.25) shows that two ingredients matter in controlling the estimation bias $\|\hat{\varphi}_n - \varphi_0\|$. First, the choice of a sequence of regularization parameters α_n will govern the speed of convergence to zero of the regularization bias $\left\|\varphi_{\alpha_n} - \varphi_0\right\|$ (for φ_0 in a given Φ_β) and the speed of convergence to infinity of $d(\alpha_n)$. Second, nonparametric estimation of K and r will determine the rate of convergence of $\left\|\hat{K}_n^*\hat{r}_n - \hat{K}_n^*\hat{K}_n\varphi_0\right\|$ and $\left\|\hat{K}_n^*\hat{K}_n - K^*K\right\|$.

4. Asymptotic properties of solutions of integral equations of the first kind 4

4.1. Consistency

Let φ_0 be the solution of $K\varphi = r$. By abuse of notation, we denote $X_n = O(c_n)$ for positive sequences $\{X_n\}$ and $\{c_n\}$, if the sequence X_n/c_n is upper bounded.

We maintain the following assumptions:

A1. $\hat{K}_n$, $\hat{r}_n$ are consistent estimators of K and r.

A2. $\left\|\hat{K}_n^*\hat{K}_n - K^*K\right\| = O\left(\frac{1}{a_n}\right)$

A3. $\left\|\hat{K}_n^*\hat{r}_n - \hat{K}_n^*\hat{K}_n\varphi_0\right\| = O\left(\frac{1}{b_n}\right)$

As before $\varphi_\alpha = A_\alpha K^* r$ is the regularized solution where A_α is a second order regularization scheme and $\hat{\varphi}_n = \hat{A}_{\alpha_n}\hat{K}_n^*\hat{r}_n$. Proposition 4.1 below follows directly from Proposition 3.9 and Definition 3.6 (with the associated normalization rule $\alpha d(\alpha) = O(1)$):

Proposition 4.1. *When applying a smooth regularization scheme, we get:*

$$\begin{aligned} & \|\hat{\varphi}_n - \varphi_0\| \\ = \; & O\left(\frac{1}{\alpha_n b_n} + \left(\frac{1}{\alpha_n a_n} + 1\right)\left\|\varphi_{\alpha_n} - \varphi_0\right\|\right). \end{aligned}$$

Discussion on the rate of convergence:

The general idea is that the fastest possible rate of convergence in probability of $\|\hat{\varphi}_n - \varphi_0\|$ to zero should be the rate of convergence of the regularization bias $\left\|\varphi_{\alpha_n} - \varphi_0\right\|$. Proposition 4.1 shows that these two rates of convergence will precisely coincide when the rate of convergence to zero of the regularization parameter α_n is chosen sufficiently slow with respect to both the rate of convergence a_n of the sequence of approximations of the true operator and the rate of convergence b_n of the estimator of the right-hand side of the operator equation. This is actually a common strategy when both the operator and the right-hand side of the inverse problem have to be estimated (see e.g. Vapnik (1998), corollary p. 299).

To get this, it is first obvious that $\alpha_n b_n$ must go to infinity at least as fast as $\left\|\varphi_{\alpha_n} - \varphi_0\right\|^{-1}$. For $\varphi_0 \in \Phi_\beta$, $0 < \beta < 2$, this means that:

$$\alpha_n^2 b_n^2 \geq \alpha_n^{-\beta}$$

that is $\alpha_n \geq b_n^{-\frac{2}{\beta+2}}$. To get the fastest possible rate of convergence under this constraint, we will choose:

$$\alpha_n = b_n^{-\frac{2}{\beta+2}}.$$

Then, the rate of convergence of $\|\hat{\varphi}_n - \varphi_0\|$ and $\left\|\varphi_{\alpha_n} - \varphi_0\right\|$ will coincide if and only if $a_n b_n^{-\frac{2}{\beta+2}}$ is bounded away from zero. With this way to define optimality, we have thus proved:

Proposition 4.2. *Consider a smooth regularization scheme, with estimators of K and r conformable to Assumptions A1, A2, A3, and $a_n b_n^{-\frac{2}{\beta+2}}$ bounded away from zero. For $\varphi_0 \in \Phi_\beta, 0 < \beta < 2$, the optimal choice of the regularization parameter is $\alpha_n = b_n^{-\frac{2}{\beta+2}}$, and then,*

$$\|\hat{\varphi}_n - \varphi_0\| = O\left(b_n^{-\frac{\beta}{\beta+2}}\right).$$

Note that the only condition about the estimator of the operator K^*K is that its rate of convergence a_n is sufficiently fast to be greater than $b_n^{\frac{2}{\beta+2}}$. But, under this condition, the rate of convergence of $\hat{\varphi}_n$ does not depend upon the accuracy of the estimator of K^*K. Of course, the more regular the unknown function φ_0 is, the larger β is and the easier it will be to meet the required condition. Closer is β to its upper bound 2, closer is the rate of convergence of $\hat{\varphi}_n$ to the parametric rate. Generally speaking, the condition will involve the relative bandwidth sizes in the nonparametric estimation of K^*K and K^*r. Note that if, as it is generally the case for a convenient bandwidth choice (see e.g. subsection 5.4), b_n is the parametric rate ($b_n = \sqrt{n}$), a_n must be at least $n^{-1/\beta+2}$. For β not too small, this condition will fulfilled by optimal nonparametric rates. For instance, the optimal unidimensional nonparametric rate $n^{-2/5}$ will work as soon as $\beta \geq 1/2$.

4.2. Asymptotic normality

Asymptotic normality of

$$\begin{aligned}\hat{\varphi}_n - \varphi_0 &= \hat{\varphi}_n - \varphi_{\alpha_n} + \varphi_{\alpha_n} - \varphi_0 \\ &= \hat{A}_{\alpha_n}\hat{K}_n^*\hat{r}_n - A_{\alpha_n}K^*K\varphi_0 + \varphi_{\alpha_n} - \varphi_0\end{aligned}$$

can be deduced from a functional central limit theorem applied to $\hat{K}_n^*\hat{r}_n - \hat{K}_n^*\hat{K}_n\varphi_0$. Therefore, we must reinforce Assumption A3 by assuming a weak convergence in $\mathcal{H}$:

Assumption WC:

$$b_n\left(\hat{K}_n^*\hat{r}_n - \hat{K}_n^*\hat{K}_n\varphi_0\right) \Rightarrow \mathcal{N}(0,\Sigma) \text{ in } \mathcal{H}.$$

According to (3.26), (3.28), and (3.29), we have in the case of Tikhonov regularization:

$$\begin{aligned}b_n(\hat{\varphi}_n - \varphi_0) &= b_n\hat{A}_{\alpha_n}\left[\hat{K}_n^*\hat{r}_n - \hat{K}_n^*\hat{K}_n\varphi_0\right] &(4.1)\\ &\quad + b_n\hat{A}_{\alpha_n}\left[\hat{K}_n^*\hat{K}_n - K^*K\right](\varphi_0 - \varphi_{\alpha_n}) &(4.2)\end{aligned}$$

while an additional term corresponding ε_n in (3.27) should be added for general regularization schemes. The term (4.1) can be rewritten as

$$\hat{A}_{\alpha_n}\xi + \hat{A}_{\alpha_n}(\xi_n - \xi)$$

where ξ denotes the random variable $\mathcal{N}(0,\Sigma)$ in $\mathcal{H}$ and

$$\xi_n = b_n\left(\hat{K}_n^*r_n - \hat{K}_n^*\hat{K}_n\varphi_0\right).$$

By definition:

$$\frac{\left\langle \hat{A}_{\alpha_n}\xi, g\right\rangle}{\left\|\Sigma^{1/2}\hat{A}_{\alpha_n}g\right\|} \xrightarrow{d} \mathcal{N}(0,1)$$

for all $g \in \mathcal{H}$. Then, we may hope to get a standardized normal asymptotic probability distribution for

$$\frac{\langle b_n (\hat{\varphi}_n - \varphi_0), g \rangle}{\left\| \Sigma^{1/2} \hat{A}_{\alpha_n} g \right\|}$$

for vectors g conformable to the following assumption:

Assumption G

$$\frac{\left\| \hat{A}_{\alpha_n} g \right\|}{\left\| \Sigma^{1/2} \hat{A}_{\alpha_n} g \right\|} = O(1).$$

Indeed, we have in this case:

$$\frac{\left| \left\langle \hat{A}_{\alpha_n} (\xi_n - \xi), g \right\rangle \right|}{\left\| \Sigma^{1/2} \hat{A}_{\alpha_n} g \right\|} \leq \frac{\|\xi_n - \xi\| \left\| \hat{A}_{\alpha_n} g \right\|}{\left\| \Sigma^{1/2} \hat{A}_{\alpha_n} g \right\|}$$

which converges to zero in probability because $\|\xi_n - \xi\| \xrightarrow{P} 0$ by WC. We are then able to show:

Proposition 4.3. *Consider a Tikhonov regularization. Suppose Assumptions A1, A2, A3, and WC hold and $\varphi_0 \in \Phi_\beta$, $0 < \beta < 2$, with $b_n \alpha_n^{\beta/2} \underset{n \to \infty}{\to} 0$, we have for any g conformable to G:*

$$\frac{\langle b_n (\hat{\varphi}_n - \varphi_0), g \rangle}{\left\| \Sigma^{1/2} \hat{A}_{\alpha_n} g \right\|} \xrightarrow{d} \mathcal{N}(0, 1).$$

Proof. ¿From the proof of Proposition 3.9, we have:

$$\begin{aligned} & \langle b_n (\hat{\varphi}_n - \varphi_{\alpha_n}), g \rangle \\ = & \left\langle \hat{A}_{\alpha_n} \xi, g \right\rangle \\ & + \left\langle \hat{A}_{\alpha_n} (\xi_n - \xi), g \right\rangle \\ & + \left\langle b_n \hat{A}_{\alpha_n} \left[\hat{K}_n^* \hat{K}_n - K^* K \right] (\varphi_0 - \varphi_{\alpha_n}), g \right\rangle \end{aligned} \tag{4.3}$$

in the case of Tikhonov regularization. We already took care of the terms in ξ and ξ_n, it

remains to deal with the bias term corresponding to (4.3):

$$\begin{aligned}
& \frac{b_n \left\langle \hat{A}_{\alpha_n} \left(\hat{K}_n^* \hat{K}_n - K^* K \right) \left(\varphi_0 - \varphi_{\alpha_n} \right), g \right\rangle}{\left\| \Sigma^{1/2} \hat{A}_{\alpha_n} g \right\|} \\
\leq & \frac{b_n \left\langle \left(\hat{K}_n^* \hat{K}_n - K^* K \right) \left(\varphi_0 - \varphi_{\alpha_n} \right), \hat{A}_{\alpha_n} g \right\rangle}{\left\| \Sigma^{1/2} \hat{A}_{\alpha_n} g \right\|} \\
\leq & b_n \left\| \hat{K}_n^* \hat{K}_n - K^* K \right\| \left\| \varphi_0 - \varphi_{\alpha_n} \right\| \frac{\left\| \hat{A}_{\alpha_n} g \right\|}{\left\| \Sigma^{1/2} \hat{A}_{\alpha_n} g \right\|} \\
= & O\left(\frac{b_n \alpha_n^{\beta/2}}{a_n} \right).
\end{aligned}$$

■

Discussion of Proposition 4.3.

(i) It is worth noticing that Proposition 4.3 does not deliver in general a weak convergence result for $b_n \left(\hat{\varphi}_n - \varphi_0 \right)$ because it does not hold for all $g \in \mathcal{H}$. However, the condition G is not so restrictive. It just amounts to assume that the multiplication by $\Sigma^{1/2}$ does not modify the rate of convergence of $\hat{A}_{\alpha_n} g$.

(ii) We remark that for $g = K^* K h$, $\hat{A}_{\alpha_n} g$ and $\Sigma^{1/2} \hat{A}_{\alpha_n} g$ converge respectively to h and $\Sigma^{1/2} h$. Moreover, if $g \neq 0$, $\Sigma^{1/2} h = \Sigma^{1/2} \left(K^* K \right)^{-1} g \neq 0$. Therefore, in this case, not only the condition G is fulfilled but we get asymptotic normality with rate of convergence b_n, that is typically root n. This result is conformable to the theory of asymptotic efficiency of inverse estimators as recently developed by Van Rooij, Ruymgaart and Van Zwet (2000). They show that there is a dense linear submanifold of functionals for which the estimators are asymptotically normal at the root n rate with optimal variance (in the sense of minimum variance in the class of the moment estimators). We do get optimal variance in Proposition 4.3 in this case since (using heuristic notations as if we were in finite dimension) the asymptotic variance is:

$$\begin{aligned}
& \lim_{n \to \infty} g' A_{\alpha_n} \Sigma A_{\alpha_n} \\
= & g' \left(K^* K \right)^{-1} \Sigma \left(K^* K \right)^{-1} g.
\end{aligned}$$

Moreover, we get this result in particular for any nonzero g in $\mathcal{R}\left(K^* K \right)$ while we know that $\mathcal{R}\left(K^* \right)$ is dense in $\mathcal{H}$ (identification condition). Generally speaking, Van Rooij, Ruymgaart and Van Zwet (2000) stress that the inner products do not converge weakly for every vector g in $\mathcal{H}$ at the same rate, if they converge at all.

(iii) The condition $b_n \alpha_n^{\beta/2} \to 0$ imposes a convergence to zero of the regularization coefficient α_n faster than the optimal rate $\alpha_n = b_n^{-2/(\beta+2)}$. This condition is needed to

show that the regularization bias multiplied by b_n converges to zero. A fortiori, the estimation bias term vanishes asymptotically.

The results of Proposition 4.3 are established under strong assumptions: convergence in $\mathcal{H}$ and restriction on g. An alternative method consists in establishing the normality of $\hat{\varphi}_n$ by the Liapunov condition (Davidson, 1994), see the example on deconvolution in Section 5 below.

5. Applications

A well-known example is that of the kernel estimator of the density. Indeed, the estimation of the pdf f of a random variable X can be seen as solving an integral equation of the first kind

$$Kf(x) = \int_{-\infty}^{+\infty} I(u \leq x) f(u) du = F(x) \tag{5.1}$$

where F is the cdf of X. Applying the Tikhonov regularization to (5.1), one obtains a kernel estimator of f. This example is detailed in Hardle and Linton (1994) and in Vapnik (1998) and will not be discussed further.

This section review the standard example of the Ridge regression and less standard examples such as the regression with an infinity of regressors, the deconvolution and the instrumental variable estimation.

5.1. Ridge regression

The Tikhonov regularization discussed in Section 3 can be seen as an extension of the well-known ridge regression. The ridge regression was introduced by Hoerl and Kennard (1970). It was initially motivated by the fact that in presence of nearly multicollinearity of the regressors, the least squares estimator may vary dramatically as the result of a small perturbation in the data. The ridge estimator may also have a lower risk than the conventional least squares estimator. For a review of this method, see Judge, Griffiths, Hill, Lutkepohl, and Lee (1980) and for a discussion in the context of inverse problems, see Ruymgaart (2001).

Consider the linear model (the notation of this paragraph are specific and corresponds of general notations of linear models).

$$y = X\beta + \varepsilon$$

where y and ε are $n \times 1-$random vectors, X is a $n \times q$ matrix of regressors of full rank, and β is an unknown $q \times 1-$vector of parameters. The classical least-squares estimator of β is

$$\hat{\beta} = (X'X)^{-1} X'y.$$

There exists an orthogonal transformation such that $X'X = P'DP$ with

$$D = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_q \end{pmatrix}$$

$\lambda_j > 0$ and $P'P = I_q$. Using the mean square error as measure of the risk, we get

$$\begin{aligned} E\left\|\hat{\beta} - \beta\right\|^2 &= E\left\|\left(X'X\right)^{-1}\left(X'\left(X\beta + \varepsilon\right) - \beta\right)\right\|^2 \\ &= E\left\|\left(X'X\right)^{-1} X'\varepsilon\right\|^2 \\ &= E\left(\varepsilon' X\left(X'X\right)^{-2} X'\varepsilon\right) \\ &= \sigma^2 trace\left(X\left(X'X\right)^{-2} X'\right) \\ &= \sigma^2 trace\left(\left(X'X\right)^{-1}\right) \\ &= \sigma^2 trace\left(P'D^{-1}P\right) \\ &= \sigma^2 \sum_{j=1}^{q} \frac{1}{\lambda_j}. \end{aligned}$$

If some of the columns of X are closely collinear, the eigenvalues may be very small and the risk very large.

A solution is to use the ridge regression estimator:

$$\hat{\beta}_a = \left(aI + X'X\right)^{-1} X'y,\ a > 0.$$

or

$$\hat{\beta}_a = \left(\alpha + \frac{1}{n}X'X\right)^{-1} \frac{1}{n}X'y,\ a > 0.$$

where $\alpha = \frac{a}{n}$ corresponds to the regularization parameter introduced in the other sections.

$\hat{\beta}_\alpha$ may have a lower risk that $\hat{\beta}$ but it is no longer unbiased. We have

$$\begin{aligned} \beta_a &= E\hat{\beta}_a \\ &= \left(aI + X'X\right)^{-1} X'X\beta. \end{aligned}$$

Using the fact that $A^{-1} - B^{-1} = A^{-1}\left[B - A\right]B^{-1}$. The bias can be rewritten as

$$\begin{aligned} \beta_a - \beta &= \left(aI + X'X\right)^{-1} X'X\beta - \left(X'X\right)^{-1} X'X\beta \\ &= a\left(aI + X'X\right)^{-1}\beta. \end{aligned}$$

The risk becomes

$$\begin{aligned} E\left\|\hat{\beta}_a-\beta\right\|^2 &= E\left\|\hat{\beta}_a-\beta_a\right\|^2+\left\|\beta_a-\beta\right\|^2 \\ &= E\left\|\left(aI+X'X\right)^{-1}X'\varepsilon\right\|^2+a^2\left\|\left(aI+X'X\right)^{-1}\beta\right\|^2 \\ &= E\left(\varepsilon'X\left(aI+X'X\right)^{-2}X'\varepsilon\right)+a^2\left\|\left(aI+X'X\right)^{-1}\beta\right\|^2 \\ &= \sigma^2 trace\left(\left(aI+X'X\right)^{-2}X'X\right)+a^2\left\|\left(aI+X'X\right)^{-1}\beta\right\|^2 \\ &= \sigma^2\sum_{j=1}^{q}\frac{\lambda_j}{\left(a+\lambda_j\right)^2}+a^2\sum_{j=1}^{q}\frac{\left((P\beta)_j\right)^2}{\left(a+\lambda_j\right)^2}. \end{aligned}$$

There is the usual trade-off between the variance (decreasing in a) and the bias (increasing in a). For each β and σ^2, there is a value of a for which the risk of $\hat{\beta}_a$ is smaller than that of $\hat{\beta}$. When q is finite, the inverse of $X'X$ is still continuous and the regularization is not absolutely necessary, however, when the number of regressors is infinite, some kind of regularization needs to be implemented as the risk of $\hat{\beta}$

$$E\left\|\hat{\beta}-\beta\right\|^2=\sigma^2\sum_{j=1}^{\infty}\frac{1}{\lambda_j}$$

is no longer bounded.

5.2. Regression with an infinity of regressors

Consider the following model

$$Y=\int Z(\tau)\varphi(\tau)\pi(d\tau)+U \tag{5.2}$$

where Z is exogenous, π is a known measure (possibly with finite support). One observes $(y_i, z_i(\tau))_{i=1,\ldots,n}$.

First approach: Ridge regression

(5.2) can be rewritten as

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}=\begin{pmatrix} \int z_1(\tau)\varphi(\tau)\pi(d\tau) \\ \vdots \\ \int z_n(\tau)\varphi(\tau)\pi(d\tau) \end{pmatrix}+\begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix}$$

or equivalently

$$y=K\varphi+u$$

where the operator K is defined in the following manner

$$\begin{aligned} K &: L^2(\pi) \to R^n \\ K\varphi &= \begin{pmatrix} \int z_1(\tau)\varphi(\tau)\pi(d\tau) \\ \vdots \\ \int z_n(\tau)\varphi(\tau)\pi(d\tau) \end{pmatrix}. \end{aligned}$$

As it is usual in the regression, the error term u is omitted and we solve

$$K\varphi = y$$

using a regularized inverse

$$\varphi^\alpha = (K^*K + \alpha I)^{-1} K^* y. \tag{5.3}$$

As an exercise, we compute K^* and K^*K. To compute K^*, we solve

$$\langle K\varphi, \psi\rangle = \langle \varphi, K^*\psi\rangle$$

for $\psi = (\psi_1, ..., \psi_n)$ and we obtain

$$\begin{aligned} (K^*y)(\tau) &= \frac{1}{n}\sum_{i=1}^n y_i z_i(\tau), \\ K^*K\varphi(\tau) &= \int \frac{1}{n}\sum_{i=1}^n z_i(\tau) z_i(s)\varphi(s)\pi(ds). \end{aligned}$$

The properties of the estimator (5.3) are further discussed in Van Rooij, Ruymgaart and Van Zwet (2000).

Second approach: moment conditions

Alternatively, (5.2) can be rewritten as

$$E[Y - \langle Z, \varphi\rangle | Z(\tau)] = 0 \text{ for all } \tau \text{ in the support of } \pi$$

Replacing the conditional moments by unconditional moments, we have

$$\begin{aligned} E[YZ(\tau) - \langle Z, \varphi\rangle Z(\tau)] &= 0 \Longleftrightarrow \\ \int E[Z(\tau)Z(s)]\varphi(s)\pi(ds) &= E[YZ(\tau)] \Longleftrightarrow \\ T\varphi &= r. \end{aligned} \tag{5.4}$$

The operator T can be estimated by $\hat{T}_n$, the operator with kernel $\frac{1}{n}\sum_{i=1}^n z_i(\tau) z_i(s)$ and r_F can be estimated by $\hat{r}_n(\tau) = \frac{1}{n}\sum_{i=1}^n y_i z_i(\tau)$. Hence (5.4) becomes

$$\hat{T}_n\varphi = \hat{r}_n \tag{5.5}$$

which is equal to

$$K^*K\varphi = K^*y.$$

If one preconditions (5.5) by applying the operator $\hat{T}_n^*$, one gets the solution

$$\hat{\varphi}_n = \left(\alpha I + \hat{T}_n^*\hat{T}_n\right)\hat{T}_n^*\hat{r}_n \tag{5.6}$$

which differs from the solution (5.3). When α goes to zero at an appropriate rate of convergence (different in both cases), the solutions of (5.3) and (5.6) will be asymptotically equivalent. Actually the preconditioning by an operator in the Tikhonov regularization has for purpose to construct an operator which is positive self-adjoint. Because $\hat{T}_n = K^*K$ is already positive self-adjoint, there is no reason to precondition here. Sometimes preconditioning more than necessary has for aim to facilitate the calculations (see Ruymgaart, 2001).

Using the results of Section 4, we can establish the asymptotic normality of $\hat{\varphi}_n$ defined in (5.6).

Assuming that

A1 - u_i has mean zero and variance σ^2 and is uncorrelated with $z_i(\tau)$ for all τ

A2 - $u_i z_i(.)$ is an iid process of $L^2(\pi)$.

A3 - $E\|u_i z_i(.)\|^2 < \infty$.

we have

(i) $\left\|\hat{T}_n^2 - T^2\right\| = O\left(\frac{1}{\sqrt{n}}\right)$

(ii) $\sqrt{n}\left(\hat{T}_n\hat{r}_n - \hat{T}_n^2\varphi_0\right) \Rightarrow \mathcal{N}(0,\Sigma)$ in $L^2(\pi)$.

(i) is straightforward. (ii) follows from

$$\begin{aligned}\hat{r}_n - \hat{T}_n\varphi_0 &= \frac{1}{n}\sum_{i=1}^n y_i z_i(\tau) - \int \frac{1}{n}\sum_{i=1}^n z_i(\tau) z_i(s)\varphi_0(s)\pi(ds) \\ &= \frac{1}{n}\sum_{i=1}^n u_i z_i(\tau).\end{aligned}$$

We have $a_n = \sqrt{n}$ and $b_n = \sqrt{n}$, hence if $\beta > 0$, the condition of Proposition **??** are satisfied. Under Assumptions A1 to A3, we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n u_i z_i(\tau) \Rightarrow \mathcal{N}\left(0,\sigma^2 T\right)$$

in $L^2(\pi)$ by Theorem 2.46.Hence

$$\sqrt{n}\left(\hat{T}_n\hat{r}_n - \hat{T}_n^2\varphi_0\right) \Rightarrow \mathcal{N}\left(0,\sigma^2 T^3\right).$$

Let us rewrite Condition G in terms of the eigenvalues λ_j and eigenfunctions ϕ_j of T :

$$\frac{\left\|\left(T^2+\alpha_n I\right)^{-1} g\right\|^2}{\left\|T^{3/2}\left(T^2+\alpha_n I\right)^{-1} g\right\|^2}=O(1)$$

$$\Leftrightarrow \frac{\sum_{j=1}^{\infty} \frac{\left\langle g,\phi_j\right\rangle^2}{\left(\lambda_j^2+\alpha\right)^2}}{\sum_{j=1}^{\infty} \frac{\lambda_j^3\left\langle g,\phi_j\right\rangle^2}{\left(\lambda_j^2+\alpha\right)^2}}=O(1).$$

Obviously the condition G will not be satisfied for all g in $L^2(\pi)$.

By Proposition 4.2., assuming that $\varphi_0 \in \Phi_\beta, 0<\beta<2$ and $\sqrt{n}\alpha_n^\beta \to 0$, we have for g conformable with Condition G,

$$\frac{\left\langle \sqrt{n}\left(\hat{\varphi}_n-\varphi_0\right), g\right\rangle}{\left\|T^{3/2}\left(T^2+\alpha_n I\right)^{-1} g\right\|} \xrightarrow{d} \mathcal{N}(0,1).$$

The asymptotic variance is given by

$$\left\|T^{-1/2} g\right\|^2=\sum_{j=1}^{\infty} \frac{\left\langle g,\phi_j\right\rangle^2}{\lambda_j}.$$

Whenever it is finite, that is whenever $g \in \mathcal{R}\left(T^{-1/2}\right)$, $\left\langle\left(\hat{\varphi}_n-\varphi_0\right), g\right\rangle$ converges at the parametric rate.

We can relate Model (5.2) to the regression with a very large number of regressors considered by Stock and Watson (1998) and Forni and Reichlin (1998). Their objective is to forecast a single time series using many explanatory variables. To do so, they extract common factors to the dependent variable y and the independent variables Z. These factors are estimated by the first l eigenvectors of the matrix $Z'Z$. A forecast of y is obtained by regressing y on these eigenvectors. This approach consists in applying a spectral cut-off regularization scheme instead of the Tikhonov regularization discussed above.

5.3. Deconvolution

This example is detailed in Carrasco and Florens (2002a). Assume we observe iid realizations $y_1, ..., y_n$ of a random variable Y with unknown pdf h, where Y satisfies

$$Y=X+Z$$

where X and Z are independent random variables with pdf φ and g respectively. The aim is to give an estimator of φ assuming g is known. This problem consists in solving in φ the equation:

$$h(y)=\int g(y-x)\varphi(x)\,dx. \tag{5.7}$$

(5.7) is an integral equation of the first kind where the operator K defined by $(K\varphi)(y) = \int g(y-x)\varphi(x)\,dx$ has a known kernel and need not be estimated. Note that K is not a compact operator with respect to Lebesgue measure and hence has a continuous spectrum. The most common approach to solving (5.7) is to use a kernel estimator, this method was pioneered by Carroll and Hall (1988) and Stefanski and Carroll (1990). It is essentially equivalent to inverting Equation (5.7) by means of the continuous spectrum of K, see Carroll, Van Rooij, and Ruymgaart (1991). In a related paper, Van Rooij and Ruymgaart (1991) propose a regularized inverse to a convolution problem of the type (5.7) where g has for support the circle. They invert the operator K using its continuous spectrum. Our approach is different. Instead of working with respect to Lebesgue measure, we define two spaces of reference, $L^2_{\pi_X}(\mathbb{R})$ and $L^2_{\pi_Y}(\mathbb{R})$, as:

$$\begin{aligned} L^2_{\pi_X}(\mathbb{R}) &= \left\{\phi(x) \text{ such that } \int \phi(x)^2 \pi_X(x)\,dx < \infty\right\}, \\ L^2_{\pi_Y}(\mathbb{R}) &= \left\{\psi(y) \text{ such that } \int \psi(y)^2 \pi_Y(y)\,dy < \infty\right\}. \end{aligned}$$

We choose π_X and π_Y so that K is a Hilbert-Schmidt operator from $L^2_{\pi_X}(\mathbb{R})$ to $L^2_{\pi_Y}(\mathbb{R})$, that is the following condition is satisfied

$$\int\int \left(\frac{\pi_Y(y)\, g(y-x)}{\pi_Y(y)\,\pi_X(x)}\right)^2 \pi_Y(y)\,\pi_X(x)\,dxdy < \infty.$$

As a result K has a discrete spectrum for these spaces of reference. Let $\{\lambda_j, \phi_j, \psi_j\}$ denote its singular value decomposition. Equation (5.7) can be approximated by a well-posed problem using Tikhonov regularization

$$(\alpha_n I + K^*K)\,\varphi_{\alpha_n} = K^*h.$$

Hence we have

$$\begin{aligned} \varphi_{\alpha_n}(x) &= \sum_{j=1}^{\infty} \frac{1}{\alpha_n + \lambda_j^2} \langle K^*h, \phi_j\rangle \phi_j(x) \\ &= \sum_{j=1}^{\infty} \frac{1}{\alpha_n + \lambda_j^2} \langle h, K\phi_j\rangle \phi_j(x) \\ &= \sum_{j=1}^{\infty} \frac{\lambda_j}{\alpha_n + \lambda_j^2} \langle h, \psi_j\rangle \phi_j(x) \\ &= \sum_{j=1}^{\infty} \frac{\lambda_j}{\alpha_n + \lambda_j^2} E\left[\psi_j(Y_i)\,\pi_Y(Y_i)\right] \phi_j(x). \end{aligned}$$

The estimator of φ is obtained by replacing the expectation by a sample mean:

$$\hat{\varphi}_n = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{\infty} \frac{\lambda_j}{\alpha_n + \lambda_j^2} \psi_j(y_i)\,\pi_Y(y_i)\,\phi_j(x).$$

Note that we avoided estimating h by a kernel estimator. In some cases, ψ_j and ϕ_j are known. For instance, if $Z \sim \mathcal{N}(0, \sigma^2)$, $\pi_Y(y) = \phi(y/\tau)$ and $\pi_X(x) = \phi\left(x/\sqrt{\tau^2+\sigma^2}\right)$ then ψ_j and ϕ_j are Hermite polynomials associated with $\lambda_j = \rho^j$. When ψ_j and ϕ_j are unknown, they can be estimated via simulations. As one can do as many simulations as one wishes, the error due to the estimation of ψ_j and ϕ_j can be considered negligeable.

Using the results of Section 3, one can establish the rate of convergence of $\|\hat{\varphi}_n - \varphi_0\|$. Assume that $\varphi_0 \in \Phi_\beta$, $0 < \beta < 2$, that is

$$\sum_{j=1}^{\infty} \frac{\langle \varphi, \phi_j \rangle^2}{\lambda_j^{2\beta}} < \infty.$$

We have $\left\|\varphi_{\alpha_n} - \varphi_0\right\| = O\left(\alpha_n^{\beta/2}\right)$ and $\left\|\hat{\varphi}_n - \varphi_{\alpha_n}\right\| = O\left(1/\left(\alpha_n\sqrt{n}\right)\right)$ as here $b_n = \sqrt{n}$. For an optimal choice of $\alpha_n = Cn^{-1/(\beta+2)}$, $\|\hat{\varphi}_n - \varphi_0\|^2$ is $O\left(n^{-\beta/(\beta+2)}\right)$. The mean integrated square error (MISE) defined as $E\|\hat{\varphi}_n - \varphi_0\|^2$ has the same rate of convergence. Fan (1993) provides the optimal rate of convergence for a minimax criterion on a Lipschitz class of functions. The optimal rate of the MISE when the error term is normally distributed is only $(\ln n)^{-2}$ when φ is twice differentiable. On the contrary, here we get an arithmetic rate of convergence. The condition $\varphi_0 \in \Phi_\beta$ has for effect to reduce the class of admissible functions and hence to improve the rate of convergence. Which type of restriction does $\varphi_0 \in \Phi_\beta$ impose? In Carrasco and Florens (2002a), it is shown that $\varphi_0 \in \Phi_1$ is satisfied if

$$\int \left| \frac{\psi_{\varphi_0}(t)}{\psi_g(t)} \right| dt < \infty \tag{5.8}$$

where ψ_{φ_0} and ψ_g denote the characteristic functions of φ_0 and g respectively. This condition can be interpreted as the noise is "smaller" than the signal. Consider for example the case where φ_0 and g are normal. Condition 5.8 is equivalent to the fact that the variance of g is smaller than that of φ_0. Note that the condition $\varphi_0 \in \Phi_1$ relates φ_0 and g while one usually imposes restrictions on φ_0 independently of those on g.

5.4. Instrumental variables

This example is detailed in Darolles, Florens and Renault (2002).

An economic relationship between a response variable Y and a vector Z of explanatory variables is often represented by an equation:

$$Y = \varphi(Z) + U, \tag{5.9}$$

where the function $\varphi(.)$ should define the relationship of interest while U is an error term. The relationship (5.9) does not characterize the function φ if the residual term is not constrained. This difficulty is solved if it is assumed that $E[U \mid Z] = 0$, or if equivalently $\varphi(Z) = E[Y \mid Z]$. However in numerous structural econometric models, the conditional expectation function is not the parameter of interest. The structural parameter is a

relation between Y and Z where some of the Z components are endogenous. This is the case in various situations: simultaneous equations, error-in-variables models, treatment model with endogenous selection etc.

The first question is to add assumptions to Equation (5.9) in order to characterize φ. Two general strategies exist in the literature, at least for linear models. The first one consists in introducing some hypotheses on the joint distribution of U and Z (for example on the variance matrix). The second one consists in increasing the vector of observables from (Y, Z) to (Y, Z, W), where W designates instrumental variables. The first approach was essentially followed in the error-in-variables models and some similarities exist with the instrumental model analysis (see e.g. Malinvaud (1970, ch. 9), Florens, Mouchart, Richard (1974) or Florens, Mouchart, Richard (1987) for the linear case). Instrumental variable analysis was proposed by Reiersol (1941), Reiersol (1945) and extended by Theil (1953), Basmann (1957) and Sargan (1958).

However, even in the instrumental variables framework, a definition of the functional parameter of interest remains ambiguous in the general nonlinear case. Three possible definitions of φ have been proposed (see Florens, Heckman, Meghir and Vytlacil (2002) for a general comparison between these three concepts and their extensions to more general treatment models).

i) The first one replaces $E[U \mid Z] = 0$ by $E[U \mid W] = 0$, or equivalently it defines φ as the solution of

$$E[Y - \varphi(Z) \mid W] = 0. \tag{5.10}$$

This definition was the foundation of the analysis of simultaneity in linear models or parametric nonlinear models (see Amemiya (1974)), but its extension to the nonparametric case raises new difficulties. The focus of this subsection is to show how to address this issue in the framework of ill-posed inverse problems (see for previous attempts, Newey and Powell (2000), quoted in Pagan and Ullah (1999));

ii) A second approach is now called *control function approach* and was systematized by Newey, Powell, and Vella (1999). This technique was previously developed in specific models (e.g. Mills ratio correction in some selection models for example). The starting point is to compute $E[Y \mid Z, W]$ which satisfies:

$$E[Y \mid Z, W] = \varphi(Z) + h(Z, W), \tag{5.11}$$

where $h(Z, W) = E[U \mid Z, W]$. Equation (5.11) does not characterize φ. However we can assume that there exists a function V (the *control function*) of (Z, W) (typically $Z - E[Z \mid W]$) which captures all the endogeneity of Z in the sense that $E[U \mid W, V] = E[U \mid V] = \tilde{h}(V)$. This implies that (5.11) may be rewritten as

$$E[Y \mid Z, W] = \varphi(Z) + \tilde{h}(V), \tag{5.12}$$

and, under some conditions, φ may be identified from (5.12), up to an additive constant term. This model is an additive model where the V are not observed but are estimated.

These models are considered in Section 7.5.
iii) A third definition follows from the literature on treatment model (see e.g. Imbens, Angrist (1994), Heckman, Ichimura, Smith, Todd (1998) and Heckman, Vytlacil (2000)). We simplify extremely this analysis by considering Z and W as scalar. *Local instrument* is defined by $\frac{\partial E[Y|W]}{\partial W} / \frac{\partial E[Z|W]}{\partial W}$, and the function of interest φ is assumed to be characterized by the relation:

$$\frac{\frac{\partial E[Y|W]}{\partial W}}{\frac{\partial E[Z|W]}{\partial W}} = E\left[\frac{\partial \varphi}{\partial Z} \mid W\right]. \tag{5.13}$$

Let us summarize the arguments which justify Equation (5.13)

Equation (5.9) is extended to a non separable model

$$Y = \varphi(Z) + Z\varepsilon + U \tag{5.14}$$

where ε and U are two random noises.

First we assume that

$$E(U|W) = E(\varepsilon|W) = 0$$

This assumption extends the instrumental variable assumption but is not sufficient to identify the parameter of interest φ. From (5.14) we get:

$$E(Y|W=w) = \int [\varphi(z) + zr(z,w)] f_Z(z|w)\, dz$$

where $f_Z(.|.)$ denote the conditional density of Z given W and $r(z,w) = E(\varepsilon|Z=z, W=w)$. Then

$$\begin{aligned} \frac{\partial}{\partial w} E(Y|W=w) &= \int \varphi(z) \frac{\partial}{\partial w} f_Z(z|w)\, dz + \int z \frac{\partial}{\partial w} r(z,w) f_Z(z|w)\, dz \\ &+ \int zr(z,w) \frac{\partial}{\partial w} f_Z(z|w)\, dz. \end{aligned}$$

We assume that the order of integration and derivative may commute (in particular the boundary of the distribution of Z given $W=w$ does not depends on w).

Let us now introduce the assumption that $V = Z - E(Z|W)$ is independent of W. In terms of density, this assumption implies that $f_Z(z|w) = \tilde{f}(z - m(w))$ where $m(w) = E(Z|W=w)$ and $\tilde{f}$ is the density of v. Then:

$$\begin{aligned} \frac{\partial}{\partial w} E(Y|W=w) &= -\frac{\partial m(w)}{\partial w} \int \varphi(z) \frac{\partial}{\partial z} f_Z(z|w)\, dz \\ &+ \int z \frac{\partial}{\partial w} r(z,w) f_Z(z|w)\, dz \\ &- \frac{\partial m(w)}{\partial w} \int zr(z,w) \frac{\partial}{\partial z} f_Z(z|w)\, dz \end{aligned}$$

An integration by part of the first and the third integrations gives

$$\begin{aligned}\frac{\partial}{\partial w}E\left(Y|W=w\right) &= \frac{\partial m\left(w\right)}{\partial w}\int\frac{\partial}{\partial z}\varphi(z)f_{Z}\left(z|w\right)dz \\ &+ \int z\left(\frac{\partial r}{\partial w}+\frac{\partial m}{\partial w}\frac{\partial r}{\partial z}\right)f_{Z}\left(z|w\right)dz \\ &+ \frac{\partial m\left(w\right)}{\partial w}\int r\left(z,w\right)f_{Z}\left(z|w\right)dz\end{aligned}$$

The last integral is zero under $E\left(\varepsilon|w\right)=0$. Finally we need to assume that the second integral is zero. This is true in particular if there exists $\tilde{r}$ such that $r\left(z,w\right)=\tilde{r}\left(z-m\left(w\right)\right)$.

Hence, Equation (5.13) is verified.

These three concepts are identical in the linear normal case but differ in general. We concentrate our presentation on this chapter on the pure instrumental variable cases defined by equation (5.10).

For a general approach of Equation (5.10) in terms of inverse problems, we introduce the following notations:

$K:L_{F}^{2}(Z)\rightarrow L_{F}^{2}(W)\quad\varphi\rightarrow K\varphi=E[\varphi\left(Z\right)\mid W]$,

$K^{*}:L_{F}^{2}(W)\rightarrow L_{F}^{2}(Z)\quad\psi\rightarrow K^{*}\psi=E[\psi\left(W\right)\mid Z]$.

All these spaces are defined relatively to the true (unknown) DGP.

These two linear operators satisfy:

$$\left\langle \varphi\left(Z\right),\psi\left(W\right)\right\rangle =E[\varphi\left(Z\right)\psi\left(W\right)]=\left\langle K\varphi\left(W\right),\psi\left(W\right)\right\rangle _{L_{F}^{2}(W)}=\left\langle \varphi\left(Z\right),K^{*}\psi\left(Z\right)\right\rangle _{L_{F}^{2}(Z)},$$

and then K^{*} is the adjoint operator of K, and reciprocally. Using these notations, the unknown instrumental regression φ corresponds to any solution of the functional equation:

$$A(\varphi,F)=K\varphi-r=0, \tag{5.15}$$

where $r\left(W\right)=E[Y\mid W]$.

In order to illustrate this construction and the central role played by the adjoint operator K^{*}, we consider first the example where Z is discrete, namely Z is binary. In that case a function $\varphi(Z)$ is characterized by two numbers $\varphi(0)$ and $\varphi(1)$ and L_{Z}^{2} is isomorphic to $\mathbb{R}^{2}$. Equation (5.10) becomes

$$\varphi\left(0\right)\text{Prob}\left(Z=0|W=w\right)+\varphi\left(1\right)\text{Prob}\ \left(Z=1|W=w\right)=E\left(Y|W=w\right)$$

The instruments W need to take at least two values in order to identify $\varphi\left(0\right)$ and $\varphi\left(1\right)$ from this equation. In general φ is overidentified and overidentification is solved by replacing (5.15) by

$$K^{*}K\varphi=K^{*}r$$

or, in the binary case, by

$$\varphi(0)\,E\left(\text{Prob}\left(Z=0|W\right)|Z\right)+\varphi(1)\,E\left(\text{Prob}\left(Z=1|W\right)|Z\right)=E\left(E\left(Y|W\right)|Z\right).$$

In the latter case, we get two equations which have in general a unique solution.

This model can be extended by considering $Z=(Z_1,Z_2)$ where Z_1 is discrete ($Z_1\in\{0,1\}$) and Z_2 be exogenous (i.e. $W=(W_1,Z_2)$). In this extended binary model, φ is characterized by two functions $\varphi(0,z_2)$ and $\varphi(1,z_2)$ solutions of

$$\begin{aligned}\varphi(0,z_2)E\left(\text{Prob}\left(Z_1=0|W\right)|Z_1=z_1,Z_2=z_2\right)+\varphi(1,z_2)\,E\left(\text{Prob}\left(Z_1=1|W\right)|Z_1=z_1,Z_2=z_2\right)\\ =E\left(E\left(Y|W\right)|Z_1=z_1,Z_2=z_2\right)\qquad\text{for } z_1=0,1\end{aligned}$$

The properties of the estimator based on the previous equation are considered in Florens and Malavolti (2002). In this case, no regularization is needed because K^*K has a continuous inverse (since the dimension is finite in the pure binary case and K^*K is not compact in the extended binary model).

We can also illustrate our approach in the case where the Hilbert spaces are not necessarily L^2 spaces. Consider the following semi parametric case. The function φ is constrained to be an element of

$$\mathcal{X}=\left\{\varphi/\varphi=\sum_{l=1}^{L}\beta_l\varepsilon_l\right\}$$

where $(\varepsilon_l)_{l=1,...,L}$ is a vector of fixed functions in $L_F^2(Z)$. Then, $\mathcal{X}$ is a finite dimensional Hilbert space. However we keep the space $\mathcal{E}$ equal to $L_F^2(W)$.The model is then partially parametric but the relation between Z and W is treated non parametrically. In this case it can be easily shown that K^* transforms any function of $L_F^2(W)$ into its best approximation in L^2 sense by a function of $\mathcal{X}$. Indeed:

If $\psi\in L_F^2(W),\forall j\in\{1,...L\}$

$$E\left(\varepsilon_j\psi\right)=\langle K\varepsilon_j,\psi\rangle=\langle\varepsilon_j,K^*\psi\rangle.$$

Moreover, $K^*\psi\in\mathcal{X}\Longrightarrow K^*\psi=\sum_{l=1}^{L}\alpha_l\varepsilon_l$, therefore

$$\left\langle\varepsilon_j,\sum_{l=1}^{L}\alpha_l\varepsilon_l\right\rangle=E\left(\psi\varepsilon_j\right)$$

$$\Leftrightarrow\sum_{l=1}^{L}\alpha_l E\left(\varepsilon_j\varepsilon_l\right)=E\left(\psi\varepsilon_j\right).$$

The function φ defined as the solution of $K\varphi = r$ is in general overidentified but the equation $K^*K\varphi = K^*r$ has always a unique solution. The finite dimension of $\mathcal{X}$ implies that $(K^*K)^{-1}$ is a finite dimensional linear operator and is then continuous. No regularization is required.

Now we introduce an assumption which is only a regularity condition when Z and W have no element in common. However, this assumption cannot be satisfied if there are some elements in common between Z and W. Extensions to this latter case are discussed in Darolles, Florens and Renault (2002).

Assumption A.1: *The joint distribution of (Z, W) is dominated by the product of its marginal distributions, and its density is square integrable w.r.t. the product of margins.*

Assumption A.1 ensures that K and K^* are Hilbert Schmidt operators, and is a sufficient condition of the compactness of K, K^*, K K^* and K^* K. (see Lancaster (1968), Darolles, Florens, Renault (1998)) and theorem 2.34..

Under Assumption A1, the instrumental regression φ is identifiable if and only if 0 is not an eigenvalue of K^*K. Then, for sake of expositional simplicity, we consider the statistical issue of estimation of this instrumental regression φ in the i.i.d. context:

Assumption A.2: *The data* (y_i, z_i, w_i) $i = 1, \cdots n$, are *i.i.d samples of* (Y, Z, W).

We estimate the joint distribution F of (Y, Z, W) using a kernel smoothing of the empirical distribution. In the applications, the bandwidths differ, but they have all the same speed represented by the notation c_n.

For economic applications, one may be interested either by the unknown function $\varphi(Z)$ itself, or only by its moments, including covariances with some known functions. These moments may for instance be useful for testing economic statements about scale economies, elasticities of substitutions, and so on.

For such tests, one will only need the empirical counterparts of these moments and their asymptotic probability distribution. An important advantage of the instrumental variable approach is to allow to estimate the covariance between $\varphi(Z)$ and $g(Z)$ for a large class of functions. Actually the identification assumption amounts to ensure that the range $\mathcal{R}(K^*)$ is dense in $L_F^2(Z)$ and for any g in this range:

$$\exists \psi \in L_F^2(W),\ g(Z) = E[\psi(W) \mid Z],$$

and then $Cov[\varphi(Z), g(Z)] = Cov[\varphi(Z), E[\psi(W) \mid Z]] = Cov[\varphi(Z), \psi(W)] = Cov[E[\varphi(Z) \mid W], \psi(W)] = Cov[Y, \psi(W)]$, can be estimated with standard parametric techniques. For instance, if $E[g(Z)] = 0$, the empirical counterpart of $Cov[Y, \psi(W)]$, i.e.:

$$\frac{1}{n}\sum_{i=1}^{n} Y_i \psi(W_i),$$

is a root-n consistent estimator of $Cov[\varphi(Z), g(Z)]$, and:

$$\sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n} Y_i \psi(W_i) - Cov[\varphi(Z), g(Z)]\right] \xrightarrow{d} \mathcal{N}(0, Var[Y\psi(W)]),$$

where $Var[Y\psi(W)]$ will also be estimated by its sample counterpart. However in practice this analysis has very limited interest because even if g is given, ψ is not known and must be estimated by solving the integral equation $g(Z) = E[\psi(W) \mid Z]$, where the conditional distribution of W given Z is also estimated.

Therefore, the real problem of interest is to estimate $Cov[\varphi(Z), g(Z)]$, or $\langle \varphi, g \rangle$ by replacing φ by an estimator. This estimator will be constructed by solving a regularized version of the empirical counterpart of (5.15) where K and r are replaced by their estimators.. In the case of kernel smoothing, the necessity of a regularization appears obviously. Using the notation of 2.5.2, the equation

$$\hat{K}_n \varphi = \hat{r}_n$$

becomes

$$\frac{\sum_{i=1}^{n} \varphi(z_i)\, \omega\left(\frac{w - w_i}{c_n}\right)}{\sum_{i=1}^{n} \omega\left(\frac{w - w_i}{c_n}\right)} = \frac{\sum_{i=1}^{n} y_i \omega\left(\frac{w - w_i}{c_n}\right)}{\sum_{i=1}^{n} \omega\left(\frac{w - w_i}{c_n}\right)}.$$

The function φ is not identified by this equation except the value $\varphi(z_i)$ equal to y_i. This solution does not define a consistent estimate. The regularized Tikhonov solution is the solution of:

$$\alpha_n \varphi(z) + \frac{\sum_{j=1}^{n} \omega\left(\frac{z - z_j}{c_n}\right) \frac{\sum_{i=1}^{n} \varphi(z_i) \omega\left(\frac{w_j - w_i}{c_n}\right)}{\sum_{i=1}^{n} \omega\left(\frac{w_j - w_i}{c_n}\right)}}{\sum_{j=1}^{n} \omega\left(\frac{z - z_j}{c_n}\right)} = \frac{\sum_{j=1}^{n} \omega\left(\frac{z - z_j}{c_n}\right) \frac{\sum_{i=1}^{n} y_i \omega\left(\frac{w_j - w_i}{c_n}\right)}{\sum_{i=1}^{n} \omega\left(\frac{w_j - w_i}{c_n}\right)}}{\sum_{j=1}^{n} \omega\left(\frac{z - z_j}{c_n}\right)}.$$

This functional equation may be solved in two steps. First, the z variable is fixed to the values z_i and the system becomes an $n \times n$ linear system, which can be solved in order to obtain the $\varphi(z_i)$. Second, the previous expression gives a value of $\varphi(z)$ for any value of z.

If n is very large this inversion method may be difficult to apply and may be replaced by a Landweber Friedman resolution. A first expression of $\varphi(z)$ may be for instance the estimated conditional expectation $E(E(Y|W)|Z)$ and this estimator will be modified a finite number of times by the formula

$$\hat{\varphi}_{l,n} = \left(I - c\hat{K}_n^* \hat{K}_n\right) \hat{\varphi}_{l-1,n} + c\hat{K}_n^* \hat{r}_n$$

(see Section 3.3).

Even if this assumption is relatively strong in a non linear context we simplify our analysis by assuming:

Assumption A.3: The error term is homoskedastic, that is:

$$Var\left(U|W\right)=0$$

In order to check the asymptotic properties of the estimator of φ, it is necessary to study to properties of the estimators of K and of r. Under regularity conditions such as the compactness of the joint distribution support and the smoothness of the density (see Darolles et al. (2002)), estimation by boundary kernels gives the following results:

i) $\left\|\hat{K}_n^*\hat{K}_n-K^*K\right\|^2\sim O\left(\frac{1}{n(c_n)^p}+(c_n)^{2\rho}\right)$ where ρ is the order of the kernel and p the dimension of Z.

ii) $\left\|\hat{K}_n^*\hat{r}_n-\hat{K}_n^*\hat{K}_n\varphi\right\|^2\sim O\left(\frac{1}{n}+(c_n)^{2\rho}\right)$

iii) A suitable choice of c_n implies

$$\sqrt{n}\left(\hat{K}_n^*\hat{r}_n-\hat{K}_n^*\hat{K}_n\varphi\right)\Longrightarrow N\left(0,\sigma^2K^*K\right)$$

This convergence is a weak convergence in $L_F^2\left(Z\right)$ (see Section 2.4).

Using results developed Section 4 and in Darolles et al. (2002) it can be deduced that:

a) If $\alpha_n\to 0$, $\frac{c_n^{2\rho}}{\alpha_n^2}\to 0$, $\frac{1}{\alpha_n^2nc_n^\rho}\sim O\left(1\right)$ the regularized estimator $\hat{\varphi}_n$ converge in probability to φ in L^2 norm.

b) If $\varphi\in\Phi_\beta$ $\left(0<\beta\leq 2\right)$, the optimal choices of α_n and c_n are:

$$\alpha_n=k_1n^{-\frac{1}{2\beta}}$$
$$c_n=k_2n^{-\frac{1}{2\rho}}$$

and, if ρ is chosen such that $\frac{p}{2\rho}\leq\frac{\beta}{2+\beta}$, we obtain the following bound of the rate of convergence

$$\left\|\hat{\varphi}_n-\varphi\right\|\sim O\left(n^{-\frac{\beta}{2+\beta}}\right)$$

c) Let us assume that α is kept constant. In that case the linear operators $(\alpha I+K_n^*\,K_n\,)^{-1}$ and $(\alpha I+K^*K)^{-1}$ are bounded and, using a functional version of the Slutsky theorem (see Chen, White (1992) and, section (2.4)), it is immediately checked that:

$$\sqrt{n}(\hat{\varphi}_n-\varphi-b_n^\alpha)\Longrightarrow\mathcal{N}(0,\Omega),\tag{5.16}$$

where

$$b_n^\alpha=\alpha\left[(\alpha I+K_n^*K_n)^{-1}-(\alpha I+K^*K)^{-1}\right]\varphi,$$

and

$$\Omega = \sigma^2(\alpha I + K^*K)^{-1}K^*K(\alpha I + K^*K)^{-1}.$$

Some comments may illustrate this first result:

i) The convergence obtained in (5.16) is still a functional distributional convergence in the Hilbert space $L_F^2(Z)$, which in particular implies the convergence of inner product $\sqrt{n}\langle\hat{\varphi}_n - \varphi - b_n^\alpha, g\rangle$ to univariate normal distribution $\mathcal{N}(0, \langle g, \Omega g\rangle)$.

ii) The convergence of $\hat{\varphi}_n$ involves two bias terms. The first bias is $\varphi_\alpha - \varphi$. This term is due to the regularization and does not decrease if α is constant. The second one, $\hat{\varphi}_n - \varphi_\alpha$ follows from the estimation error of K. This bias decreases to zero when n increases, but at a lower speed than $\sqrt{n}$.

iii) The asymptotic variance in (5.16) can be seen as the generalization of the two stage least squares asymptotic variance. An intuitive (but not correct) interpretation of this result could be the following: if α is small, the asymptotic variance is approximately $\sigma^2(K^*K)^{-1}$, which is the functional extension of $\sigma^2(E(ZW')E(WW')^{-1}E(WZ'))^{-1}$.

d) Let us now consider the case where $\alpha \to 0$. For any $\delta \in \Phi_\beta$ $(\beta \geq 1)$, if α_n is optimal $(= k_1 n^{-\frac{1}{2\beta}})$and if $c_n = k_2 n^{-\left(\frac{1}{2\rho}+\varepsilon\right)}$ $(\varepsilon > 0)$, we have

$$\sqrt{\nu_n(\delta)}\,\langle\hat{\varphi}_n - \varphi, \delta\rangle - B_n \Longrightarrow N\left(0, \sigma^2\right)$$

where the speed of convergence is equal to

$$\nu_n(\delta) = \frac{n}{\left\|K\left(\alpha_n I + K^*K\right)^{-1}\delta\right\|^2} \geq O\left(n^{\frac{2\beta}{2+\beta}}\right)$$

and the bias B_n is equal to $\sqrt{\nu_n(\delta)}\,\langle\varphi_\alpha - \varphi, \delta\rangle$, which in general does not vanish. If $\delta = 1$ for example, this bias is $O(n\alpha_n^2)$ and diverges.

The notion of Φ_β space gives a rigorous content to the concept of weak or strong instruments. Indeed any φ is identified by Equation (5.15) if λ_j are not zero for any j and $\hat{\varphi}_n$ is then a consistent estimator. The speed of convergence may be bounded is φ is in a Φ_β space with $\beta > 0$. This means that the rate of decline of the Fourier coefficients of φ in the basis of ϕ_j is faster than the rate of decline of the λ_j^β(which measures the dependence). In order to have an asymptotic normality we need to assume that $\beta > 1$. In that case if $\varphi \in \Phi_\beta$ we have weak asymptotic normality in the vector space Φ_β. Then a natural definition of strong instruments is to assume $\beta \geq 1$. This may have two equivalent interpretations. Given Z and W, the set of instrumental regression for which W is a strong instrument is Φ_1 or given Z and φ, any set of instruments is strong if φ is an element of the set Φ_1 defined using this instruments.

We may complete this short presentation by two final remarks. First the optimal choice of c_n and α_n implies that the speed of convergence and the asymptotic distribution are not

affected by the fact that K is not known and is estimated. The accuracy of the estimation is governed by the estimation of the right hand side term K^*r. Secondly the usual "curse of dimensionality" of nonparametric estimation appears in a complex way. The dimension of Z appears in many places but the dimension of W is less explicit. The value and the rate of decline of the λ_j depend on the dimension of W: given Z, the reduction of the number of instruments reduces the λ_j and affects negatively the properties of the estimator.

6. Reproducing kernel and GMM in Hilbert spaces

6.1. Reproducing kernel

Models based on reproducing kernels are the foundation for penalized likelihood estimation and regularization methods (Wahba, 2000). However it has been little used in econometrics so far. The theory of reproducing kernels becomes very useful when the econometrician has an infinity of moment conditions and want to exploit all of them in an efficient way. For illustration, let $\theta \in \mathbb{R}$ be the parameter of interest and consider an $L \times 1-$vector h that gives L moment conditions satisfying $E^{\theta_0}(h(\theta)) = 0 \Rightarrow \theta = \theta_0$. Let $h_n(\theta)$ be the sample estimate of $E^{\theta_0}(h(\theta))$. The (optimal) generalized method of moments (GMM) estimator of θ is the minimizer of $h_n(\theta)' \Sigma^{-1} h_n(\theta)$ where Σ is the covariance matrix of h. $h_n(\theta)' \Sigma^{-1} h_n(\theta)$ can be rewritten as $\left\|\Sigma^{-1/2} h_n(\theta)\right\|^2$ and coincides with the norm of $h_n(\theta)$ in a particular space called the reproducing kernel Hilbert space (RKHS). When h is finite dimensional, the computation of the GMM objective function does not raise any particular difficulty, however when h is infinite dimensional (for instance is a function) then the theory of RKHS becomes very handy. A second motivation for the introduction of the RKHS of a self-adjoint operator K is the following. Let T be such that $K = TT^*$ then the RKHS of K corresponds to the $1-$regularity space of T (denoted Φ_1 in Section 3.1).

6.1.1. Definitions and basic properties of RKHS

This section presents the theory of reproducing kernels, as described in Aronszajn (1950) and Parzen (1959, 1970). Let $L_C^2(\pi) = \left\{\varphi : I \subset \mathbb{R}^L \to \mathbf{C} : \int_I |\varphi(s)|^2 \pi(s)\, ds < \infty\right\}$ where π is a pdf (π may have a discrete or continuous support) and denote $\|.\|$ and $\langle , \rangle$ the norm and inner product on $L_C^2(\pi)$.

Definition 6.1. *A space $\mathcal{H}(K)$ of complex-valued functions defined on a set $I \subset \mathbb{R}^L$ is said to be a reproducing kernel Hilbert space $\mathcal{H}(K)$ associated with the integral operator $K : L_C^2(\pi) \to L_C^2(\pi)$ with kernel $k(t,s)$ if the three following conditions hold*
(i) it is a Hilbert space (with inner product denoted $\langle , \rangle_K$),
(ii) for every $s \in I$, $k(t,s)$ as a function of t belongs to $\mathcal{H}(K)$,
(iii) (reproducing property) for every $s \in I$ and $\varphi \in \mathcal{H}(K)$, $\varphi(s) = \langle \varphi(.), k(.,s) \rangle_K$.
The kernel k is then called reproducing kernel.

The following properties are listed in Aronszajn (1950):

1 - If the RK k exists, it is unique.

2 - A Hilbert space $\mathcal{H}$ of functions defined on $I \subset \mathbb{R}^L$ is a RKHS if and only if all functionals $\varphi \to \varphi(s)$ for all $\varphi \in \mathcal{H}$, $s \in I$, are bounded.

3 - K is a self-adjoint positive operator on $L_C^2(\pi)$.

4 - To a self-adjoint positive operator K on I, there corresponds a unique RKHS $\mathcal{H}(K)$ of complex-valued functions.

5 - Every sequence of functions $\{\varphi_n\}$ which converges weakly to φ in $\mathcal{H}(K)$ (that is $\langle \varphi_n, g \rangle \to \langle \varphi, g \rangle$ for all $g \in \mathcal{H}(K)$) converges also pointwise, that is $\lim \varphi_n(s) = \varphi(s)$.

Note that (2) is a consequence of Riesz theorem 2.18: there exists a representor k such that for all $\varphi \in \mathcal{H}$

$$\varphi(t) = \langle \varphi, k_t \rangle_K .$$

Let $k_t = k(t, .)$ so that $\langle k_t, k_s \rangle_K = k(t, s)$. (5) follows from the reproducing property. Indeed, $\langle \varphi_n(t) - \varphi(t), k(t, s) \rangle_K = \varphi_n(s) - \varphi(s)$.

Example (finite dimensional case). Let $I = \{1, 2, ..., L\}$, let Σ be a positive definite $L \times L$ matrix with principal element $\sigma_{t,s}$. Σ defines an inner product on $\mathbb{R}^L$: $\langle \varphi, \psi \rangle_\Sigma = \varphi' \Sigma^{-1} \psi$. Let $(\sigma_1, ..., \sigma_L)$ be the columns of Σ. Let $\varphi = (\varphi(1), ..., \varphi(L))'$, then we have the reproducing property

$$\langle \varphi, \sigma_t \rangle_\Sigma = \varphi(t), \tau = 1, ..., L$$

because $\varphi \Sigma \Sigma^{-1} = \varphi$. Now we diagonalize Σ, $\Sigma = PDP'$ where P is the $m \times m$ matrix with (t, j) element $\phi_j(t)$ (ϕ_j are the orthonormal eigenvectors of Σ) and D is the diagonal matrix with diagonal element λ_j (the eigenvalues of Σ). The (t, s)th element of Σ can be rewritten as

$$\sigma(t, s) = \sum_{j=1}^{m} \lambda_j \phi_j(t) \phi_j(s) .$$

We have

$$\langle \varphi, \psi \rangle_\Sigma = \varphi' \Sigma^{-1} \psi = \sum_{j=1}^{m} \frac{1}{\lambda_j} \langle \varphi, \phi_j \rangle \langle \psi, \phi_j \rangle$$

where $\langle , \rangle$ is the euclidean inner product.

From this small example, we see that the norm in a RKHS can be characterized by the spectral decomposition of an operator. Let K be a positive self-adjoint Hilbert Schmidt operator with spectrum $\{\phi_j, \lambda_j : j = 1, 2, ...\}$. By contrast to Section 3.1, we do not assume that $\mathcal{N}(K) = 0$. In order to write series involving $1/\lambda_j$, we use the convention

$1/0 = 0$. It turns out that $\mathcal{H}(K)$ coincides with the 1/2-regularization space of the operator K :

$$\mathcal{H}(K) = \left\{ \varphi : \varphi \in L^2(\pi) \text{ and } \sum_{j=1}^{\infty} \frac{\left|\langle \varphi, \phi_j \rangle\right|^2}{\lambda_j} < \infty \right\} = \Phi_{1/2}(K).$$

We can check that

(i) $\mathcal{H}(K)$ is a Hilbert space with inner product

$$\langle \varphi, \psi \rangle_K = \sum_{j=1}^{\infty} \frac{\langle \varphi, \phi_j \rangle \overline{\langle \psi, \phi_j \rangle}}{\lambda_j}$$

and norm

$$\|\varphi\|_K^2 = \sum_{j=1}^{\infty} \frac{\left|\langle \varphi, \phi_j \rangle\right|^2}{\lambda_j}.$$

(ii) $k(.,t)$ belongs to $\mathcal{H}(K)$

(iii) $\langle \varphi, k(.,t) \rangle_K = \varphi(t)$.

Proof. (ii) follows from Mercer's formula (Theorem 2.42 (iii)) that is $k(t,s) = \sum_{j=1}^{\infty} \lambda_j \phi_j(t) \overline{\phi_j(s)}$. Hence $\|k(.,t)\|_K^2 = \sum_{j=1}^{\infty} \left|\langle \phi_j, k(.,t) \rangle\right|^2 / \lambda_j = \sum_{j=1}^{\infty} \left|\lambda_j \phi_j(t)\right|^2 / \lambda_j = \sum_{j=1}^{\infty} \lambda_j \phi_j(t) \overline{\phi_j(t)} = k(t,t) < \infty$. For (iii), we use again Mercer's formula. $\langle \varphi(.), k(.,t) \rangle_K = \sum_{j=1}^{\infty} \langle \phi_j, k(.,t) \rangle \langle \varphi, \phi_j \rangle / \lambda_j = \sum_{j=1}^{\infty} \langle \varphi, \phi_j \rangle K\phi_j(t) / \lambda_j = \sum_{j=1}^{\infty} \langle \varphi, \phi_j \rangle \phi_j(t) = \varphi(t)$. ■

There is a link between calculating a norm in a RKHS and solving an integral equation $K\varphi = \psi$. We follow Nashed and Wahba (1974) to enlighten this link. We have

$$K\varphi = \sum_{j=1}^{\infty} \lambda_j \langle \varphi, \phi_j \rangle \phi_j.$$

Define $K^{1/2}$ the square root of K:

$$K^{1/2}\varphi = \sum_{j=1}^{\infty} \sqrt{\lambda_j} \langle \varphi, \phi_j \rangle \phi_j.$$

Note that $\mathcal{N}(K) = \mathcal{N}\left(K^{1/2}\right)$, $\mathcal{H}(K) = K^{1/2}\left(L_C^2(\pi)\right)$. Define $K^{-1/2} = \left(K^{1/2}\right)^{-1}$ where $()^{-1}$ is a generalized inverse. First we explain what we mean by generalized inverse. Consider solving in φ the integral equation $K\varphi = \psi$. The study of the properties of this equation has been the object of Section 3.1. We know that the solution will exist only if $\psi \in \mathcal{R}(K) \oplus \mathcal{R}(K)^{\perp} = \Psi_1 \oplus \mathcal{N}(K)$. When the solution exists, it is not unique because for any solution φ, one can find another solution $\varphi + \varphi_0$ where $\varphi_0 \in \mathcal{N}(K)$. In Section 3.1, we just assumed $\mathcal{N}(K) = 0$. Another way to solve the non uniqueness issue is by

considering only the solution with minimal variance, this solution is called generalized inverse and takes the form:

$$K^{-1}\psi = \sum_{j=1}^{\infty} \frac{1}{\lambda_j} \langle \psi, \phi_j \rangle \phi_j.$$

Similarly, the generalized inverse of $K^{1/2}$ takes the form:

$$K^{-1/2}\psi = \sum_{j=1}^{\infty} \frac{1}{\sqrt{\lambda_j}} \langle \psi, \phi_j \rangle \phi_j.$$

From Nashed and Wahba (1974), we have the relations

$$\begin{aligned} \|\varphi\|_K^2 &= \inf \left\{ \|p\| : p \in L_C^2(\pi) \text{ and } \varphi = K^{1/2}p \right\}, \\ \langle \varphi, \psi \rangle_K &= \left\langle K^{-1/2}\varphi, K^{-1/2}\psi \right\rangle, \text{ for all } \varphi, \psi \in \mathcal{H}(K). \end{aligned} \tag{6.1}$$

The following result follows from Proposition 3.4.

Proposition 6.2. *Let $T : \mathcal{E} \to L_C^2(\pi)$ be an operator such that $K = TT^*$ then*

$$\mathcal{H}(K) = \mathcal{R}\left(K^{1/2}\right) = \mathcal{R}(T^*) = \Phi_1(T).$$

Note that $T^* : L_C^2(\pi) \to \mathcal{E}$ and $K^{1/2} : L_C^2(\pi) \to L_C^2(\pi)$ are not equal because they take their values in different spaces.

6.1.2. RKHS for covariance operators of stochastic processes

In the previous section, we have seen how to characterize $\mathcal{H}(K)$ using the spectral decomposition of K. When K is known to be the covariance kernel of a stochastic process, then $\mathcal{H}(K)$ admits a simple representation. The main results of this section come from Parzen (1959). Consider a random element (r.e.). $\{h(t), t \in I \subset \mathbb{R}^p\}$ defined on a probability space $(\Omega, \mathcal{F}, P)$ and observed for all values of t. Assume $h(t)$ is a second order random function that is $E\left(|h(t)|^2\right) = \int_\Omega |h(t)|^2 dP < \infty$ for every $t \in I$. Let $L_2(\Omega, \mathcal{F}, P)$ be the set of all r.v. U such that $E|U|^2 = \int_\Omega |U|^2 dP < \infty$. Define the inner product $\langle U, V \rangle_{L_2(\Omega,\mathcal{F},P)}$ between any two r.v. U and V of $L_2(\Omega, \mathcal{F}, P)$ by $\langle U, V \rangle_{L_2(\Omega,\mathcal{F},P)} = E\left(U\overline{V}\right) = \int_\Omega UV dP$. Let $L_2(h(t), t \in I)$ be the Hilbert space spanned by the r.e. $\{h(t), t \in I\}$. Define K the covariance operator with kernel $k(t,s) = E\left(h(t)\overline{h(s)}\right)$. The following theorem implies that any symmetric nonnegative kernel can be written as a covariance kernel of a particular process.

Theorem 6.3. *K is a covariance operator of a r.e. if and only if K is a positive self-adjoint operator.*

The following theorem can be found in Parzen (1959) for real-valued functions. The complex case is treated in Saitoh (1997).

Theorem 6.4. *Let $\{h(t), t \in I\}$ be a r.e. with mean zero and covariance kernel k. Then*

(i) $L_2(h(t), t \in I)$ is isometrically isomorphic or congruent to the RKHS $\mathcal{H}(K)$. Denote J this congruence.

(ii) For every function φ in $\mathcal{H}(K)$, $J(\varphi)$ satisfies

$$\langle J(\varphi), h(t)\rangle_{L_2(\Omega,\mathcal{F},P)} = E\left(J(\varphi)\overline{h(t)}\right) = \langle \varphi, k(.,t)\rangle_K = \varphi(t), \text{ for all } t \in I \tag{6.2}$$

where $J(\varphi)$ is unique in $L_2(h(t), t \in I)$ and has mean zero and variance such that

$$\|\varphi\|_K^2 = \|J(\varphi)\|_{L_2(\Omega,\mathcal{F},P)}^2 = E\left(|J(\varphi)|^2\right).$$

Note that, by (6.2), the congruence is such that $J(k(.,t)) = h(t)$. The r.v. $U \in L_2(h(t), t \in I)$ corresponding to $\varphi \in \mathcal{H}(K)$ is denoted below as $\langle \varphi, h\rangle_K$ (or $J(\varphi)$). As $L_2(h(t), t \in I)$ and $\mathcal{H}(K)$ are isometric, we have by Definition 2.19

$$cov\left[\langle \varphi, h\rangle_K, \langle \psi, h\rangle_K\right] = E\left[J(\varphi)\overline{J(\psi)}\right] = \langle \varphi, \psi\rangle_K$$

for every φ, $\psi \in \mathcal{H}(K)$. Note that $\langle \varphi, h\rangle_K$ is not a correct notation because $h = \sum_j \langle h, \phi_j\rangle \phi_j$ a.s. does not belong to $\mathcal{H}(K)$. If it were the case, we should have $\sum_j \langle h, \phi_j\rangle^2 / \lambda_j < \infty$ a.s.. Unfortunately $\langle h, \phi_j\rangle$ are independent with mean 0 and variance $\langle K\phi_j, \phi_j\rangle = \lambda_j$. Hence, $E\left[\sum_j \langle h, \phi_j\rangle^2 / \lambda_j\right] = \infty$ and by Kolmogorov's theorem $\sum_j \langle h, \phi_j\rangle^2 / \lambda_j = \infty$ with nonzero probability. The r.v. $J(\varphi)$ itself is well-defined, only the notation $\langle \varphi, h\rangle_K$ is not adequate; as Kailath (1971) explains, it should be regarded only as a mnemonic for finding $J(\varphi)$ in a closed form. The rest of this section is devoted on the calculation of $\|\varphi\|_K$. Note that the result (6.2) is valid when t is multidimensional, $t \in \mathbb{R}^L$. In the next section, $h(t)$ will be a moment function indexed by an arbitrary index parameter t.

Assume that the kernel k on $I \times I$ can be represented as

$$k(s,t) = \int h(s,x)\overline{h(t,x)}P(dx) \tag{6.3}$$

where P is a probability measure and $\{h(s,.), s \in I\}$ is a family of functions on $L_2(\Omega, \mathcal{F}, P)$. By Theorem 6.4, $\mathcal{H}(K)$ consists of functions φ on I of the form

$$\varphi(t) = \int \psi(x)\overline{h(t,x)}P(dx) \tag{6.4}$$

for some unique ψ in $L_2(h(t,.), t \in I)$, the subspace of $L_2(\Omega, \mathcal{F}, P)$ spanned by $\{h(t,.), t \in I\}$. The RKHS norm of φ is given by

$$\|\varphi\|_K^2 = \|\psi\|_{L_2(\Omega,\mathcal{F},P)}^2.$$

When calculating $\|\varphi\|_K^2$ in practice, one looks for the solutions of (6.4). If there are several, it is not always obvious to see which one is spanned by $\{h(t,.), t \in I\}$. However, the right solution ψ is the solution with minimum norm (Parzen, 1970).

Theorem 6.4 can be reinterpreted in terms of range. Let T and T^* be

$$\begin{aligned} T &: L^2(\pi) \to L_2(h(t,.), t \in I) \\ \varphi &\to T\varphi(x) = \int \varphi(t) h(t,x) \pi(t) dt. \end{aligned}$$

and

$$\begin{aligned} T^* &: L_2(h(t,.), t \in I) \to L^2(\pi) \\ \psi &\to T^*\psi(s) = \int \psi(x) h(s,x) P(dx). \end{aligned}$$

To check that T^* is indeed the dual of T, it suffices to check $\langle T\varphi, \psi \rangle_{L_2(\Omega,\mathcal{F},P)} = \langle \varphi, T^*\psi \rangle_{L^2(\pi)}$ for $\varphi \in L^2(\pi)$ and $\psi(x) = h(t,x)$ as $h(t,.)$ spans $L_2(h(t,.), t \in I)$. Using the fact that $K = T^*T$ and Proposition 6.2, we have $\mathcal{H}(K) = \mathcal{R}(T^*)$, which gives Equation (6.4).

Example. Let $k(t,s) = t \wedge s$. k can be rewritten as

$$k(t,s) = \int_0^1 (t-x)_+^0 (s-x)_+^0 du$$

with

$$(s-x)_+^0 = \begin{cases} 1 \text{ if } x < s \\ 0 \text{ if } x \geq s \end{cases}.$$

It follows that $\mathcal{H}(K)$ consists of functions φ of the form:

$$\begin{aligned} \varphi(t) &= \int_0^1 \psi(x) (t-x)_+^0 dx = \int_0^t \psi(x) dx, \ 0 \leq t \leq 1 \\ &\Rightarrow \psi(t) = \varphi'(t). \end{aligned}$$

Hence, we have

$$\|\varphi\|_K^2 = \int_0^1 |\psi(x)|^2 dx = \int_0^1 |\varphi'(x)|^2 dx.$$

Example. Let k be defined as in (6.3) with $h(t,x) = e^{itx}$. Assume P admits for pdf $f_{\theta_0}(x)$ positive everywhere. To compute Λ one needs to solve

$$\begin{aligned} \varphi(t) &= \int \psi(x) e^{-itx} P(dx) \\ &= \int \psi(x) e^{-itx} f_{\theta_0}(x) dx. \end{aligned}$$

By the Fourier Inversion formula, we have

$$\psi(x) = \frac{1}{2\pi} \frac{1}{f_{\theta_0}(x)} \int e^{itx} \varphi(t) dt.$$

6.2. GMM in Hilbert spaces

First introduced by Hansen (1982), the Generalized Method of Moments (GMM) became the cornerstone of modern structural econometrics. In Hansen, the number of moment conditions is supposed to be finite. The method proposed in this section permits to deal with moment functions that take their values in a finite or infinite dimensional Hilbert space. It was initially proposed by Carrasco and Florens (2000) and further developed in Carrasco and Florens (2001) and Carrasco, Chernov, Florens, and Ghysels (2001).

6.2.1. Definition and examples

Let $\{x_i : i = 1, 2, ..., n\}$ be an iid sample of a random vector $X \in \mathbb{R}^p$. The case where X is a time-series will be discussed later. The distribution of X is indexed by a parameter $\theta \in \Theta \subset \mathbb{R}^d$. Denote E^θ the expectation with respect to this distribution. The unknown parameter θ is identified from the function $h(X;\theta)$ (called moment function) defined on $\mathbb{R}^p \times \Theta$, so that the following is true.

Identification Assumption

$$E^{\theta_0}(h(X;\theta)) = 0 \Rightarrow \theta = \theta_0. \tag{6.5}$$

It is assumed that $h(X;\theta)$ takes its values in a Hilbert space $\mathcal{H}$ with inner product $\langle .,. \rangle$ and norm $\|.\|$. When $f = (f_1, ..., f_L)$ and $g = (g_1, ..., g_L)$ are vectors of functions of $\mathcal{H}$, we use the convention that $\langle f, g' \rangle$ denotes the $L \times L-$matrix with (l, m) element $\langle f_l, g_m \rangle$. Let $B_n : \mathcal{H} \to \mathcal{H}$ be a sequence of random bounded linear operators and

$$\hat{h}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} h(x_i;\theta).$$

We define the GMM estimator associated with B_n as

$$\hat{\theta}_n(B_n) = \arg\min_{\theta \in \Theta} \left\| B_n \hat{h}_n(\theta) \right\|. \tag{6.6}$$

Such an estimator will be in general suboptimal; we will discuss the optimal choice of B_n later. Below, we give four examples that can be handled by the method discussed in this section. They illustrate the versatility of the method as it can deal with a finite number of moments (Example 1), a continuum (Examples 2 and 3) and a countable infinite sequence (Example 4).

Example 1 (Traditional GMM). Let $h(x;\theta)$ be a vector of $\mathbb{R}^L$, B_n be a $L \times L-$matrix and $\|.\|$ denote the Euclidean norm. The objective function to minimize is

$$\left\| B_n \hat{h}_n(\theta) \right\|^2 = \hat{h}_n(\theta)' B_n' B_n \hat{h}_n(\theta)$$

and corresponds to the usual GMM quadratic form $\hat{h}_n(\theta)' W_n \hat{h}_n(\theta)$ with weighting matrix $W_n = B_n' B_n$.

Example 2 (Continuous time process). Suppose we observe independent replications of a continuous time process

$$X^i(t) = G(\theta, t) + u^i(t),\ 0 \leq t \leq T,\ i = 1, 2, ..., n \tag{6.7}$$

where G is a known function and $u^i = \{u^i(t) : 0 \leq t \leq T\}$ is a zero mean Gaussian process with continuous covariance function $k(t,s) = E[u(t)\,u(s)]$, $t, s \in [0,T]$. Denote $X^i = \{X^i(t) : 0 \leq t \leq T\}$, $G(\theta) = \{G(\theta, t) : 0 \leq t \leq T\}$, and $\mathcal{H} = L^2([0,T])$. The unknown parameter θ is identified from the moment of the function

$$h\left(X^i; \theta\right) = X^i - G(\theta).$$

Assume $h(X^i; \theta) \in L^2([0,T])$ with probability one. Candidates for B_n are arbitrary bounded operators on $L^2([0,T])$ including the identity. For $B_n f = f$, we have

$$\left\| B_n \hat{h}_n(\theta) \right\|^2 = \int_0^T \hat{h}_n(\theta)^2\, dt.$$

Estimation of Model (6.7) is discussed in Kutoyants (1984).

Example 3 (Characteristic function). Denote $\psi_\theta(t) = E^\theta\left[e^{it'X}\right]$ the characteristic function of X. Inference can be based on

$$h(t, X; \theta) = e^{it'X} - \psi_\theta(t),\ t \in \mathbb{R}^L.$$

Note that contrary to the former examples, $h(t, X; \theta)$ is complex valued and $|h(t, X; \theta)| \leq \left|e^{it'X}\right| + |\psi_\theta(t)| \leq 2$. Let Π be a probability measure on $\mathbb{R}^L$ and $\mathcal{H} = L^2_C\left(\mathbb{R}^L, \Pi\right)$. As $h(., X; \theta)$ is bounded, it belongs to $L^2_C\left(\mathbb{R}^L, \Pi\right)$ for any Π. Feuerverger and McDunnough (1981) and more recently Singleton (2001) show that an efficient estimator of θ is obtained from $h(., X; \theta)$ by solving an empirical counterpart of $\int Eh(t, X; \theta)\,\omega(t)\,dt = 0$ for an adequate weighting function ω which turns out to be a function of the pdf of X. This efficient estimator is not implementable as the pdf of X is unknown. They suggest estimating θ by GMM using moments obtained from a discrete grid $t = t_1, t_2, ..., t_L$. An alternative strategy put forward in this section is to use the full continuum of moment conditions by considering a moment function h as an element of $L^2_C\left(\mathbb{R}^L, \Pi\right)$.

Example 4 (Conditional moment restrictions). Let $X = (Y, Z)$. For a known function $\rho \in \mathbb{R}$, we have the conditional moment restrictions

$$E^{\theta_0}[\rho(Y, Z, \theta)\,|Z] = 0.$$

Hence for any function $g(Z)$, we can construct unconditional moment restrictions

$$E^{\theta_0}[\rho(Y, Z, \theta)\, g(Z)] = 0.$$

Chamberlain (1987) shows that the semiparametric efficiency bound can be approached by a GMM estimator based on a sequence of moment conditions using as instruments

the power function of $Z: 1, Z, Z^2, ..., Z^L$ for a large L. Let π be the Poisson probability measure $\pi(l) = e^{-1}/l!$ and $\mathcal{H} = L^2(\mathbf{N},\pi) = \{f: \mathbf{N} \to \mathbb{R}: \sum_{l=1}^{\infty} g(l)\pi(l) < \infty\}$. Let

$$h(l, X; \theta) = \rho(Y, Z, \theta) Z^l,\ l = 1, 2, ...$$

If $h(l, X; \theta)$ is bounded with probability one, then $h(., X; \theta) \in L^2(\mathbf{N},\pi)$ with probability one. Instead of using an increasing sequence of moments as suggested by Chamberlain, it is possible to handle $h(., X; \theta)$ as a function. The efficiency of the GMM estimator based on the countable infinity of moments $\{h(l, X; \theta) : l \in \mathbf{N}\}$ will be discussed later.

6.2.2. Asymptotic properties

Let $\mathcal{H} = L_C^2(I, \Pi) = \left\{f: I \to \mathbf{C}: \int_I |f(t)|^2 \Pi(dt) < \infty\right\}$ where I is a subset of $\mathbb{R}^L$ for some $L \geq 1$ and Π is a probability measure. This choice of $\mathcal{H}$ is consistent with Examples 1 to 4. Under some weak assumptions, $\sqrt{n}\hat{h}_n(\theta_0)$ converges to a Gaussian process $\mathcal{N}(0, K)$ in $\mathcal{H}$ where K denotes the covariance operator of $h(X; \theta_0)$. K is defined by

$$\begin{aligned} K &: \mathcal{H} \to \mathcal{H} \\ f &\to Kf(s) = \langle f, k(., t)\rangle = \int_I k(t, s) f(s) \Pi(ds) \end{aligned}$$

where the kernel k of K satisfies $k(t, s) = E^{\theta_0}\left[h(t, X; \theta_0)\overline{h(s, X; \theta_0)}\right]$ and $k(t, s) = \overline{k(s, t)}$. Assume moreover that K is a Hilbert Schmidt operator and hence admits a discrete spectrum. Suppose that B_n converges to a bounded linear operator B defined on H and that θ_0 is the unique minimizer of $\left\|BE^{\theta_0}h(X; \theta)\right\|$ then $\hat{\theta}_n(B_n)$ is consistent and asymptotically normal. The following result is proved in Carrasco and Florens (2000).

Proposition 6.5. *Under Assumptions 1 to 11 of Carrasco and Florens (2000), we have*

$$\sqrt{n}\left(\hat{\theta}_n(B_n) - \theta_0\right) \xrightarrow{L} \mathcal{N}(0, V)$$

with

$$\begin{aligned} V &= \left\langle BE^{\theta_0}(\nabla_\theta h), BE^{\theta_0}(\nabla_\theta h)'\right\rangle^{-1} \\ &\quad \times \left\langle BE^{\theta_0}(\nabla_\theta h), (BKB^*) BE^{\theta_0}(\nabla_\theta h)'\right\rangle \\ &\quad \times \left\langle BE^{\theta_0}(\nabla_\theta h), BE^{\theta_0}(\nabla_\theta h)'\right\rangle^{-1} \end{aligned}$$

where B^* is the adjoint of B.

6.2.3. Optimal choice of the weighting operator

Carrasco and Florens (2000) show that the asymptotic variance V given in Proposition 6.5 is minimal for $B = K^{-1/2}$. In that case, the asymptotic covariance becomes $\left\langle K^{-1/2}E^{\theta_0}(\nabla_\theta h), K^{-1/2}E^{\theta_0}(\nabla_\theta h)\right\rangle^{-1}$.

Example 1 (continued). K is the $L \times L-$covariance matrix of $h(X;\theta)$. Let K_n be the matrix $\frac{1}{n}\sum_{i=1}^{n} h\left(x_i;\hat{\theta}^1\right) h\left(x_i;\hat{\theta}^1\right)'$ where $\hat{\theta}^1$ is a consistent first step estimator of θ. K_n is a consistent estimator of K. Then the objective function becomes

$$\left\langle K_n^{-1/2}\hat{h}_n(\theta), K_n^{-1/2}\hat{h}_n(\theta)\right\rangle = \hat{h}_n(\theta)' K_n^{-1}\hat{h}_n(\theta)$$

which delivers the optimal GMM estimator.

When $\mathcal{H}$ is infinite dimensional, we have seen in Section 3.1 that the inverse of K, K^{-1}, is not bounded. Similarly $K^{-1/2} = \left(K^{1/2}\right)^{-1}$ is not bounded on $\mathcal{H}$ and its domain has been shown in Subsection 6.1.1 to be the subset of $\mathcal{H}$ which coincides with the RKHS associated with K and denoted $\mathcal{H}(K)$.

To estimate the covariance operator K, we need a first step estimator $\hat{\theta}^1$ that is $\sqrt{n}-$consistent. It may be obtained by letting B_n equal the identity in (6.6). Let K_n be the operator with kernel

$$k_n(t,s) = \frac{1}{n}\sum_{i=1}^{n} h\left(t,x_i;\hat{\theta}^1\right)\overline{h\left(s,x_i;\hat{\theta}^1\right)}.$$

Then K_n is a consistent estimator of K and $\|K_n - K\| = O\left(1/\sqrt{n}\right)$. As $K^{-1}f$ is not continuous in f, we estimate K^{-1} by the Tykhonov regularized inverse of K_n :

$$\left(K_n^{\alpha_n}\right)^{-1} = \left(\alpha_n I + K_n^2\right)^{-1} K_n$$

for some penalization term $\alpha_n \geq 0$. If $\alpha_n > 0$, $\left(K_n^{\alpha_n}\right)^{-1} f$ is continuous in f but is a biased estimator of $K^{-1}f$. There is a trade-off between the stability of the solution and its bias. Hence, we will let α_n decrease to zero at an appropriate rate. We define $\left(K_n^{\alpha_n}\right)^{-1/2} = \left(\left(K_n^{\alpha_n}\right)^{-1}\right)^{1/2}$.

The optimal GMM estimator is given by

$$\hat{\theta}_n = \arg\min_{\theta\in\Theta} \left\|\left(K_n^{\alpha_n}\right)^{-1/2}\hat{h}_n(\theta)\right\|.$$

Interestingly, the optimal GMM estimator minimizes the norm of $\hat{h}_n(\theta)$ in the RKHS associated with $K_n^{\alpha_n}$. Under certain regularity conditions, we have

$$\left\|\left(K_n^{\alpha_n}\right)^{-1/2}\hat{h}_n(\theta)\right\| \xrightarrow{P} \left\|E^{\theta_0}\left(h(\theta)\right)\right\|_K.$$

A condition for applying this method is that $E^{\theta_0}(h(\theta)) \in \mathcal{H}(K)$. This condition can be verified using results from 6.1.1.

Proposition 6.6. *Under the regularity conditions of Carrasco and Florens (2000, Theorem 8), we have*

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{L} \mathcal{N}\left(0, \left\langle E^{\theta_0}\left(\nabla_\theta h(\theta_0)\right), E^{\theta_0}\left(\nabla_\theta h(\theta_0)\right)'\right\rangle_K^{-1}\right)$$

as n and $n\alpha_n^{5/2} \to \infty$ and $\alpha_n \to 0$.

The stronger condition $n\alpha_n^3 \to \infty$ of Carrasco and Florens (2000) has been relaxed into $n\alpha^{5/2} \to \infty$ in Carrasco, Chernov, Florens, and Ghysels (2001). Proposition 6.6 does not indicate how to select α_n in practice. A data-driven method is desirable. Carrasco and Florens (2001) propose to select the α_n that minimizes the mean square error (MSE) of the GMM estimator $\hat{\theta}_n$. As $\hat{\theta}_n$ is consistent for any value of α_n, it is necessary to compute a higher order MSE.

6.2.4. Implementation

There are two equivalent ways to compute the objective function

$$\left\| (K_n^{\alpha_n})^{-1/2} \hat{h}_n(\theta) \right\|^2, \tag{6.8}$$

1) using the spectral decomposition of K_n,or

2) using a simplified formula that involves only vectors and matrices.

The first method discussed in Carrasco and Florens (2000) requires calculating the eigenvalues and eigenfunctions of K_n using the method described in 2.5.2. Let $\hat{\phi}_j$ denote the orthonormalized eigenfunctions of K_n and $\hat{\lambda}_j$ the corresponding eigenvalues. The objective function in Equation (6.8) becomes

$$\sum_{j=1}^{n} \frac{\hat{\lambda}_j}{\hat{\lambda}_j^2 + \alpha_n} \left| \left\langle \hat{h}_n(\theta), \hat{\phi}_j \right\rangle \right|^2.$$

The second method is more attractive by its simplicity. Carrasco and al. (2001) show that (6.8) can be rewritten as

$$\overline{\underline{v}(\theta)}' \left[I_n - C \left[\alpha_n I_n + C^2 \right] C \right] \underline{v}(\theta)$$

where C is a $n \times n-$matrix with (i,j) element c_{ij}, I_n is the $n \times n$ identity matrix and $\underline{v}(\theta) = (v_1(\theta), ..., v_n(\theta))'$ with

$$\begin{aligned} v_i(\theta) &= \int \overline{h\left(t, x_i; \hat{\theta}^1\right)}' \hat{h}_n(t;\theta) \Pi(dt) \\ c_{ij} &= \frac{1}{n} \int \overline{h\left(t, x_i; \hat{\theta}^1\right)}' h\left(t, x_j; \hat{\theta}^1\right) \Pi(dt). \end{aligned}$$

Note that the dimension of C is the same whether $h \in \mathbb{R}$ or $h \in \mathbb{R}^L$.

6.2.5. Asymptotic Efficiency

Assume that the pdf of X, f_θ, is differentiable with respect to θ. Let $L^2(h)$ be the closure of the subspace of $L^2(\Omega, \mathcal{F}, P)$ spanned by $\{h(t, X_i; \theta_0) : t \in I\}$.

Proposition 6.7. *Under certain regularity conditions, the GMM estimator based on* $\{h(t, x_i; \theta) : t \in I\}$ *is asymptotically as efficient as the MLE if and only if*

$$\nabla_\theta \ln f_\theta(x_i; \theta_0) \in L^2(h).$$

This result is proved in Carrasco and Florens (2002) in a more general setting where X_i is Markov of order L. A similar efficiency result can be found in Hansen (1985), Tauchen (1997) and Gallant and Long (1997).

Example 2 (continued). Let K be the covariance operator of $\{u(t)\}$ and $\mathcal{H}(K)$ the RKHS associated with K. Kutoyants (1984) shows that if $G(\theta) \in \mathcal{H}(K)$, the likelihood ratio of the measure induced by $X(t)$ with respect to the measure induced by $u(t)$ equals

$$LR(\theta) = \prod_{i=1}^{n} \exp\left\{\left\langle G(\theta), x^i\right\rangle_K - \frac{1}{2}\|G(\theta)\|_K^2\right\}$$

where $\langle G, X\rangle_K$ has been defined in Subsection 6.1.2 and denotes the element of $L^2(X(t) : 0 \leq t \leq T)$ under the mapping J^{-1} of the function $G(\theta)$ (J is defined in Theorem 6.4). The score function with respect to θ is

$$\nabla_\theta \ln(LR(\theta)) = \left\langle \nabla_\theta G(\theta), \frac{1}{n}\sum_{i=1}^{n}\left(x^i - G(\theta)\right)\right\rangle_K.$$

For $\theta = \theta_0$ and a single observation, the score is equal to

$$\langle \nabla_\theta G(\theta_0), u\rangle_K$$

which is an element of $L^2(u(t) : 0 \leq t \leq T) = L^2(h(X(t); \theta_0) : 0 \leq t \leq T)$. Hence, by Proposition 6.7, the GMM estimator based on $h(X; \theta_0)$ is asymptotically efficient. This efficiency result is corroborated by the following. The GMM objective function is

$$\|h(x; \theta)\|_K^2 = \left\langle \frac{1}{n}\sum_{i=1}^{n}\left(x^i - G(\theta)\right), \frac{1}{n}\sum_{i=1}^{n}\left(x^i - G(\theta)\right)\right\rangle_K.$$

The first order derivative equals to

$$\begin{aligned}\nabla_\theta \|h(x; \theta)\|_K^2 &= 2\left\langle \nabla_\theta G(\theta), \frac{1}{n}\sum_{i=1}^{n}\left(x^i - G(\theta)\right)\right\rangle_K \\ &= 2\nabla_\theta \ln(LR(\theta)).\end{aligned}$$

Therefore, the GMM estimator coincides with the MLE in this particular case as they are solutions of the same equation.

Example 3 (continued). Under minor conditions on the distribution of X_i, the closure of the linear span of $\left\{h(t, X_i; \theta_0) : t \in \mathbb{R}^L\right\}$ contains all functions of $L^2(X) =$

$\left\{g : E^{\theta_0}\left[g\left(X\right)^2\right] < \infty\right\}$ and hence the score $\nabla_\theta \ln f_\theta\left(X_i;\theta_0\right)$ itself. Therefore the GMM estimator is efficient. Another way to prove efficiency is to calculate explicitly the asymptotic covariance of $\hat{\theta}_n$. To simplify, assume that θ is scalar. By Theorem 6.4, we have

$$\left\|E^{\theta_0}\left(\nabla_\theta h\left(\theta_0\right)\right)\right\|_K^2 = \left\|\overline{E^{\theta_0}\left(\nabla_\theta h\left(\theta_0\right)\right)}\right\|_K^2 = E\left|U\right|^2$$

where U satisfies

$$E^{\theta_0}\left[U\overline{h\left(t;\theta_0\right)}\right] = \overline{E^{\theta_0}\left(\nabla_\theta h\left(t;\theta_0\right)\right)} \text{ for all } t \in \mathbb{R}^L$$

which is equivalent to

$$E^{\theta_0}\left[\overline{U\left(X\right)}\left(e^{it'X} - \psi_{\theta_0}\left(t\right)\right)\right] = -\nabla_\theta \psi_{\theta_0}\left(t\right) \text{ for all } t \in \mathbb{R}^L. \tag{6.9}$$

As U has mean zero, $\overline{U}$ has also mean zero and we can replace (6.9) by

$$\begin{aligned} E^{\theta_0}\left[\overline{U\left(X\right)}e^{it'X}\right] &= -\nabla_\theta \psi_{\theta_0}\left(t\right) \text{ for all } t \in \mathbb{R}^L \Leftrightarrow \\ \int \overline{U\left(x\right)}e^{it'x} f_{\theta_0}\left(x\right) dx &= -\nabla_\theta \psi_{\theta_0}\left(t\right) \text{ for all } t \in \mathbb{R}^L \Leftrightarrow \\ \overline{U\left(x\right)} f_{\theta_0}\left(x\right) &= -\frac{1}{2\pi}\int e^{-it'x}\nabla_\theta \psi_{\theta_0}\left(t\right) dt. \end{aligned} \tag{6.10}$$

The last equivalence follows from the Fourier inversion formula. Assuming that we can exchange the integration and derivation in the right hand side of (6.10), we obtain

$$\begin{aligned} \overline{U\left(x\right)} f_{\theta_0}\left(x\right) &= -\nabla_\theta f_{\theta_0}\left(x\right) \Leftrightarrow \\ U\left(x\right) &= -\nabla_\theta \ln f_{\theta_0}\left(x\right). \end{aligned}$$

Hence $E^{\theta_0}\left|U\right|^2 = E^{\theta_0}\left[\left(\nabla_\theta \ln f_{\theta_0}\left(X\right)\right)^2\right]$. And the asymptotic variance of $\hat{\theta}_n$ coincides with the Cramer Rao efficiency bound even if, contrary to Example 3, $\hat{\theta}_n$ differs from the MLE.

Example 4 (continued). As in the previous example, we intend to calculate the asymptotic covariance of $\hat{\theta}_n$ using Theorem 6.4. We need to find U the $p-$vector of r.v. such that

$$\begin{aligned} E^{\theta_0}\left[U\rho\left(Y,Z;\theta_0\right)Z^l\right] &= E^{\theta_0}\left[\nabla_\theta \rho\left(Y,Z;\theta_0\right)Z^l\right] \text{ for all } l \in \mathbf{N}, \Leftrightarrow \\ E^{\theta_0}\left[E^{\theta_0}\left[U\rho\left(Y,Z;\theta_0\right)\left|Z\right.\right]Z^l\right] &= E^{\theta_0}\left[E^{\theta_0}\left[\nabla_\theta \rho\left(Y,Z;\theta_0\right)\left|Z\right.\right]Z^l\right] \text{ for all } l \in \mathbf{N} \end{aligned} \tag{6.11}$$

(6.11) is equivalent to

$$E^{\theta_0}\left[U\rho\left(Y,Z;\theta_0\right)\left|Z\right.\right] = E^{\theta_0}\left[\nabla_\theta \rho\left(Y,Z;\theta_0\right)\left|Z\right.\right]. \tag{6.12}$$

by the completeness of polynomials (Sansone, 1959) under some mild conditions on the distribution of Y. A solution is

$$U_0 = E^{\theta_0}\left[\nabla_\theta \rho\left(Y,Z;\theta_0\right)\left|Z\right.\right] E^{\theta_0}\left[\rho\left(Y,Z;\theta_0\right)^2\left|Z\right.\right]^{-1} \rho\left(Y,Z;\theta_0\right).$$

We have to check that this solution has minimal norm among all the solutions. Consider an arbitrary solution $U = U_0 + U_1$. U solution of (6.12) implies

$$E^{\theta_0}\left[U_1 \rho\left(Y, Z; \theta_0\right) | Z\right] = 0.$$

Hence $E^{\theta_0}\left(UU'\right) = E^{\theta_0}\left(U_0 U_0'\right) + E^{\theta_0}\left(U_1 U_1'\right)$ and is minimal for $U_1 = 0$. Then

$$\begin{aligned} & \left\| E^{\theta_0}\left(\nabla_\theta h\left(\theta_0\right)\right)\right\|_K^2 \\ = \; & E^{\theta_0}\left(U_0 U_0'\right) \\ = \; & E^{\theta_0}\left\{E^{\theta_0}\left[\nabla_\theta \rho\left(Y, Z; \theta_0\right) | Z\right] E^{\theta_0}\left[\rho\left(Y, Z; \theta_0\right)^2 | Z\right]^{-1} E^{\theta_0}\left[\nabla_\theta \rho\left(Y, Z; \theta_0\right) | Z\right]'\right\}. \end{aligned}$$

Its inverse coincides with the semi-parametric efficiency bound derived by Chamberlain (1987).

Note that in Examples 2 and 3, the GMM estimator reaches the Cramer Rao efficiency bound asymptotically while in Example 4, it reaches the semi-parametric efficiency bound.

6.2.6. Extension to time series

So far, the data were assumed to be iid. Now we relax this assumption. Let $\{x_1, ..., x_T\}$ be the observations of a time series $\{X_t\}$ that satisfies some mixing conditions. Inference will be based on moment functions $\{h\left(\tau, X_t; \theta_0\right)\}$ indexed by a real, possibly multidimensional, index τ. $\{h\left(\tau, X_t; \theta_0\right)\}$ are in general autocorrelated, except in some special cases, an example of which will be discussed below.

Example 5 (Conditional characteristic function). Let Y_t be a (scalar) Markov process and assume that the conditional characteristic function (CF) of Y_{t+1} given Y_t, $\psi_\theta\left(\tau | Y_t\right) \equiv E^\theta\left[\exp\left(i\tau Y_{t+1}\right) | Y_t\right]$, is known. The following conditional moment condition holds

$$E^\theta\left[e^{i\tau Y_{t+1}} - \psi_\theta\left(\tau | Y_t\right) | Y_t\right] = 0.$$

Denote $X_t = \left(Y_t, Y_{t+1}\right)'$. Let $g\left(Y_t\right)$ be an instrument so that

$$h\left(\tau, X_t; \theta\right) = \left(e^{i\tau Y_{t+1}} - \psi_\theta\left(\tau | Y_t\right)\right) g\left(Y_t\right)$$

satisfies the identification condition (6.5). $\{h\left(\tau, X_t; \theta\right)\}$ is a martingale difference sequence and is therefore uncorrelated. The use of the conditional CF is very popular in finance. Assume that $\{Y_t,\ t = 1, 2, ..., T\}$ is a discretely sampled diffusion process, then Y_t is Markov. While the conditional likelihood of Y_{t+1} given Y_t does not have a closed form expression, the conditional CF of affine diffusions is known. Hence GMM can replace MLE to estimate these models where MLE is difficult to implement. For an adequate choice of the instrument $g\left(Y_t\right)$, the GMM estimator is asymptotically as efficient as the MLE. The conditional CF has been recently applied to the estimation of diffusions by Singleton (2001), Chacko and Viceira (2001), and Carrasco et al. (2001). The first three

papers use GMM based on a finite grid of values for τ, whereas the last paper advocates using the full continuum of moments which permits to achieve efficiency asymptotically.

Example 6 (Joint characteristic function). Assume Y_t is not Markov. In that case, the conditional CF is usually unknown. On the other hand, the joint characteristic function may be calculated explicitly (for instance when Y_t is an ARMA process with stable error, see Knight and Yu, 2002; or Y_t is the growth rate of a stochastic volatility model, see Jiang and Knight, 2000) or may be estimated via simulations (this technique is developed in Carrasco et al., 2001). Denote $\psi_\theta(\tau) \equiv E^\theta[\exp(\tau_1 Y_t + \tau_2 Y_{t+1} + ... + \tau_{L+1} Y_{t+L})]$ with $\tau = (\tau_1, ..., \tau_L)'$, the joint CF of $X_t \equiv (Y_t, Y_{t+1}, ..., Y_{t+L})'$ for some integer $L \geq 1$. Assume that L is large enough for

$$h(\tau, X_t; \theta) = e^{i\tau' X_t} - \psi_\theta(\tau)$$

to identify the parameter θ. Here $\{h(\tau, X_t; \theta)\}$ are autocorrelated. Knight and Yu estimate various models by minimizing the following norm of $h(\tau, X_t; \theta)$:

$$\int \left(\frac{1}{T} \sum_{t=1}^{T} e^{i\tau' x_t} - \psi_\theta(\tau) \right)^2 e^{-\tau'\tau} d\tau.$$

This is equivalent to minimizing $\left\| B \frac{1}{T} \sum_{t=1}^{T} h(\tau, X_t; \theta) \right\|^2$ with $B = e^{-\tau'\tau/2}$. This choice of B is suboptimal but has the advantage to be easy to implement. The optimal weighting operator is, as before, the square root of the inverse of the covariance operator. Its estimation will be discussed shortly.

Under some mixing conditions on $\{h(\tau, X_t; \theta_0)\}$, the process $\hat{h}_T(\theta_0) = \frac{1}{T} \sum_{t=1}^{T} h(\tau, X_t; \theta_0)$ follows a functional CLT (see Subsection 2.4.2):

$$\sqrt{T} \hat{h}_T(\theta_0) \xrightarrow{L} \mathcal{N}(0, K)$$

where the covariance operator K is an integral operator with kernel

$$k(\tau_1, \tau_2) = \sum_{j=-\infty}^{+\infty} E^{\theta_0} \left[h(\tau_1, X_t; \theta_0) \overline{h(\tau_2, X_{t-j}; \theta_0)} \right].$$

k can be estimated using a kernel-based estimator as those described in Andrews (1991) and references therein. Let $\omega : \mathbb{R} \to [-1, 1]$ be a kernel satisfying the conditions stated by Andrews. Let q be the largest value in $[0, +\infty)$ for which

$$\omega_q = \lim_{u \to \infty} \frac{1 - \omega(u)}{|u|^q}$$

is finite. In the sequel, we will say that ω is a $q-$kernel. Typically, $q = 1$ for the Bartlett kernel and $q = 2$ for Parzen, Tuckey-Hanning and quadratic spectral kernels. We define

$$\hat{k}_T(\tau_1, \tau_2) = \frac{T}{T - d} \sum_{j=-T+1}^{T-1} \omega\left(\frac{j}{S_T} \right) \hat{\Gamma}_T(j) \tag{6.13}$$

with

$$\hat{\Gamma}_T(j) = \begin{cases} \frac{1}{T}\sum_{t=j+1}^{T} h\left(\tau_1, X_t; \hat{\theta}_T^1\right) \overline{h\left(\tau_2, X_{t-j}; \hat{\theta}_T^1\right)}, & j \geq 0 \\ \frac{1}{T}\sum_{t=-j+1}^{T} h\left(\tau_1, X_{t+j}; \hat{\theta}_T^1\right) \overline{h\left(\tau_2, X_t; \hat{\theta}_T^1\right)}, & j < 0 \end{cases} \tag{6.14}$$

where S_T is some bandwidth that diverges with T and $\hat{\theta}_T^1$ is a $T^{1/2}-$consistent estimator of θ. Let K_T be the integral estimator with kernel $\hat{k}_T$. Under some conditions on ω and $\{h(\tau, X_t; \theta_0)\}$, Carrasco et al. (2001) establish the rate of convergence of K_T to K. Assuming $S_T^{2q+1}/T \to \gamma \in (0, +\infty)$, we have

$$\|K_T - K\| = O_p\left(T^{-q/(2q+1)}\right).$$

The inverse of K is estimated using the regularized inverse of K_T, $K_T^{\alpha_T} = \left(K_T^2 + \alpha_T I\right)^{-1} K_T$ for a penalization term $\alpha_T \geq 0$. As before, the optimal GMM estimator is given by

$$\hat{\theta}_T = \arg\min_{\theta \in \Theta} \left\| \left(K_T^{\alpha_T}\right)^{-1/2} \hat{h}_T(\theta) \right\|.$$

Carrasco et al. (2001) show the following result.

Proposition 6.8. *Assume that ω is a $q-$kernel and that $S_T^{2q+1}/T \to \gamma \in (0, +\infty)$. We have*

$$\sqrt{T}(\hat{\theta}_T - \theta_0) \xrightarrow{L} \mathcal{N}\left(0, \left(\left\langle E^{\theta_0}(\nabla_\theta h), E^{\theta_0}(\nabla_\theta h)' \right\rangle_K\right)^{-1}\right) \tag{6.15}$$

as T and $T^{q/(2q+1)}\alpha_T^{5/4}$ go to infinity and α_T goes to zero.

Note that the implementation of this method requires two smoothing parameters α_T and S_T. No cross-validation method for selecting these two parameters simultaneously has been derived yet. If $\{h_t\}$ is uncorrelated, then K can be estimated using the sample average and the resulting estimator satisfies $\|K_T - K\| = O_p\left(T^{-1/2}\right)$. When $\{h_t\}$ are correlated, the convergence rate of K_T is slower and accordingly the rate of convergence of α_T to zero is slower.

7. Estimation of the solution of an integral equation of the second kind

7.1. Introduction

The objective of this section is to study the properties of the solution of an integral equation of the second kind (also called Fredholm equation of the second type) defined by:

$$(I - K)\varphi = r \tag{7.1}$$

where φ is an element of an Hilbert space $\mathcal{H}$, K is a compact operator from $\mathcal{H}$ to $\mathcal{H}$ and r is an element of $\mathcal{H}$. As in the previous sections, K and r are known functions of a data generating process characterized by its c.d.f. F and the functional parameter of interest is the function φ.

In most cases, $\mathcal{H}$ is a functional space and K is an integral operator defined by its kernel k and Equation (7.1) becomes:

$$\varphi(t) - \int k(t,s)\varphi(s)\Pi(ds) = r(t) \tag{7.2}$$

The estimated operators are often degenerated, see Subsection 2.6.2. and, in that case, Equation (7.2) simplifies into:

$$\varphi(t) - \sum_{\ell=1}^{L} a_\ell(\varphi)\varepsilon_\ell(t) = r(t) \tag{7.3}$$

where the $a_\ell(\varphi)$ are linear forms on $\mathcal{H}$ and ε_ℓ belongs to $\mathcal{H}$ for any ℓ.

The essential difference between equations of the first kind and of the second kind is the compactness of the operator. In (7.1), K is compact but $I - K$ is not compact. Moreover, if $I - K$ is one-to-one, its inverse is bounded. In that case, the inverse problem is well-posed. Even if $I - K$ is not one-to-one the ill-poseness of equation (7.1) is less severe than in the first kind case because the solutions are stable in r.

In most cases, K is a self-adjoint operator (and hence $I-K$ is also self-adjoint) but we will not restrict our presentation to this case. On the other hand, Equation (7.1) could be extended by considering an equation $(S - K)\varphi = r$ where K is now a compact operator from $\mathcal{H}$ to $\mathcal{E}$ and S is a bounded operator from $\mathcal{H}$ to $\mathcal{E}$, one-to-one with a bounded inverse. This extension will not be considered in this paper.

This section will be organized in the following way. The next paragraph recalls the main mathematical properties of the equations of the second kind. The two following paragraphs present the statistical properties of the solution in the cases of well-posed and of ill-posed problems and the last paragraph applies these results to the two examples given in Section 1.

The implementation of the estimation procedures is not discussed here because this issue is similar to the implementation of the estimation of a regularized equation of the first kind (see Section 3). Actually, regularizations transform first kind equations into second kind equations and the numerical methods are then formally equivalent, even though statistical properties are fundamentally different.

7.2. Riesz theory and Fredholm alternative

We first briefly recall the main results about equations of the second kind as they were developed at the beginning of the 20th century by Fredholm and Riesz. The statements are given without proofs (see e.g. Kress, 1999, Chapters 3 and 4).

Let K be a compact operator from $\mathcal{H}$ to $\mathcal{H}$ and I be the identity on $\mathcal{H}$ (which is compact only if $\mathcal{H}$ is finite dimensional). Then the operator $I-K$ has a finite dimensional

null space and its range is closed. Moreover $I-K$ is injective if and only if it is surjective. In that case $I-K$ is invertible and its inverse $(I-K)^{-1}$ is a bounded operator.

An element of the null space of $I-K$ verifies $K\varphi = \varphi$ and if $\varphi \neq 0$, it is an eigenfunction of K associated with the eigenvalue equal to 1. Equivalently the inverse problem (7.1) is well-posed if and only if 1 is not an eigenvalue of K. The Fredholm alternative follows from the previous results.

Theorem 7.1 (Fredholm alternative). *Let us consider the two equations of the second kind:*

$$(I-K)\varphi = r \tag{7.4}$$

and

$$(I-K^*)\psi = s \tag{7.5}$$

where K^* is the adjoint of K. Then:

i) Either the two homogeneous equations $(I-K)\varphi = 0$ and $(I-K^*)\psi = 0$ only have the trivial solutions $\varphi = 0$ and $\psi = 0$ and in that case (7.4) and (7.5) have a unique solution for any r and s in $\mathcal{H}$

ii) or the two homogeneous equations $(I-K)\varphi = 0$ and $(I-K^*)\psi = 0$ have the same finite number m of linearly independent solutions φ_j and ψ_j $(j = 1, ..., m)$ respectively and the solutions of (7.4) and (7.5) exist if and only if $\langle \psi_j, r \rangle = 0$ and $\langle \varphi_j, s \rangle = 0$ for any $j = 1, ..., m$.

7.3. Statistical properties of the solution of a well-posed equation of the second kind

In the case of a one to one equation of the second kind, the asymptotic properties are easily deduced from the properties of the estimation of the operator K and of the right-hand side r.

The starting point of this analysis is the relation:

$$\begin{aligned}
\hat{\varphi}_n - \varphi_0 &= \left(I - \hat{K}_n\right)^{-1} \hat{r}_n - (I-K)^{-1} r \\
&= \left(I - \hat{K}_n\right)^{-1} (\hat{r}_n - r) + \left[\left(I - \hat{K}_n\right)^{-1} - (I-K)^{-1}\right] r \\
&= \left(I - \hat{K}_n\right)^{-1} \left[\hat{r}_n - r + \left(\hat{K}_n - K\right)(I-K)^{-1} r\right] \\
&= \left(I - \hat{K}_n\right)^{-1} \left[\hat{r}_n - r + \left(\hat{K}_n - K\right)\varphi_0\right]
\end{aligned} \tag{7.6}$$

where the third equality follows from $A^{-1} - B^{-1} = A^{-1}(B-A)B^{-1}$.

Theorem 7.2. ??*If*

i) $\left\|\hat{K}_n - K\right\| = o(1)$

ii) $\left\|\left(\hat{r}_n + \hat{K}_n\varphi_0\right) - (r + K\varphi_0)\right\| = O\left(\frac{1}{a_n}\right)$

Then $\|\hat{\varphi}_n - \varphi_0\| = O\left(\frac{1}{a_n}\right)$

Proof. As $I - K$ is invertible and admits a continuous inverse, i) implies that $\| \left(I - \hat{K}_n\right)^{-1} \|$ converges to $\left\|(I-K)^{-1}\right\|$ and the result follows from (7.6). ■

In some cases $\|r - \hat{r}_n\| = O(\frac{1}{b_n})$ and $\|\hat{K}_n - K\| = O(\frac{1}{d_n})$. Then $\frac{1}{a_n} = \frac{1}{b_n} + \frac{1}{d_n}$. In some particular examples, as it will be illustrated in the last subsection the asymptotic behavior of $\hat{r}_n - \hat{K}_n\varphi$ is directly considered.

Asymptotic normality can be obtained from different sets of assumptions. The following theorems illustrate two kinds of asymptotic normality.

Theorem 7.3. *If*

i) $\left\|\hat{K}_n - K\right\| = o(1)$

ii) $a_n\left(\left(\hat{r}_n + \hat{K}_n\varphi_0\right) - (r + K\varphi_0)\right) \Longrightarrow N(0, \Sigma)$ *(weak convergence in* $\mathcal{H}$*)*

Then

$$a_n\left(\hat{\varphi}_n - \varphi_0\right) \Longrightarrow \mathcal{N}\left(0, (I-K)^{-1}\Sigma\left(I - K^*\right)^{-1}\right)$$

Proof. The proof follows immediately from (7.6) and Theorem 2.47 in Section 2. ■

Theorem 7.4. *We consider the case where* $\mathcal{H} = L^2(\mathbb{R}^p, \pi)$*. If*

i) $\|\hat{K}_n - K\| = o(1)$

ii) $\exists a_n \; s.t \; a_n\left[\left(\hat{r}_n + \hat{K}_n\varphi_0\right) - (r + K\varphi_0)\right](x) \xrightarrow{d} N\left(0, \sigma^2(x)\right), \quad \forall x \in \mathbb{R}^p$

iii) $\exists b_n \; s.t \; \frac{a_n}{b_n} = o(1)$ *and*

$$b_n\hat{K}\left[\left(\hat{r}_n + \hat{K}_n\varphi\right) - (r + K\varphi_0)\right] \Longrightarrow N(0, \Omega)$$

(weak convergence in $\mathcal{H}$*)*

Then

$$a_n\left(\hat{\varphi}_n - \varphi_0\right)(x) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2(x)\right) \qquad \forall x$$

Proof. Using

$$(I-K)^{-1} = I + (I-K)^{-1}K$$

we deduce from (7.6):

$$\begin{aligned} a_n(\hat{\varphi}_n - \varphi_0)(x) &= a_n\left\{(I-\hat{K}_n)^{-1}\left[\hat{r}_n + \hat{K}\varphi_0 - r - K\varphi_0\right]\right\} \\ &= a_n(\hat{r} + \hat{K}\varphi_0 - r - K\varphi_0)(x) \\ &+ \frac{a_n}{b_n}\left\{b_n(I-\hat{K})^{-1}\hat{K}(\hat{r} + \hat{K}\varphi_0 - r - K\varphi_0)\right\}(x) \end{aligned} \tag{7.7}$$

The last term into bracket converges (weakly in L^2). to a $N(0,(I-K)^{-1}\Omega(I-K)^{-1})$ and the value of this function at any point x also converges to a normal distribution (weak convergence implies finite dimensional convergences). Then the last term into brackets is bounded and the result is verified. ■

Note that condition (iii) is justified by circumstances when K is an integral operator which increases the rate of convergence of $\hat{r}_n + \hat{K}_n\varphi$.

We illustrate these results by the following three examples even if the first one appears to be a little artificial.

Example. Consider $L^2(\mathbb{R},\pi)$ and (Y,Z) is a random element of $\mathbb{R}\times L^2(\mathbb{R},\pi)$. We study the integral equation of the second kind defined by

$$\varphi(x) + \int E^F\left(Z(x)Z(y)\right)\varphi(y)\pi\left(dy\right) = E^F(YZ(x)) \tag{7.8}$$

denoted by $\varphi + V\varphi = r$.

This equation defines a well posed inverse problem because the covariance operator is positive. We assume that an i.i.d. sample of (Y,Z) is available and the estimated equation (7.8) defines the parameter of interest as the solution of an integral equation having the following form:

$$\varphi(x) + \frac{1}{n}\sum_{i=1}^{n} z_i(x)\int z_i(y)\varphi(y)\pi\left(dy\right) = \frac{1}{n}\sum_{i=1}^{n} y_i z_i(x) \tag{7.9}$$

Under regularity conditions one can check that $\|\hat{V}_n - V\| = O\left(\frac{1}{\sqrt{n}}\right)$ and that

$$\sqrt{n}\frac{1}{n}\sum_i\left\{z_i(\cdot)\left[y_i - \int z_i(y)\varphi(y)\pi(dy)\right] - E^F\left(YZ\left(\cdot\right)\right) - \int E^F(Z(.)Z(y))\varphi(y)\pi(dy)\right\}$$

$\Rightarrow$ $N(0,\Sigma)$ in $L^2(\mathbb{R},\pi)$.

If for instance $E^F(Y|Z) = \int Z(y)\varphi(y)\pi(y)dy$ and under a homoscedasticity hypothesis the operator Σ is a covariance operator with kernel $\sigma^2 E^F(Z(x)Z(y))$ where

$$\sigma^2 = Var\left(Y - \int Z(y)\varphi(y)\pi(dy)|Z\right).$$

Then, from Theorem 7.3,

$$\sqrt{n}\left(\hat{\varphi}_n - \varphi_0\right) \Rightarrow N\left(0, \sigma^2(I+V)^{-1}V(I+V)^{-1}\right) \tag{7.10}$$

Example. **Rational expectations asset pricing models**:

Following Lucas (1978), such models characterize the pricing functional as a function φ of the Markov state solution of an integral equation:

$$\varphi(x) - \int a(x,y)\varphi(y)\, f(y|x)\, dy = \int a(x,y)b(y)f(y|x)\, dy \tag{7.11}$$

While f is the transition density of the Markov state, the function a denotes the marginal rate of substitution and b the dividend function. For sake of expositional simplicity, we assume here that the functions a and b are both known while f is estimated nonparametrically by a kernel method. Note that if the marginal rate of substitution a involves some unknown preference parameters (subjective discount factor, risk aversion parameter), they will be estimated, for instance by GMM, with a parametric root n rate of convergence. Therefore, the nonparametric inference about φ (deduced by solution of (7.11) of a kernel estimation of f) is not contaminated by this parametric estimation; all the statistical asymptotic theory can be derived as if the preference parameters were known.

As far as kernel density estimation is concerned, it is well known that under mild conditions (see e.g. Bosq (1998)) it is possible to get with stationary strongly mixing stochastic processes the same convergence rates and the same asymptotic distribution as in the i.i.d. case. Then, we do not make explicit in this presentation the assumed dynamic properties of the observations y and x of present and lagged values of a Markov process.

Let us then consider a n-dimensional stationary stochastic process X_t and $\mathcal{H}$ the space of square integrable functions of one realization of this process. In this example, $\mathcal{H}$ is defined with respect to the true distribution. The operator K is defined by

$$K\varphi(x) = E^F\left(a\left(X_{t-1}, X_t\right)\varphi\left(X_t\right)|X_{t-1} = x\right) \tag{7.12}$$

and

$$r(x) = E^F\left(a\left(X_{t-1}, X_t\right)b(X_t)|X_{t-1} = x\right) \tag{7.13}$$

We will assume that K is compact though for example a Hilbert-Schmidt condition (see assumption A.1 of Section 5.4 for such a condition). A common assumption in rational

expectation models is that K is a contraction mapping, due to discounting. Then, 1 is not an eigenvalue of K and (7.11) is a well-posed Fredholm integral equation.

Under these hypotheses, both numerical and statistical issues associated with the solution of (7.11) are well documented. See Rust, Traub and Wozniakowski (2002) and references therein for numerical issues. The statistical consistency of the estimator $\hat{\varphi}_n$ deduced from the kernel estimator $\hat{K}_n$ is deduced from Theorem **??** above. Assumption $i)$ is satisfied because $\hat{K}_n - K$ has the same behavior as the conditional expectation operator and

$$\begin{aligned}\hat{r}_n + \hat{K}_n\varphi - r - K\varphi \\ = E^{F_n}\left(a\left(X_{t-1}, X_t\right)\left(b(X_t) + \varphi\left(X_t\right)\right)|X_{t-1}\right) \\ -E^{F}\left(a\left(X_{t-1}, X_t\right)\left(b(X_t) + \varphi\left(X_t\right)\right)|X_{t-1}\right)\end{aligned}$$

converges at the speed $\frac{1}{a_n} = \left(\frac{1}{nc_n^m} + c_n^4\right)^{1/2}$ if c_n is the bandwidth of the (second order) kernel estimator and m is the dimension of X.

The weak convergence obtained through Theorem 7.4, Assumption ii) is the usual result on the normality of kernel estimation of conditional expectation. As K is an integral operator, the transformation by K increases the speed of convergence which implies iii).

Example: Partially Nonparametric forecasting model:

Nonparametric prediction of a stationary ergodic scalar random process X_t is often performed by looking for a predictor $\varphi\left(X_{t-1}, ..., X_{t-d}\right)$ able to minimize the mean square error of prediction:

$$E\left[X_t - \varphi\left(X_{t-1}, ..., X_{t-d}\right)\right]^2$$

In other words, if φ can be any squared integrable function, the optimal predictor is the conditional expectation

$$\varphi_0\left(X_{t-1,...,}X_{t-d}\right) = E\left[X_t|X_{t-1,...,}X_{t-d}\right]$$

and can be estimated by kernel smoothing or any other nonparametric way to estimate a regression function. The problems with this kind of approach are twofold. First, it is often necessary to include many lagged variables and the resulting nonparametric estimation surface suffers from the well-known "curse of dimensionality". Second, it is hard to describe and interpret the estimated regression surface when the dimension is more than two.

A solution to deal with these problems is to think about a kind of nonparametric generalization of ARMA processes. For this purpose, let us consider semiparametric predictors of the following form

$$E\left[X_t|I_{t-1}\right] = m_\varphi\left(\theta, I_{t-1}\right) = \sum_{j=1}^{\infty} a_j\left(\theta\right)\varphi\left(X_{t-j}\right) \tag{7.14}$$

where θ is an unknown finite dimensional vector of parameters, $a_j(.)$, $j \geq 1$, are known given scalar functions and $\varphi(.)$ is the unknown functional parameter of interest. The notation

$$E[X_t|I_{t-1}] = m_\varphi(\theta, I_{t-1})$$

stresses the fact that the predictor depends on the true unknown value of the parameters θ and φ and of the information I_{t-1} available at time $(t-1)$ as well. This information is actually the σ-field generated by $X_{t-j}, j \geq 1$. A typical example is

$$a_j(\theta) = \theta^{j-1} \text{ for } j \geq 1 \text{ with } 0 < \theta < 1. \tag{7.15}$$

Then, the predictor (7.14) is actually characterized by

$$m_\varphi(\theta, I_{t-1}) = \theta m_\varphi(\theta, I_{t-2}) + \varphi(X_{t-1}) \tag{7.16}$$

In the context of volatility modelling, X_t would denote a squared asset return over period $[t-1, t]$ and $m_\varphi(\theta, I_{t-1})$ the so-called squared volatility of this return as expected at the beginning of the period. Engle and Ng (1993) have studied such a partially nonparametric (PNP for short) model of volatility and called the function φ the "news impact function". They proposed an estimation strategy based on piecewise linear splines. Note that the PNP model includes several popular parametric volatility models as special cases. For instance, the GARCH (1,1) model of Bollerslev (1986) corresponds to $\varphi(x) = w + \alpha x$ while the Engle (1990) asymmetric model is obtained for $\varphi(x) = w + \alpha(x + \delta)^2$. See also Linton and Mammen (2003) and references therein.

The nonparametric identification and estimation of the news impact function can be derived for a given value of θ. After that, a profile criterion can be calculated to estimate θ. In any case, since θ will be estimated with a parametric rate of convergence, the asymptotic distribution theory of a nonparametric estimator of φ is the same as if θ were known. For sake of notational simplicity, the dependence on unknown finite dimensional parameters θ is no longer made explicit.

At least in the particular case (7.15)-(7.16), φ is easily characterized as the solution of a linear integral equation of the first kind

$$E[X_t - \theta X_{t-1}|I_{t-2}] = E[\varphi(X_{t-1})|I_{t-2}] \tag{7.17}$$

Except for its dynamic features, this problem is completely similar to the nonparametric instrumental regression example described in Section 5.4. However, as already mentioned, problems of the second kind are often preferable since they may be well-posed. As shown by Linton and Mammen (2003) in the particular case of a PNP volatility model, it is actually possible to identify and consistently estimate the function φ of interest in (**??**) from a well-posed linear inverse problem of the second kind. The main trick is to realize that φ is characterized by the first order conditions of the least squares problem

$$\min_\varphi E\left[X_t - \sum_{j=1}^{\infty} a_j \varphi(X_{t-j})\right]^2 \tag{7.18}$$

Then, when φ is an element of the Hilbert space $L_F^2(X)$, its true unknown value is characterized by the first order conditions obtained by differentiating in the direction of any vector h

$$E\left[\left(X_t - \sum_{j=1}^{\infty} a_j \varphi\left(X_{t-j}\right)\right)\left(\sum_{l=1}^{\infty} a_l h\left(X_{t-l}\right)\right)\right] = 0$$

In other words, for any h in $L_F^2(X)$

$$\begin{aligned}
&\sum_{j=1}^{\infty} a_j E^X\left[E\left[X_t | X_{t-j} = x\right] h(x)\right] \\
&- \sum_{j=1}^{\infty} a_j^2 E^X\left[\varphi(x) h(x)\right] \\
&- \sum_{j=1}^{\infty} \sum_{\substack{l=1 \\ l \neq j}}^{\infty} a_j a_l E^X\left[E\left[\varphi\left(X_{t-l}\right) | X_{t-j} = x\right] h(x)\right] = 0
\end{aligned} \tag{7.19}$$

where E^X denotes the expectation with respect to the stationary distribution of X_t. As the equality (7.19) is true for all h, it is in particular true for a complete sequence of functions of $L_F^2(X)$. It follows that

$$\begin{aligned}
&\sum_{j=1}^{\infty} a_j E\left[X_t | X_{t-j}\right] - \left(\sum_{l=1}^{\infty} a_l^2\right) \varphi\left(X_{t-j}\right) \\
&- \sum_{j=1}^{\infty} \sum_{l \neq j}^{\infty} a_j a_l E\left[\varphi\left(X_{t-l}\right) | X_{t-j}\right] = 0
\end{aligned}$$

P^X- almost surely. Let us denote

$$r_j(X_t) = E\left[X_{t+j} | X_t\right] \text{ and } H_k(\varphi)(X_t) = E\left[\varphi\left(X_{t+k}\right) | X_t\right].$$

Then, we have proved that the unknown function φ of interest must be the solution of the linear inverse problem of the second kind

$$A(\varphi, F) = (I - K)\varphi - r = 0 \tag{7.20}$$

where

$$\begin{aligned}
r &= \left(\sum_{j=1}^{\infty} a_j^2\right)^{-1} \sum_{j=1}^{\infty} a_j r_j \\
K &= -\left(\sum_{j=1}^{\infty} a_j^2\right)^{-1} \sum_{j=1}^{\infty} \sum_{l \neq j} a_j a_l H_{j-l}
\end{aligned},$$

and, with a slight change of notation, F characterizes now the probability distribution of the stationary process (X_t).

To study the inverse problem (7.20), it is first worth noticing that K is a self adjoint integral operator. Indeed, while

$$K = \left(\sum_{j=1}^{\infty} a_j^2 \right)^{-1} \sum_{h=\pm 1}^{+\infty} H_k \left(\sum_{l=\max[1,1-k]}^{+\infty} a_l a_{l+k} \right)$$

we immediately deduce from Subsection 2.5.1 that the conditional expectation operator H_k is such that

$$H_k^* = H_{-k}$$

and thus $K = K^*$, since

$$\sum_{l=\max[1,1-k]}^{+\infty} a_l a_{l+k} = \sum_{l=\max[1,1+k]}^{+\infty} a_l a_{l-k}$$

As noticed by Linton and Mammen (2003), this property greatly simplify the practical implementation of the solution of a sample counterpart of equation (7.19). But, even more importantly, the inverse problem (7.19) will be well-posed as soon as one maintains the following identification assumption about the news impact function φ

Assumption A: There exists no θ and $\varphi \in L_F^2(X)$ with $\varphi \neq 0$ such that $\sum_{j=1}^{\infty} a_j(\theta)\, \varphi(X_{t-j}) = 0$ almost certainly.

To see this, note that assumption A means that for any non-zero function φ

$$0 < E\left[\sum_{j=1}^{\infty} a_j \varphi(X_{t-j}) \right]^2$$

that is

$$\begin{array}{c} 0 < \sum_{j=1}^{\infty} a_j^2 \langle \varphi, \varphi \rangle \\ + \sum_{j=1}^{\infty} \sum_{\substack{l=1 \\ l \neq j}}^{\infty} a_l a_j \langle \varphi, H_{j-l} \varphi \rangle \end{array}$$

Therefore

$$0 < \langle \varphi, \varphi \rangle - \langle \varphi, K\varphi \rangle \tag{7.21}$$

for non zero φ. In other words, there is no non-zero φ such that

$$K\varphi = \varphi$$

and the operator $(I-K)$ is one-to-one. It is also worth noticing that the operator K is Hilbert-Schmidt and a fortiori compact under reasonable assumptions. As already mentioned in subsection 2.5.1, the Hilbert-Schmidt property for the conditional expectation operator H_k is tantamount to the integrability condition

$$\int\int \left[\frac{f_{X,X_{t-k}}(x,y)}{f_{X_t(x)}f_{X_t}(y)}\right]^2 f_{X_t}(x)\, f_{X_t}(y)\, dxdy < \infty$$

It amounts to say that there is not too much dependence between X_t and X_{t-k}. This should be tightly related to the ergodicity or mixing assumptions about the stationary process X_t. Then, if all the conditional expectation operator H_k, $k \geq 1$, are Hilbert-Schmidt, the operator K will also be Hilbert-Schmidt insofar as

$$\sum_{j=1}^{\infty}\sum_{l\neq j} a_j^2 a_l^2 < +\infty$$

Note that (7.21) implies that $(I-K)$ has eigenvalues bounded from below by a positive number.

Up to a straightforward generalization to stationary mixing processes of results only stated in the i.i.d. case, the general asymptotic theory of this subsection 7.3 can then be easily applied to nonparametric estimators of the new impact function φ based on the Fredholm equation of the second kind (7.19). An explicit formula for the asymptotic variance of $\hat{\varphi}_n$ as well as a practical implementation by solution of matricial equations similar to subsection 3.5 (without need of a Tikhonov regularization) is provided by Linton and Mammen (2003) in the particular case of volatility modelling.

However, an important difference with the i.i.d. case (see for instance assumption A.3 in section 5.4 about instrumental variables) is that the conditional homoskedasticity assumption cannot be maintained about conditional probability distribution of X_t given its own past. This should be particularly detrimental in the case of volatility modelling since, when X_t denotes a squared return, it will be in general even more conditionally heteroskedastic than returns themselves. Such a severe conditional heteroskedasticity will likely imply poor finite sample performance and large asymptotic variance of the estimator $\hat{\varphi}_n$ defined from the inverse problem (7.19), that is from the least squares problem (7.18). Indeed, $\hat{\varphi}_n$ is basically kind of OLS estimator in infinite dimension. In order to better take into account conditional heteroskedasticity of X_t in the context of volatility modelling, Linton and Mammen (2003) propose to replace the least squares problem (7.18) by a quasi-likelihood kind of approach where the criterion to optimize is defined from the density function of a normal conditional probability distribution of returns, with variance $m_\varphi(\theta, I_{t-1})$. Then, the difficulty is that the associated first order conditions now characterize the news impact function φ as solution of a nonlinear inverse problem. Linton and Mammen (2003) suggest to work with a version of this problem which is locally linearized around the previously described least squares estimator $\hat{\varphi}_n$ (and associated consistent estimator of θ).

7.4. Regularized solution of an ill posed equation of the second kind and statistical implications

The objective of this section is to study equations $(I-K)\varphi = r$ where 1 is an eigenvalue of K, i.e. where $I-K$ is not injective (or one-to-one). For simplicity we restrict our analysis to the case where the order of multiplicity of the eigenvalue 1 is one and the operator K is self-adjoint. This implies that the dimension of the null spaces of $I-K$ is one and using the results of Section 7.2, the space $\mathcal{H}$ may be decomposed into

$$\mathcal{H} = \mathcal{N}(I-K) \oplus \mathcal{R}(I-K)$$

i.e. $\mathcal{H}$ is the direct sum between the null space and the range of $I-K$, both closed. We denote by $P_{\mathcal{N}}r$ the projection of r on $\mathcal{N}(I-K)$ and by $P_{\mathcal{R}}r$ the projection of r on the range $\mathcal{R}(I-K)$.

Using ii) of Theorem 7.1, a solution of $(I-K)\varphi = r$ exists in the non injective case only if r is orthogonal to $\mathcal{N}(I-K)$ or, equivalently, if r belongs to $\mathcal{R}(I-K)$. In other words, a solution exists if and only if $r = P_{\mathcal{R}}r$. However in this case, this solution is not unique and there exists a one dimensional linear manifold of solutions. Obviously, if φ is a solution, φ plus any element of $\mathcal{N}(I-K)$ is again a solution. This non uniqueness problem will be solved by a normalization rule which selects a unique element in the set of solutions. The normalization we adopt is

$$\langle \varphi, \phi_0 \rangle = 0 \tag{7.22}$$

where ϕ_0 is the eigenfunction of K corresponding to the eigenvalue equal to 1.

In most statistical applications of equations of the second kind, the r element corresponding to the true data generating process is assumed to be in the range of $I-K$ where K is also associated with the true DGP. However this property is no longer true if F is estimated and we need to extend the resolution of $(I-K)\varphi = r$ to cases where $I-K$ is not injective and r is not in the range of this operator. This extension must be done in such a way that the continuity properties of inversion are preserved.

For this purpose we consider the following generalized inverse of $(I-K)$. As K is a compact operator it has a discrete spectrum $\lambda_0 = 1, \lambda_1, ...$ where only 0 may be an accumulation point (in particular 1 cannot be an accumulation point). The associated eigenfunctions are $\phi_0, \phi_1, ...$ Then we define:

$$Lu = \sum_{j=1}^{\infty} \frac{1}{1-\lambda_j} \langle u, \phi_j \rangle \phi_j, \quad u \in \mathcal{H} \tag{7.23}$$

This operator computes the unique solution of $(I-K)\varphi = P_{\mathcal{R}}u$ satisfying the normalization rule (7.22). It can be easily verified that L satisfies:

$$\begin{aligned} LP_{\mathcal{R}} &= L = P_{\mathcal{R}}L \\ L(I-K) &= (I-K)L = P_{\mathcal{R}} \end{aligned} \tag{7.24}$$

It can easily be checked that L is the generalized inverse of $I - K$ as it was defined in Luenberger (1969).

We now consider estimation. For an observed sample, we obtain the estimator F_n of F (that may be built from a kernel estimator of the density) and then the estimators $\hat{r}_n$ and $\hat{K}_n$ of r and K respectively. Let $\hat{\phi}_0$, $\hat{\phi}_1$, ... denote the eigenfunctions of $\hat{K}_n$ associated with $\hat{\lambda}_0$, $\hat{\lambda}_1$, ... We restrict our attention to the cases where 1 is also an eigenvalue of multiplicity one of $\hat{K}_n$ (i.e. $\hat{\lambda}_0 = 1$). However $\hat{\phi}_0$ may be different from ϕ_0.

We have to make a distinction between two cases. First assume that the Hilbert space $\mathcal{H}$ of reference is known and in particular the inner product is given (for example $\mathcal{H} = L^2(\mathbb{R}^p, \pi)$ with π given), the normalization rule imposed to $\hat{\varphi}_n$ is

$$\langle \hat{\varphi}_n, \hat{\phi}_0 \rangle = 0$$

and $\hat{L}_n$ is the generalized inverse of $I - \hat{K}_n$ in $\mathcal{H}$ (which depends on the Hilbert space structure) where

$$\hat{L}_n u = \sum_{j=1}^{\infty} \frac{1}{1 - \hat{\lambda}_j} \langle u, \hat{\phi}_j \rangle \hat{\phi}_j, \quad u \in \mathcal{H}$$

Formula (7.24) applies immediately for F_n.

If however the Hilbert space $\mathcal{H}$ depends on F (e.g. $\mathcal{H} = L^2(\mathbb{R}^p, F)$), we need to assume that $L^2(\mathbb{R}, F_n) \subset L^2(\mathbb{R}^p, F)$. The orthogonality condition which defines the normalization rule (7.22) is related to $L^2(\mathbb{R}^p, F)$ but the estimator $\hat{\varphi}_n$ of φ will be normalized by

$$\langle \hat{\varphi}_n, \hat{\phi}_0 \rangle_n = 0$$

where $\langle \ , \ \rangle_n$ denotes the inner product relative to F_n. This orthogonality is different from an orthogonality relative to $\langle \ , \ \rangle$.

In the same way $\hat{L}_n$ is now defined as the generalized inverse of $I - \hat{K}_n$ with respect to the estimated Hilbert structure, i.e.

$$\hat{L}_n u = \sum_{j=1}^{\infty} \frac{1}{1 - \hat{\lambda}_j} \langle u, \hat{\phi}_j \rangle_n \hat{\phi}_j$$

and $\hat{L}_n$ is not the generalized inverse of $I - \hat{K}_n$ in the original space $\mathcal{H}$. The advantages of this definition is that $\hat{L}_n$ may be effectively computed and satisfies the formula (7.24) where F_n replaces F. In the sequel $P_{\mathcal{R}_n}$ denotes the projection for the inner product $<, >_n$on $\mathcal{R}_n = \mathcal{R}\left(I - \hat{K}_n\right)$.

¿From (7.24) one can deduce that:

$$\begin{aligned} \hat{L}_n - L &= \hat{L}_n(\hat{K}_n - K)L \\ &+ \hat{L}_n(P_{\mathcal{R}_n} - P_{\mathcal{R}}) + (P_{\mathcal{R}_n} - P_{\mathcal{R}})L \end{aligned} \tag{7.25}$$

since $\hat{L}_n - L = \hat{L}_n P_{\mathcal{R}_n} - P_{\mathcal{R}} L = \hat{L}_n(P_{\mathcal{R}_n} - P_{\mathcal{R}}) + (P_{\mathcal{R}_n} - P_{\mathcal{R}})L - P_{\mathcal{R}_n} L + \hat{L}_n P_r$

and $\hat{L}_n (K_n - K) L = \hat{L}_n (K_n - I) L + \hat{L}_n (I - K) L = P_{\mathcal{R}_n} L + \hat{L}_n P_r$.

The convergence property is given by the following theorem:

Theorem 7.5. *Let us define* $\varphi_0 = Lr$ *and* $\hat{\varphi}_n = \hat{L}_n \hat{r}_n$. *If*

i) $\left\| \hat{K}_n - K \right\| = o(1)$

ii) $\|P_{\mathcal{R}_n} - P_{\mathcal{R}}\| = O\left(\frac{1}{b_n}\right)$

iii) $\left\| (\hat{r}_n + \hat{K}_n \varphi_0) - (r + K\varphi_0) \right\| = O\left(\frac{1}{a_n}\right)$

Then

$$\|\hat{\varphi}_n - \varphi_0\| = O\left(\frac{1}{a_n} + \frac{1}{b_n}\right)$$

Proof. The proof is based on:

$$\begin{aligned} \hat{\varphi}_n - \varphi_0 &= \hat{L}_n \hat{r}_n - Lr \\ &= \hat{L}_n(\hat{r}_n - r) + (\hat{L}_n - L)r \\ &= \hat{L}_n(\hat{r}_n - r) + \hat{L}_n(\hat{K}_n - K)\varphi_0 \\ &+ \hat{L}_n (P_{\mathcal{R}_n} - P_{\mathcal{R}})\, r + (P_{\mathcal{R}_n} - P_{\mathcal{R}})\varphi_0 \end{aligned} \tag{7.26}$$

deduced from (7.25). Then

$$\begin{aligned} \|\hat{\varphi}_n - \varphi_0\| &\leq \|\hat{L}_n\| \|(\hat{r}_n + \hat{K}_n \varphi_0) - (r + K\varphi_0)\| \\ &+ (\|\hat{L}_n\| \|r\| + \|\varphi\|) \|P_{\mathcal{R}_n} - P_{\mathcal{R}}\| \end{aligned} \tag{7.27}$$

Under i) and ii) $\|\hat{L}_n - L\| = o(1)$ from (7.25). This implies $\|\hat{L}_n\| \to \|L\|$ and the result follows. ■

If $\dfrac{a_n}{b_n} \sim O(1)$, the actual speed of convergence is bounded by $\dfrac{1}{a_n}$. This will be the case in the two examples of 7.5 where $\dfrac{a_n}{b_n} \to 0$.

We consider asymptotic normality in this case. By (7.24), we have $\hat{L}_n = P_{\mathcal{R}_n} + \hat{L}_n \hat{K}_n$, hence:

$$\begin{aligned} \hat{\varphi}_n - \varphi_0 &= P_{\mathcal{R}_n} \left[(\hat{r}_n + \hat{K}_n \varphi_0) - (r + K\varphi_0) \right] \\ &+ \hat{L}_n \hat{K}_n \left[(\hat{r}_n + \hat{K}_n \varphi_0) - (r + K\varphi_0) \right] \\ &+ \hat{L}_n (P_{\mathcal{R}_n} - P_{\mathcal{R}}) r + (P_{\mathcal{R}_n} - P_{\mathcal{R}})\varphi_0 \end{aligned} \tag{7.28}$$

Let us assume that there exists a sequence a_n such that i) and ii) below are satisfied

i) $a_n P_{\mathcal{R}_n}\left[(\hat{r}_n + \hat{K}_n\varphi_0) - (r + K\varphi_0)\right](x)$ has an asymptotic normal distribution

ii) $a_n\left[\hat{L}_n\hat{K}_n(\hat{r}_n + \hat{K}_n\varphi_0 - r - K\varphi_0)\right](x) \to 0$, $a_n\left[\hat{L}_n\left(P_{\mathcal{R}_n} - P_{\mathcal{R}}\right)r\right](x) \to 0$ and $[(P_{\mathcal{R}_n} - P_{\mathcal{R}})\varphi_0](x) \to 0$

Then the asymptotic normality of $a_n(\hat{\varphi}_n - \varphi_0)$ is driven by the behavior of the first term of the decomposition (7.28). This situation occurs in the non parametric estimation as illustrated in the next section.

7.5. Two examples: backfitting estimation in additive models and panel model

7.5.1. Backfitting estimation in additive models

Let us recall that in an additive model defined by

$$\begin{array}{l}(Y, Z, W) \in \mathbb{R} \times \mathbb{R}^p \times \mathbb{R}^q \\ Y = \varphi(Z) + \psi(W) + U \\ E(U|Z, W) = 0,\end{array} \tag{7.29}$$

in which case (see 1.24), the function φ is solution of the equation:

$$\varphi - E\left[E(\varphi(Z)|W)|Z\right] = E(Y|Z) - E\left[E(Y|W)|Z\right]$$

and ψ is the solution of an equation of the same nature obtained by a permutation of W and Z. We focus our presentation on the estimation of φ. It appears as the resolution of a linear equation of the second kind. More precisely, we have in that case :

- $\mathcal{H}$ is the space of the square integrable functions of Z with respect to the true data generating process. This definition simplifies our presentation but an extension to different spaces is possible.
- The unknown function φ is an element of $\mathcal{H}$. Actually asymptotic considerations will restrict the class of functions φ by smoothness restrictions.
- The operator K is defined by $K\varphi = E\left[E(\varphi(Z)|W)|Z\right]$. This operator is self adjoint and we assume its compactness. This compactness may be obtained through the Hilbert Schmidt assumption A.1 of section 5.
- The function r is equal to $E(Y|Z) - E\left[E(Y|W)|Z\right]$.

The operator $I - K$ is not one-to-one because the constant functions belong to the null space of this operator. Indeed the additive model (7.29) does not identify φ and ψ. We introduce the following hypothesis which warrants that φ (and ψ) are exactly identified up to an additive constant or, equivalently, that the null space of $I - K$ only contains the constants.

Identification assumption. Z and W are measurably separated w.r.t. the distribution F i.e. a function of Z almost surely equal to a function of W is almost surely constant.

This assumption implies that if $\varphi_1, \varphi_2, \psi_1, \psi_2$ are such that $E(Y|Z,W) = \varphi_1(Z) + \psi_1(W) = \varphi_2(Z) + \psi_2(W)$ then $\varphi_1(Z) - \varphi_2(Z) = \psi_2(W) - \psi_1(W)$ which implies that $\varphi_1 - \varphi_2$ and $\psi_2 - \psi_1$ are a.s. constant. In terms of the null set of $I - K$ we have:

$$\begin{aligned} K\varphi &= \varphi \Longleftrightarrow E\left[E(\varphi(Z)|W)|Z\right] = \varphi(Z) \\ &\Longrightarrow E\left[\left(E\left[\varphi(Z)|W\right]\right)^2\right] \\ &= E\left[\varphi(Z) E\left(\varphi(Z)|W\right)\right] = E\left(\varphi^2(Z)\right). \end{aligned}$$

But, by Pythagore theorem:

$$\begin{aligned} \varphi(Z) &= E(\varphi(Z)|W) + \upsilon \\ E\left(\varphi^2(Z)\right) &= E\left(\left(E\left(\varphi(Z)|W\right)\right)^2\right) + E\upsilon^2 \end{aligned}$$

Then:

$$\begin{aligned} K\varphi &= \varphi \Longrightarrow \upsilon = 0 \\ &\Leftrightarrow \varphi(Z) = E\left[\varphi(Z)|W\right]. \end{aligned}$$

Then if φ is an element of the null set of $I - K$, φ is almost surely equal to a function of W and is therefore constant.

The eigenvalues of K are real positive and smaller than 1 except for the first one. We have $1 = \lambda_0 > \lambda_1 > \lambda_2 ... > .$[1] The eigenfunctions are such that $\phi_0 = 1$ and the condition $\langle \varphi, \phi_0 \rangle = 0$ means that φ has an expectation equal to zero. The range of $I - K$ is the set of functions with a mean equal to 0 and the projection of u, $P_{\mathcal{R}}u$, equals $u - E(u)$.

It should be noticed that under the hypothesis of additive model, r has zero mean and is then an element of $\mathcal{R}(I - K)$. Then a unique (up to the normalization condition) solution of the structural equation $(I - K)\varphi = r$ exists.

The estimation may be done by kernel smoothing. The joint density is estimated by

$$f_n(y,z,w) = \frac{1}{nc_n^{1+p+q}} \sum_{i=1}^{n} \omega\left(\frac{y - y_i}{c_n}\right) \omega\left(\frac{z - z_i}{c_n}\right) \omega\left(\frac{w - w_i}{c_n}\right) \tag{7.30}$$

and F_n is the c.d.f. associated to f_n. The estimated $\hat{K}_n$ operator verifies:

$$(\hat{K}_n\varphi)(z) = \int \varphi(u)\, \hat{a}_n(u,z)\, du \tag{7.31}$$

[1] Actually $K = T^*T$ when $T\varphi = E(\varphi|W)$ and $T^*\psi = E(\psi|Z)$ when ψ is a function of W. The eigenvalues of K correspond to the squared singular values of the T and T^* defined in Section 2.

where

$$\hat{a}_n(u,z) = \int \frac{\hat{f}_n(.,u,w)\,\hat{f}_n(.,z,w)}{\hat{f}_n(.,.,w)\,\hat{f}_n(.,z,.)} dw.$$

The operator $\hat{K}_n$ must be an operator from $\mathcal{H}$ to $\mathcal{H}$ (it is by construction an operator from $L^2_Z(F_n)$ into $L^2_Z(F_n)$). Therefore $\frac{\omega\left(\frac{z-z_\ell}{c_n}\right)}{\sum_\ell \omega\left(\frac{z-z_\ell}{c_n}\right)}$ must be square integrable w.r.t. F.

The estimation of r by $\hat{r}_n$ verifies

$$\hat{r}_n(z) = \frac{1}{\sum_{\ell=1}^{n} \omega\left(\frac{z-z_\ell}{c_n}\right)} \sum_{\ell=1}^{n} \left(y_\ell - \sum_{i=1}^{n} y_i \omega_{li} \right) \omega\left(\frac{z-z_\ell}{c_n}\right)$$

where $\omega_{li} = \dfrac{\omega\left(\dfrac{w_l - w_i}{c_n}\right)}{\sum_{i=1}^{n} \omega\left(\dfrac{w_l - w_i}{c_n}\right)}$.

The operator $\hat{K}_n$ has also 1 as the greatest eigenvalue corresponding to an eigenfunction equal to 1. Since F_n is a mixture of probabilities for which z and w are independent, the measurable separability between Z and W is fulfilled. Then the null set of $I - \hat{K}_n$ reduces to a.s. (w.r.t. F_n) constant functions. The generalized inverse of an operator depends on the inner product of the Hilbert space because it is defined as the function φ of minimal norm which minimizes the norm of $\hat{K}_n\varphi - \hat{r}_n$. The generalized inverse in the space $L^2_Z(F)$ cannot be used for the estimation because it depends on the actual unknown F. Then we construct $\hat{L}_n$ as the generalized inverse in $L^2_Z(F_n)$ of $I - \hat{K}_n$. The practical computation of $\hat{L}_n$ can be done by computing the n eigenvalues of $\hat{K}_n$, $\hat{\lambda}_0 = 1, ..., \hat{\lambda}_{n-1}$ and the n eigenfunctions $\hat{\phi}_0 = 1, \hat{\phi}_1, ..., \hat{\phi}_{n-1}$. Then

$$\hat{L}_n u = \sum_{j=1}^{n-1} \frac{1}{1 - \hat{\lambda}_j} \left\{ \int u(z) \hat{\phi}_j(z) \hat{f}_n(z) dz \right\} \hat{\phi}_j \tag{7.32}$$

It can be easily checked that property (7.24) is verified where $P_{\mathcal{R}_n}$ is the projection (w.r.t. F_n) on the orthogonal of the constant function. This operator subtracts to any functions its empirical mean computed through the smoothed density :

$$P_{\mathcal{R}_n} u = u - \frac{1}{n c_n^p} \sum_i \int u(z) \omega\left(\frac{z - z_i}{c_n}\right) dz \tag{7.33}$$

The right hand side of the equation $(I - \hat{K}_n)\varphi = \hat{r}_n$ has a mean equal to 0 (w.r.t. F_n). Hence, this equation has a unique solution $\hat{\varphi}_n = \hat{L}_n \varphi_0$ which satisfies the normalization condition $\frac{1}{n c_n^p} \sum_i \int \hat{\varphi}_n(z) \omega\left(\frac{z-z_i}{c_n}\right) dz = 0$.

The general results of Section 7.4 apply.

1) Under very general assumptions, $\|\hat{K}_n - K\| \to 0$ in probability.

2) We have to check the properties of $P_{\mathcal{R}_n} - P_{\mathcal{R}}$

$$(P_{\mathcal{R}_n} - P_{\mathcal{R}})\varphi = \frac{1}{nc_n^p} \sum_i \int \varphi(z)\omega\left(\frac{z - z_i}{c_n}\right) dz - \int \varphi(z) f(z) dz$$

The asymptotic behavior of $\|(P_{\mathcal{R}_n} - P_{\mathcal{R}})\varphi\|^2 = \left| \frac{1}{nc_n^p} \sum_{i=1}^n \int \varphi(z)\omega\left(\frac{z-z_i}{c_n}\right) dz - E(\varphi) \right|^2$ is the same as the asymptotic behavior of the expectation of this positive random variable :

$$E\left(\frac{1}{nc_n^p} \sum_{i=1}^n \int \varphi(z)\omega\left(\frac{z - z_i}{c_n}\right) dz - E(\varphi) \right)^2$$

Standard computation on this expression shows that this mean square error is $O\left(\frac{1}{n} + c_n^{2min(d,d')}\right) \|\varphi\|^2$, where d is the smoothness degree of φ and d' the order of the kernel.

3) The last term we have to consider is actually not computable but its asymptotic behavior is easily characterized. We simplify the notation by denoting $E^{F_n}(.|.)$ the estimation of a conditional expectation. The term we have to consider is

$$\begin{array}{rcl} (\hat{r}_n + \hat{K}_n\varphi) - (r + K\varphi) & = & E^{F_n}(Y|Z) - E^{F_n}(E^{F_n}(Y|W)|Z) + E^{F_n}(E^{F_n}(\varphi(Z)|W)|Z) \\ & - & E^F(Y|Z) + E^F(E^F(Y|W)|Z) - E^F(E^F(\varphi(Z)|W)|Z) \\ & = & E^{F_n}\left(Y - E^F(Y|W) + E^F(\varphi(Z)|W)|Z\right) \\ & - & E^F\left(Y - E^F(Y|W) + E^F(\varphi(Z)|W)|Z\right) \\ & - & R \end{array}$$

where $R = E^F\left\{E^{F_n}(Y - \varphi(Z)|W) - E^F(Y - \varphi(Z)|W)\right\}$

1. Moreover

$$E^F(Y|W) = E^F(\varphi(Z)|W) + \psi|W)$$

Then

$$\begin{array}{rcl} (r_n + K_n\varphi) - (r + K\varphi) & = & E^{F_n}(Y - \psi(W)|Z) - E^F(Y - \psi(W)|Z) \\ & - & R \end{array}$$

The R term converges at a faster speed than the first part of the r.h.s. of this equation and can be neglected.

We have seen in the other parts of this chapter that

$$\|E^{F_n}(Y-\psi(W)|Z)-E^{F}(Y-\psi(W)|Z)\|^2 \sim 0\left(\frac{1}{nc_n^p}+c_n^{2\rho}\right)$$

where ρ depends on the regularity assumptions.

We can conclude that $\|\hat{\varphi}_n-\varphi_0\| \to 0$ in probability and that $\|\hat{\varphi}_n-\varphi_0\| \sim 0\left(\frac{1}{\sqrt{nc_n^p}}+c_n^{\rho}\right)$.

The pointwise asymptotic normality is now easy to verify. Consider $\sqrt{nc_n^{\rho}}(\hat{\varphi}_n(z)-\varphi_0(z))$. We adapt in this framework the formula (7.28) and Theorem 7.4.

1) Under a suitable condition on c_n (typically $nc_n^{\rho+2\min(d,r)} \to 0$), we have:
$\sqrt{nc_n^p}\left\{\hat{L}_n(P_{\mathcal{R}_n}-P_{\mathcal{R}})r+(P_{\mathcal{R}_n}-P_{\mathcal{R}})\varphi\right\} \to 0$ in probability.

2) Using the same argument as in 7.4, a suitable choice of c_n implies that

$$\sqrt{nc_n^{\rho}}\hat{L}_n\hat{K}_n\left[(\hat{r}_n+\hat{K}_n\varphi_0)-(r+K\varphi_0)\right] \to 0$$

Actually, while $E^{F_n}(Y-\psi(W)|Z)-E^{F}(Y-\psi(W)|Z)$ only converges pointwise at a non parametric speed, the transformation by the operator $\hat{K}_n$ transforms this convergence into a functional convergence at a parametric speed. Then

$$\sqrt{nc_n^p}\left\|\hat{K}_n\left[E^{F_n}(Y-\psi(W)|Z)-E^{F}(Y-\psi(W)|Z\right]\right\| \to 0$$

Moreover $\hat{L}_n$ converge in norm to L which is a bounded operator and the result follows.

3) The convergence of $\sqrt{nc_n^p}(\varphi_{F_n}(z)-\varphi_F(z))$ is then identical to the convergence of

$$\sqrt{nc_n^p}P_{\mathcal{R}_n}\left[E^{F_n}(Y-\psi(W)|Z)-E^{F}(Y-\psi(W)|Z\right]$$

$$=\sqrt{nc_n^p}\Bigg[E^{F_n}(Y-\psi(W)|Z)-E^{F}(Y-\psi(W)|Z)$$

$$-\tfrac{1}{n}\sum_i(Y_i-\psi(W_i))-\tfrac{1}{nc_n^p}\sum_i\int(Y-\psi(W))f(Y,W|Z)\omega\left(\frac{z-z_i}{c_n}\right)dz\Bigg]$$

Then also it can be easily checked that the difference between the two sample means converge to zero at a higher speed than $\sqrt{nc_n^p}$ and these two last terms can be cancelled. Then using standard results on nonparametric estimation, we obtain:

$$\sqrt{nc_n^p}(\varphi_{F_n}(z) - \varphi_F(z)) \xrightarrow{d} \mathcal{N}\left(0, \frac{\int \omega^2}{f_Z(z)} Var(Y - \psi(W)|Z = z)\right)$$

where the 0 mean of the asymptotic distribution is obtained thanks to a suitable choice of the bandwidth, which needs to converge to 0 faster than the optimal speed.

7.5.2. Estimation of the bias function in a measurement error equation

We have introduced in Example 1.3.6, Section 1, the measurement error model:

$$\begin{cases} Y_1 = \eta + \varphi(Z_1) + U_1 & Y_1, Y_2 \in \mathbb{R} \\ Y_2 = \eta + \varphi(Z_2) + U_2 & Z_1, Z_2 \in \mathbb{R}^p \end{cases}$$

when η, U_i are random unknown elements and Y_1 and Y_2 are two measurements of η contaminated by a bias term depending on observable elements Z_1 and Z_2. The unobservable component η is eliminated by difference and we get the model under consideration :

$$Y = \varphi(Z_2) - \varphi(Z_1) + U \tag{7.34}$$

when $Y = Y_2 - Y_1$ and $E(Y|Z_1, Z_2) = \varphi(Z_2) - \varphi(Z_1)$. We assume that i.i.d. observations of (Y, Z_1, Z_2) are available. Moreover the order of measurements is arbitrary or equivalently (Y_1, Y_2, Z_1, Z_2) is distributed identically to (Y_2, Y_1, Z_2, Z_1). This implies that (Y, Z_1, Z_2) and $(-Y, Z_2, Z_1)$ have the same distribution. In particular, Z_1 and Z_2 are identically distributed.

- The reference space $\mathcal{H}$ is the space of random variables defined on R^p that are square integrable with respect to the true marginal distribution on Z_1 (or Z_2). We are in a case where the Hilbert space structure depends on the unknown distribution
- The function φ is an element of $\mathcal{H}$ but this set has to be reduced by smoothness condition in order to obtain asymptotic properties of the estimation procedure.
- The operator K is the conditional expectation operator

$$\begin{aligned} (K\varphi)(z) &= E^F(\varphi(Z_2)|Z_1 = z) \\ &= E^F(\varphi(Z_1)|Z_2 = z) \end{aligned}$$

  from $\mathcal{H}$ to $\mathcal{H}$. The two conditional expectations are equal because (Z_1, Z_2) and (Z_2, Z_1) are identically distributed (by the exchangeability property). This operator is self-adjoint and we suppose that K is compact. This property may be deduced as in previous cases from an Hilbert Schmidt argument.

Equation (7.34) introduces an overidentification property because it constrains the conditional expectation of Y given Z_1 and Z_2. In order to define φ for any F (and in particular for the estimated one), the parameter φ is now defined as the solution of the minimization problem:

$$\varphi = \arg\min_{\varphi} E\left(Y - \varphi\left(Z_2\right) + \varphi\left(Z_1\right)\right)^2$$

or, equivalently as the solution of the first-order conditions:

$$E^F\left[\varphi\left(Z_2\right)|Z_1 = z\right] - \varphi\left(z\right) = E\left(Y|Z_1 = z\right)$$

because $\left(Y, Z_1, Z_2\right) \sim \left(-Y, Z_2, Z_1\right)$.

Then the integral equation which defines the functions of interest φ may be denoted by

$$\left(I - K\right)\varphi = r$$

where $r = E\left(Y|Z_2 = z\right) = -E\left(Y|Z_1 = z\right)$. As in the additive models, this inverse problem is ill-posed because $I - K$ is not one-to-one. Indeed, 1 is the greatest eigenvalue of K and the eigenfunctions associated with 1 are the constant functions. We need an extra assumption to warranty that the order of multiplicity is one, or, in more statistical terms, that φ is identified up to a constant. This property is obtained if Z_1 and Z_2 are measurably separated i.e. if the functions of Z_1 almost surely equal to some functions of Z_2 are almost surely constant.

Then the normalization rule is

$$\langle \varphi, \phi_0 \rangle = 0$$

where ϕ_0 is constant. This normalization is then equivalent to

$$E^F\left(\varphi\right) = 0.$$

If F is estimated using standard kernel procedure, the estimated F_n does not satisfied, in general, the exchangeability assumption ($\left(Y, Z_1, Z_2\right)$ and $\left(-Y, Z_2, Z_1\right)$ are identically distributed). A simple way to incorporate this constraint is to estimate F using a sample of size $2n$ by adding to the original sample $\left(y_i, z_{1i}, z_{2i}\right)_{i=1,\ldots,n}$ a new sample $\left(-y_i, z_{2i}, z_{1i}\right)_{i=1,\ldots,n}$. For simplicity we do not follow this method here and we consider an estimation of F which does not verify the exchangeability. In that case $\hat{r}_n$ is not, in general, an element of $\mathcal{R}\left(I - \hat{K}_n\right)$ and the estimator $\hat{\varphi}_n$ is defined as the unique solution of

$$\left(I - \hat{K}_n\right)\varphi = P_{\mathcal{R}_n}\hat{r}_n,$$

which satisfies the normalization rule

$$E^{F_n}\left(\varphi\right) = 0.$$

Equivalently we have seen that the functional equation $\left(I - \hat{K}_n\right)\varphi = \hat{r}_n$ reduces to a n dimensional linear system, which is solved by a generalized inversion. The asymptotic properties of this procedure follows immediately from the theorems of Section 7.4 and are obtained identically to the case of additive models.

References

[1] Ai, C. and X. Chen (1999) "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions", mimeo, New York University.

[2] Aronszajn, N. (1950) "Theory of Reproducing Kernels", *Transactions of the American Mathematical Society*, Vol. 68, 3, 337-404.

[3] Basmann, R.L. (1957), *A Generalized Classical Method of Linear Estimation of Coefficients in a Structural Equations*, Econometrica, 25, 77-83.

[4] Bollerslev, T. (1986), "Generalized Autoregressive Conditional Heteroskedasticity", *Journal of Econometrics* 31, 307-327.

[5] Bosq, D. (1998) "Nonparametric Statistics for Stochastic Processes. Estimation and Prediction", Springer, Berlin.

[6] Carrasco, M., M. Chernov, J.-P. Florens, and E. Ghysels (2001) "Efficient estimation of jump diffusions and general dynamic models with a continuum of moment conditions", mimeo, University of Rochester.

[7] Carrasco, M. and J. P. Florens (2000) "Generalization of GMM to a continuum of moment conditions", *Econometric Theory*, 16, 797-834.

[8] Carrasco, M. and J. P. Florens (2001) "Efficient GMM Estimation Using the Empirical Characteristic Function", mimeo, University of Rochester.

[9] Carrasco, M. and J.-P. Florens (2002a) "Spectral method for deconvolving a density", mimeo, University of Rochester.

[10] Carrasco, M. and J. P. Florens (2002b) "On the Asymptotic Efficiency of GMM", mimeo, University of Rochester.

[11] Carroll, R. and P. Hall (1988) "Optimal Rates of Convergence for Deconvolving a Density", *Journal of American Statistical Association*, 83, No.404, 1184-1186.

[12] Carroll, R., A. Van Rooij, and F. Ruymgaart (1991) "Theoretical Aspects of Ill-posed Problems in Statistics", *Acta Applicandae Mathematicae*, 24, 113-140.

[13] Chacko, G. and L. Viceira (2000) "Spectral GMM Estimation of Continuous-Time Processes", forthcoming in *Journal of Econometrics*.

[14] Chen, X., L.P. Hansen and J. Scheinkman (1998) "Shape-preserving Estimation of Diffusions", mimeo, University of Chicago.

[15] Chen, X., and H. White (1992), *Central Limit and Functional Central Limit Theorems for Hilbert Space-Valued Dependent Processes*, Working Paper, University of San Diego.

[16] Chen, X. and H. White (1996) "Law of Large Numbers for Hilbert Space-Valued mixingales with Applications", *Econometric Theory,* 12, 284-304.

[17] Chen, X. and H. White (1998) "Central Limit and Functional Central Limit Theorems for Hilbert Space-Valued Dependent Processes", *Econometric Theory,* 14, 260-284.

[18] Darolles, S., J.P. Florens, and C. Gourieroux (2000) "Kernel Based Nonlinear Canonical Analysis and Time Reversibility", forthcoming in *Journal of Econometrics.*

[19] Darolles, S., J.P. Florens, and E. Renault (1998), "Nonlinear Principal Components and Inference on a Conditional Expectation Operator with Applications to Markov Processes", presented at Garchy, France, September 1998.

[20] Darolles, S., J.P. Florens, and E. Renault (2002), "Nonparametric Instrumental Regression", Working paper 05-2002, CRDE.

[21] Dautray, R. and J.-L. Lions (1988) *Analyse mathématique et calcul numérique pour les sciences et les techniques.* Vol. 5. Spectre des opérateurs, Masson, Paris.

[22] Davidson, J. (1994) *Stochastic Limit Theory*, Oxford University Press, Oxford.

[23] Debnath, L. and P. Mikusinski (1999) *Introduction to Hilbert Spaces with Applications*, Academic Press. San Diego.

[24] Dunford, N. and J. Schwartz (1988) *Linear Operators, Part II: Spectral Theory*, Wiley, New York.

[25] Engle R.F., (1990), Discussion: Stock Market Volatility and the Crash of '87, *Review of Financial Studies* 3, 103-106.

[26] Engle, R.F., D.F. Hendry and J.F. Richard, (1983), "Exogeneity", *Econometrica*, 51 (2) 277-304.

[27] Engle, R.F., and V.K. Ng (1993), "Measuring and Testing the Impact of News on Volatility", *The Journal of Finance* XLVIII, 1749-1778.

[28] Fan, J. (1993) "Adaptively local one-dimentional subproblems with application to a deconvolution problem", *The Annals of Statistics*, 21, 600-610.

[29] Feuerverger, A. and P. McDunnough (1981), "On the Efficiency of Empirical Characteristic Function Procedures", *J. R. Statist. Soc.* B, 43, 20-27.

[30] Florens, J.P., J. Heckman, C. Meghir and E. Vytlacil (2002), " Instrumental Variables, Local Instrumental Variables and Control Functions", Manuscript, University of Toulouse.

[31] Florens, J.P. and Malavolti (2002) "Instrumental Regression with Discrete Variables", mimeo University of Toulouse, presented at ESEM 2002, Venice.

[32] Florens, J.P. and M. Mouchart (1985), "Conditioning in Dynamic Models", *Journal of Time Series Analysis*, 53 (1), 15-35.

[33] Florens, J.P., M. Mouchart, and J.F. Richard (1974), *Bayesian Inference in Error-in-variables Models*, Journal of Multivariate Analysis, 4, 419-432.

[34] Florens, J.P., M. Mouchart, and J.F. Richard (1987), *Dynamic Error-in-variables Models and Limited Information Analysis*, Annales d'Economie et Statistiques, 6/7, 289-310.

[35] Florens, J.P., C. Protopopescu, and J.F. Richard, (1997), "Identification and Estimation of a Class of Game Theoretic Models", GREMAQ-University of Toulouse.

[36] Forni, M. and L. Reichlin (1998) "Let's Get Real: A Factor Analytical Approach to Disaggregated Business Cycle Dynamics", *Review of Economic Studies*, 65, 453-473.

[37] Gallant, A. R. and J. R. Long (1997) "Estimating Stochastic Differential Equations Efficiently by Minimum Chi-squared", *Biometrika*, 84, 125-141.

[38] Gaspar, P. and J.P. Florens, (1998), "Estimation of the Sea State Biais in Radar Altimeter Measurements of Sea Level: Results from a Nonparametric Method", *Journal of Geophysical Research*, 103 (15), 803-814.

[39] Guerre, E., I. Perrigne, and Q. Vuong, (2000), "Optimal Nonparametric Estimation of First-Price Auctions", *Econometrica*, 68 (3), 525-574.

[40] Hansen, L., (1982), "Large Sample Properties of Generalized Method of Moments Estimators", *Econometrica*, 50, 1029-1054.

[41] Hansen, L. (1985) "A method for calculating bounds on the asymptotic covariance matrices of generalized method of moments estimators", *Journal of Econometrics*, 30, 203-238.

[42] Hardle, W. and O. Linton (1994) "Applied Nonparametric Methods", *Handbook of Econometrics*, Vol. IV, edited by R.F. Engle and D.L. McFadden, North Holland, Amsterdam.

[43] Hastie, T.J. and R.J. Tibshirani (1990), *Generalized Additive Models*, Chapman and Hall, London.

[44] Hausman, J., (1981), "Exact Consumer's Surplus and Deadweight Loss" *American Economic Review*, 71, 662-676.

[45] Hausman, J. (1985), "The Econometrics of Nonlinear Budget sets" *Econometrica*, 53, 1255-1282.

[46] Hausman, J. and W.K. Newey, (1995) "Nonparametric Estimation of Exact Consumers Surplus and Deadweight Loss", *Econometrica*, 63, 1445-1476.

[47] Heckman, J., H. Ichimura, J. Smith, and P. Todd (1998), *Characterizing Selection Bias Using Experimental Data*, Econometrica, 66, 1017-1098.

[48] Heckman, J., and V. Vytlacil (2000), *Local Instrumental Variables*, in Nonlinear Statistical Modeling: Proceedings of the Thirteenth International Symposium in Economic Theory and Econometrics: Essays in Honor of Takeshi Amemiya, ed. by C. Hsiao, K. Morimune, and J. Powells. Cambridge: Cambridge University Press, 1-46.

[49] Hoerl, A. E. and R. W. Kennard (1970) "Ridge Regression: Biased Estimation of Nonorthogonal Problems", *Technometrics*, 12, 55-67.

[50] Imbens, G., and J. Angrist (1994), *Identification and Estimation of Local Average Treatment Effects*, Econometrica, 62, 467-476.

[51] Jiang, G. and J. Knight (2000) "Estimation of Continuous Time Processes Via the Empirical Characteristic Function", forthcoming in *Journal of Business & Economic Statistics*.

[52] Judge, G., W. Griffiths, R. C. Hill, H. Lutkepohl, and T-C. Lee (1980) *The Theory and Practice of Econometrics*, John Wiley and Sons, New York.

[53] Kitamura, Y. and M. Stutzer (1997), "An Information Theoretic Alternative to Generalized Method of Moments Estimation", *Econometrica*, 65, 4, 861-874.

[54] Knight, J. L. and J. Yu (2002) "Empirical Characteristic Function in Time Series Estimation", *Econometric Theory*, 18, 691-721.

[55] Kress, R. (1999), *Linear Integral Equations,* Springer.

[56] Kutoyants, Yu. (1984), *Parameter estimation for stochastic processes,* Heldermann Verlag, Berlin.

[57] Lancaster, H. (1968), *The Structure of Bivariate Distributions,* Ann. Math. Statist., 29, 719-736.

[58] Linton, O. and E. Mammen (2003), "Estimating Semiparametric ARCH(∞) models by kernel smoothing methods", Discussion Paper, LSE. Invited Paper ESEM 2003, Stockholm.

[59] Loubes, J.M. and A. Vanhems (2001), "Differential Equation and Endogeneity", Discussion Paper, GREMAQ, University of Toulouse, presented at ESEM 2002, Venice.

[60] Loubes, J.M. and A. Vanhems (2003), "Saturation Spaces for Regularization Methods in Inverse Problems", Discussion Paper, GREMAQ, University of Toulouse, presented at ESEM 2003, Stockholm.

[61] Lucas, R. (1978) "Asset Prices in an Exchange Economy", *Econometrica*, 46, 1429-1446.

[62] Luenberger, D. G. (1969) *Optimization by Vector Space Methods*, Wiley, New York.

[63] Malinvaud, E. (1970), *Methodes Statistiques de l'Econométrie*, Dunod, Paris.

[64] Nashed, N. Z. and G. Wahba (1974) "Generalized inverses in reproducing kernel spaces: An approach to regularization of linear operator equations", *SIAM J. Math. Anal.* 5, 974-987.

[65] Newey, W., and J. Powell (2000), *Instrumental Variables for Nonparametric Models,* MIT Discussion Paper.

[66] Newey, W., Powell, J. and F. Vella (1999), *Nonparametric Estimation of Triangular Simultaneous Equations Models*, Econometrica, **67**, 565-604.

[67] Owen, A. (2001) *Empirical likelihood,* Monographs on Statistics and Applied Probability, vol. 92. Chapman and Hall, London.

[68] Pagan, A. and A. Ullah (1999), *Nonparametric Econometrics*, Cambridge University Press.

[69] Parzen, E. (1959) "Statistical Inference on time series by Hilbert Space Methods,I.", Technical Report No.23, Applied Mathematics and Statistics Laboratory, Stanford. Reprinted in (1967) *Time series analysis papers*, Holden-Day, San Francisco.

[70] Parzen, E. (1970) "Statistical Inference on time series by RKHS methods", Proc. 12th Biennal Canadian Mathematical Seminar, R. Pyke, ed. American Mathematical Society, Providence.

[71] Politis, D. and J. Romano (1994) "Limit theorems for weakly dependent Hilbert space valued random variables with application to the stationary bootstrap", Statistica Sinica, 4, 451-476.

[72] Qin, J. and J. Lawless, (1994), "Empirical Likelihood and General Estimating Equations", *The Annals of Statistics*, 22, 1, 300-325.

[73] Reiersol, O. (1941), *Confluence Analysis of Lag Moments and other Methods of Confluence Analysis*, Econometrica, 9, 1-24.

[74] Reiersol, O. (1945), *Confluence Analysis by Means of Instrumental Sets of Variables*, Arkiv for Mathematik, Astronomie och Fysik, 32.

[75] Rust, J., J. F. Traub, and H. Wozniakowski (2002) "Is There a Curse of Dimensionality for Contraction Fixed Points in the Worst Case?", *Econometrica*, 70, 285-330.

[76] Ruymgaart, F. (2001) "A short introduction to inverse statistical inference", lecture given at the conference "L'Odyssée de la Statistique", Institut Henri Poincaré, Paris.

[77] Saitoh, S. (1997) *Integral transforms, reproducing kernels and their applications*, Longman.

[78] Sansone, G. *Orthogonal Functions*, Dover Publications, New York.

[79] Sargan, J.D. (1958), *The Estimation of Economic Relationship using Instrumental Variables,* Econometrica, 26, 393-415.

[80] Singleton, K. (2001) "Estimation of Affine Pricing Models Using the Empirical Characteristic Function", *Journal of Econometrics*, 102, 111-141.

[81] Stefanski, L. and R. Carroll (1990) "Deconvoluting Kernel Density Estimators", *Statistics*, 2, 169-184.

[82] Stock, J. and M. Watson (1998) "Diffusion Indexes", NBER working paper 6702.

[83] Tauchen, G. (1997) "New Minimum Chi-Square Methods in Empirical Finance", in *Advances in Econometrics*, *Seventh World Congress*, eds. D. Kreps and K. Wallis, Cambridge University Press, Cambridge.

[84] Tauchen, G. and R. Hussey (1991) "Quadrature-Based Methods for Obtaining Approximate Solutions to Nonlinear Asset Pricing Models", *Econometrica*, 59, 371-396.

[85] Theil, H.(1953), *Repeated Least Squares Applied to complete Equations System*, The Hague: Central Planning Bureau (mimeo).

[86] Vanhems, A. (2000), "Nonparametric Solutions to Random Ordinary Differential Equations of First Orders", GREMAQ-University of Toulouse.

[87] Van Rooij and F. Ruymgaart (1991) "Regularized Deconvolution on the Circle and the Sphere", in *Nonparametric Functional Estimation and Related Topics*, edited by G. Roussas, 679-690, Kluwer Academic Publishers, the Netherlands.

[88] Van Rooij, A., F. Ruymgaart, and W. Van Zwet (2000) "Asymptotic Efficiency of Inverse Estimators", *Theory Probab. Appl.*, 44, 4, 722-738.

[89] Vapnik A.C.M. (1998), *Statistical Learning Theory*, Wiley, New York.