Is there an optimal weighting for linear inverse problems?*

Jean-Pierre FLORENS† Senay SOKULLU‡
Toulouse School of Economics University of Bristol

October 19, 2015

Abstract

This paper considers linear equations $\hat{r} = K\varphi + U$ in functional spaces where K and the variance of U, Σ are given. The function φ is estimated by minimizing a Tikhonov functional $\|A\hat{r} - AK\varphi\|^2 + \alpha\|L\varphi\|^2$ where α is a regularization parameter and A and L are two chosen operators. We analyse the optimal mean integrated square error, $\min_\alpha \mathbb{E}\|\hat{\varphi}_\alpha - \varphi\|^2$ in order to determine the optimal choice of A (and L). Contrary to the finite dimensional case $A = \Sigma^{-1/2}$ is not optimal and the best choice depends on the regularity of φ and the degree of ill-posedness of Σ.

Keywords: Nonparametric IV Regression, Inverse problems, Tikhonov Regularization, Regularization Parameter, GMM
JEL Classification: C13, C14, C30

*Preliminary, please do not cite or circulate.

†jean-pierre.florens@tse-fr.eu
‡senay.sokullu@bristol.ac.uk

1 Introduction

Following Hansen (1982) it is established that in a GMM approach with more moment conditions than the dimension of the parameter, optimal variance of the estimator is obtained if the empirical moments are weighted by the inverse of the square root of the variance matrix. This result is asymptotic in the general case but it may not be exact in particular case. In this paper we investigate this result for linear inverse problems such as infinite dimensional GMM. By means of simulation, we show that "*the optimal*" weighting is no longer optimal once we have an infinite dimensional parameter of interest.

We can illustrate Hansen (1982)'s result by a simple example. Let X be a random element, $m(X) \in \mathbb{R}^q$ be a set of integrable linear functions and assume that $r = E(m(X)) = M\theta$ where M is an $q \times k$ matrix and $\theta \in \mathbb{R}^k$ is an unknown parameter. Let $\hat{r}$ be the empirical moments of m ($\hat{r} = \frac{1}{n}\sum_{i=1}^{n} m(x_i)$) and $\frac{1}{n}\Sigma$ be the variance of $\hat{r}$. It is straightforward to show that the optimal estimator of θ is $(M'\Sigma^{-1}M)^{-1}M'\Sigma^{-1}\hat{r}$ in the following sense: Let θ be estimated by minimization of the Euclidian norm of $A\hat{r} - AM\theta$ where A is an arbitrary matrix. The choice of A which minimizes the variance of $\hat{\theta}$ is $A = \Sigma^{-1/2}$ and for all A, $\hat{\theta}$ is an unbiased estimator. This GMM problem can be extended to the infinite dimensional case. The problem is then rewritten as $r = K\varphi$ where r is an infinite set of moments, φ is an element of a functional space and K is a linear operator. Then the estimate of φ is given by the argument that minimizes the following: $\|A\hat{r} - AK\varphi\|^2$ where A is a chosen linear operator. However it is well-known in the nonparametric IV literature that if r is estimated by $\hat{r}$ this minimization problem is ill-posed and the minimization does not give a consistent estimator; see Carrasco, Florens, and Renault (2007). Among many solutions the Tikhonov Regularization provides a good solution to this problem where the minimization is modified by an L^2 penalty. Then the estimator minimizes $\|A\hat{r} - AK\varphi\|^2 + \alpha\|L\varphi\|^2$ where L is also a suitably chosen linear operator and α is the regularization parameter. For any A and L, the mean integrated squared error, MISE, $E(\|\hat{\varphi}_\alpha - \varphi\|^2)$ can be computed and an optimal value of α (α_{opt}) can be selected. One may then compute optimal MISE, $E(\|\hat{\varphi}_{\alpha_{opt}} - \varphi\|^2)$

which depends on the choice of A and L. Hence the optimality question becomes: *What is the operator A which minimizes the MISE?* In other words, does the variance operator $\Sigma^{-1/2}$ lead to the most efficient estimator? In this paper we show that the MISE is not minimized for the choice of $A = \Sigma^{-1/2}$ as it is in the finite dimensional case, even if the estimator is well defined for this particular choice.

Infinite dimensional GMM considered in this paper, i.e., the parameter of interest belongs to an infinite dimensional functional space, is closely related to nonparametric instrumental variables literature, see Darolles, Fan, Florens, and Renault (2011); Newey and Powell (2003); Ai and Chen (2003) and Horowitz (2011) among others. Darolles, Fan, Florens, and Renault (2011) use a Tikhonov regularized kernel based estimator while Newey and Powell (2003); Ai and Chen (2003) and Horowitz (2011) use sieve minimum distance (SMD) estimator. All these mentioned papers show that the estimators they use is consistent however none of them considers efficiency.

This paper may also be viewed as an extension to the usual GLS method to infinite dimensional case. In GLS approach optimal estimators are obtained by weighting the sum of squares by the inverse of the variance of the residual. We will later show that this property is no longer true if the infinite dimension requires a penalisation.

With the growth of nonparametic IV literature in the recent years, the attention has also been given to models that are semiparametric, i.e., the parameter of interest is composed of an infinite dimensional function as well as a finite dimensional vector. Florens, Johannes, and Bellegem (2012); Ai and Chen (2003); Chen and Pouzo (2009) considers the nonparametric estimation of these semiparametric models. Ai and Chen (2003) and Chen and Pouzo (2009) focus on the efficiency of the estimator of the finite dimensional parameter and show that it reaches semiparametric efficiency bound when the weighting matrix is equal to the inverse of variance covariance matrix of moment conditions. To the best of our knowledge efficiency of the nonparametric estimator in terms of mean integrated squared error (MISE) has only been considered by Gagliardini and Scaillet (2012) within the framework of Tikhonov regularised nonparametric IV estimation. However, they do not investigate the optimality of the estimator with respect to the choice of weighting matrix.

Moreover, the use of nonparametric techniques in structural models has increased the interest in nonparametric estimation of simultaneous equations. Most recent work on this topic comes from Berry and Haile (2014) where they show the identification without the use of completeness condition and Matzkin (2015) which introduces an easy to implement estimation technique for nonparametric, nonadditive simultaneous equations models. We believe that the light we shed on the optimal weighting matrix in linear inverse problems will also contribute to nonparametric estimation of simultaneous equations by leading to development of techniques such as nonparametric three stage least squares.

Apart from nonparametric IV literature, this paper also relates to the literature on GMM with finite dimensional parameter of interest and with a continuum of moment conditions, see Carrasco and Florens (2000, 2014). Both of these papers consider a continuum of moment conditions however the parameter of interest is a finite dimensional vector. In such a case Carrasco and Florens (2000) show that the optimal weighting matrix is not invertible and this leads to an ill-posed inverse problem in the estimation. Hence, they propose to use a regularized inverse. Carrasco and Florens (2014) show that this GMM estimator with a continuum of moment conditions which uses the regularized inverse of the optimal weighting matrix reaches the efficiency of the MLE. Note that although seems similar, the problem we investigate in this paper is different since in our case the ill-posed inverse problem is not caused by the choice of the weighting matrix but it is the result of the estimation problem itself.

From a mathematical viewpoint, the weighting problem can be considered as follows. The equation $r = K\varphi$ is an integral equation. If A is an integral operator (Σ at some positive power for example) then K becomes AK which has a larger degree of ill-posedness. Then an intuitive approach would be to select A as a differential operator (such as $\Sigma^{-1/2}$) in order to reduce the degree of ill-posedness. If Ar is defined, weighting means a derivation of the equation before its resolution. However, the impact of A is not so clear if we select the regularization parameter in an optimal way. For example, the rate of decline of the bias is lower for an integral operator A but the optimal values of α is smaller and the effect is then ambiguous. Up to our knowledge, a theoretical conclusion on optimal weighting is not tractable and we have adopted a numerical analysis. This analysis is not a Monte Carlo simulation; for a given design, we have an explicit form of the MISE and we optimize numerically with respect to α and A.

The paper proceeds as follows. In *Section 2* we introduce our model. In *Sections 3* and *4* we examine the optimisation of the MISE with geometric and exponential spectra respectively. In *Section 5*, we look for a solution to the problem of optimality introduced in *Sections 3* and *4* by means of simulation. In *Section 6* we look at the optimal weighting problem in non-parametric IV case. Finally, in *Section 6* we conclude.

2 The Model

Let us consider a linear inverse problem in the form:

$$\hat{r} = K\varphi + U$$

$\varphi \in \mathcal{E}$ and $\hat{r}$ and $U \in \mathcal{F}$ where $\mathcal{E}$ and $\mathcal{F}$ are Hilbert spaces. The operator $K : \mathcal{E} \mapsto \mathcal{F}$ is a compact operator and U is a random element in $\mathcal{F}$ such that $\mathbb{E}(U) = 0$ and $\mathbb{V}(U) = \frac{1}{n}\Sigma$ where n is the sample size. The value $\hat{r}$ is a noisy observation of $r = K\varphi$ with a variance of $\frac{1}{n}\Sigma$. The element $\hat{r}$ is observed and K and Σ are given (possibly estimated). Let L be a differential operator defined on $\mathcal{E}$ (L is densely defined, self adjoint and L^{-1} is a compact operator from $\mathcal{E} \mapsto \mathcal{E}$). We also consider a weighting operator A which is compact. In case of a well-posed inverse problem, the strategy would be to minimize $\|A\hat{r} - AK\varphi\|^2$ and the optimal choice is to take $A^*A = \Sigma^{-1}$. Let us consider the general ill-posed inverse problem. In the sequel, $\mathcal{D}(S)$ and $\mathcal{R}(S)$ denote respectively the domain and range of an operator S.

Assumption 1 *L is chosen such that $\varphi \in \mathcal{D}(L)$.*

Under this assumption the Tikhonov estimator using a Hilbert scale penalty is defined as the solution of:

$$\min \|A\hat{r} - AK\varphi\|^2 + \alpha\|L\varphi\|^2 \tag{1}$$

and it is equal to:

$$\begin{aligned} \hat{\varphi}_\alpha &= L^{-1}(\alpha I + L^{-1}K^*A^*AKL^{-1})^{-1}L^{-1}K^*A^*A\hat{r} & (2) \\ &= (\alpha L^{-2} + K^*A^*AK)^{-1}K^*A^*A\hat{r} & (3) \end{aligned}$$

see Florens and Van Bellegem (2015). Let us introduce the following assumption:

Assumption 2 *K, L^{-1} and A have the same countable family of singular vectors denoted $(\varphi_j)_{j=1}^{\infty}$ and their singular values are λ_{Kj}, $\lambda_{L^{-1}j}$ and λ_{Aj}, respectively.*

Using these notation, Assumption 2 is equivalent to:

$$\sum_{j=1}^{\infty} \frac{\langle \varphi, \varphi_j \rangle^2}{\lambda_{L^{-1}j}^2} < \infty$$

Under these assumptions we may compute the mean square error of $\hat{\varphi}_\alpha$.

Proposition 1 *The MISE of $\hat{\varphi}_\alpha$ is given by:*

$$\mathbb{E}\|\hat{\varphi}_\alpha - \varphi\|^2 = \frac{1}{n} \sum_{j=1}^{\infty} \frac{\langle \Sigma\varphi_j, \varphi_j \rangle^2 \lambda_{Kj}^2 \lambda_{Aj}^4 \lambda_{L^{-1}j}^4}{(\alpha + \lambda_{Kj}^2 \lambda_{Aj}^4 \lambda_{L^{-1}j}^4)^2} + \alpha^2 \sum_{j=1}^{\infty} \frac{\langle \varphi, \varphi_j \rangle^2}{(\alpha + \lambda_{Kj}^2 \lambda_{Aj}^4 \lambda_{L^{-1}j}^4)^2} \tag{4}$$

Proof.

$$\mathbb{E}\|\hat{\varphi}_\alpha - \varphi\|^2 = tr[\mathbb{V}(\hat{\varphi}_\alpha)] + \|\varphi_\alpha - \varphi\|^2$$

where $\varphi_\alpha = L^{-1}(\alpha I + L^{-1}K^*A^*AKL^{-1})^{-1}L^{-1}K^*A^*AK\varphi$. Using some elementary manipulations and the property that L^{-1} commute with K^*A^*AK we get:

$$\mathbb{E}\|\hat{\varphi}_\alpha - \varphi\|^2 = \frac{1}{n}tr[(\alpha I + B)^{-1}L^{-1}K^*A^*A\Sigma A^*AKL^{-1}(\alpha I + B)^{-1}] + \|\alpha(\alpha I + B)^{-1}\varphi\|^2$$

where $B = L^{-1}K^*A^*AKL^{-1}$. Using the property that $tr(\Omega) = \sum_{j=1}^{\infty}\langle\Omega\varphi_j, \varphi_j\rangle$, we get the result. ■

One can remark that A and L^{-1} play the same role because only the product $\lambda_{Aj}\lambda_{L^{-1}j}$ appears in the MISE formula. Then the same value may be obtained either by weighting by A or by penalizing by LA^{-1}. Hence, in the following we just consider A but our result may be reinterpreted in terms of Hilbert scale penalisation.

The estimation strategy which minimizes the risk measured by the MISE consists of the choice of α and A which minimise $\mathbb{E}\|\hat{\varphi}_\alpha - \varphi\|^2$ at n, K and Σ fixed. Actually, this computation cannot be done analytically up to our knowledge so we will perform this optimisation numerically. We will adopt two frameworks: one with the geometric decline of spectrums and the other with the exponential decline. In the first case the problem is mildly ill-posed while it is severely ill-posed in the second case. However, we consider the case that the true function is analytic which preserves polynomial rates of convergence.

We conclude this section by examples:

Example 1 (Density Estimation): Let $X \in [0, 1]$ be a real random variable with a cdf denoted by F and a density denoted by f. Both F and f are assumed to be the elements of $L^2_{[0,1]}$ provided with the uniform distribution. If an iid sample of X is available, we denote by $\hat{F}$ the empirical cdf. The model can be written:

$$\hat{F}(t) = \int_0^t f(t)dt + U$$

where the parameter of interest is now f, K is the integral operator and Σ is characterized by:

$$(\Sigma g)(s) = \int_0^1 [F(s \wedge t) - F(s)F(t)]g(t)dt$$

Example 2 (Functional Linear IV Model): Let $Y \in \mathbb{R}$, $Z \in \mathcal{E}$ and $W \in \mathcal{F}$ three random elements such that $Y = \langle Z, \varphi \rangle + \varepsilon$ and $E(\varepsilon W) = 0$, see Florens and Van Bellegem (2015). If an iid sample is observed, we can define $r = E(YW) \in \mathcal{F}$, $\hat{r} = \frac{1}{n} \sum y_i w_i \in \mathcal{F}$ and $K\varphi = \frac{1}{n} \sum w_i \langle z_i, \varphi \rangle$. The model is then written $\hat{r} = K\varphi + U$ and U is a zero mean random element (conditional on w_i) with a conditional variance equal to $\frac{1}{n}\Sigma$. The operator Σ is characterized by:

$$\Sigma g = \frac{1}{n} \sum w_i \langle w_i, g \rangle Var(\varepsilon | w_i)$$

3 MISE Optimisation in the case of geometric spectrum

Let us underline that we do not proceed to a Monte Carlo simulation but we will evaluate numerically equation (4) and compute its optimal value for α and A (in a given parametrized family). In order to evaluate equation (4) we fix different elements in the following way:

$$\lambda_{Kj} = \frac{1}{j^a}, \quad \lambda_{Aj} = \frac{1}{j^d} \quad \langle \Sigma\varphi_j, \varphi_j \rangle = \frac{1}{j^{2c}}, \quad \langle \varphi, \varphi_j \rangle = \frac{1}{j^b}$$

where a, b, c are strictly larger than $\frac{1}{2}$. The spectrum of A may be characterized by $d < 0$ which is to say that A is a differential operator.

In this framework, equation (4) becomes:

$$\mathbb{E}\|\hat{\varphi}_\alpha - \varphi\|^2 = \frac{1}{n} \sum_{j=1}^{\infty} \frac{j^{2(a-c)}}{(\alpha j^{2(a+d)} + 1)^2} + \alpha^2 \sum_{j=1}^{\infty} \frac{j^{4(a+d)-2b}}{(\alpha j^{2(a+d)} + 1)^2} \tag{5}$$

We will optimize numerically this risk with respect to α and d.

Even if the asymptotic analysis is not our objective, we may give an approximation of equation (5) from which the rate may be analysed. Note that the first term on the right hand side is the variance and the second term is the bias. Let us first consider the variance term. Suppose that there exists a function $h(x)$:

$$h(x) = \frac{x^{2(a-c)}}{(\alpha x^{2(a+d)} + 1)^2}$$

if we approximate the sum with the integral we can write, see Florens and Simoni (2014):

$$= \frac{1}{n} \int_1^\infty \frac{x^{2(a-c)}}{(\alpha x^{2(a+d)} + 1)^2} dx$$

After making the following change of variable to evaluate the integral:

$$y = \alpha x^{2(a+d)} + 1 \Rightarrow \left(\frac{y}{\alpha}\right)^{\frac{1}{2(a+d)}} = x$$

we can write:

$$= \frac{1}{n} \left(\frac{1}{\alpha}\right)^{\frac{2(a-c)+1}{2(a+d)}} \frac{1}{2(a+d)} \int_0^\infty \frac{y^{\frac{-2d-2c+1}{2(a+d)}}}{(y+1)^2} dy$$

the following condition is needed for the integral to converge:

$$1 - 2(c+d) < 2(a+d) \Rightarrow 2(a+c) + 4d > 1$$

Equivalently, for bias term, given the following function $g(x) = \frac{x^{4(a+d)-2b}}{(\alpha x^{2(a+d)}+1)^2}$, and under the same change of variable, we can approximate the sum with the following integral:

$$= \frac{1}{2(a-d)} \alpha^{\frac{2b-1}{2(a-d)}} \int_0^\infty \frac{y^{\frac{2(a-d)-2b+1}{2(a-d)}}}{(y+1)^2} dy$$

Then the condition for integral to converge is given by:

$$\frac{2(a+d)-2b+1}{2(a+d)} < 1 \Rightarrow b > 1/2$$

which is already satisfied by the previous assumptions.

One can now obtain the optimal regularization parameter, α_{opt}, by minimizing the MISE with respect to α. Call the integral in the variance term V and the integral in the bias term B. Then the MISE can be written as the following:

$$MISE = \frac{1}{n}\left(\frac{1}{\alpha}\right)^{\frac{2(a-c)+1}{2(a+d)}} \frac{1}{2(a+d)}V + \frac{1}{2(a+d)}\alpha^{\frac{2b-1}{2(a+d)}}B$$

$$\frac{\partial MISE}{\partial \alpha} = -\frac{1}{n}\frac{1}{2(a+d)}\frac{2(a-c)+1}{2(a+d)}\alpha^{1+\frac{2(a-c)+1}{2(a+d)}}V + \frac{1}{2(a+d)}\frac{2b-1}{2(a+d)}\alpha^{1-\frac{2b-1}{2(a+d)}}B = 0$$

Then the optimal α, α_{opt} is given by:

$$\alpha_{opt} = \left[\frac{1}{n}\frac{V}{B}\frac{2(a-c)+1}{2b-1}\right]^{\frac{2(a+d)}{2(a-c)+2b}} \tag{6}$$

If one replaces α_{opt} back in the MISE:

$$MISE = \frac{1}{2(a+d)}\left(\frac{1}{n}\right)^{\frac{2b-1}{2(a-c)+2b}} V^{\mu}B^{1-\mu}(R_1+R_2) \tag{7}$$

where $\mu = \frac{2b-1}{2(a-c)+2b}$, $R_1 = \left[\frac{2(a-c)+1}{2b-1}\right]^{\frac{-(2(a-c)+1)}{2(a-c)+2b}}$ and $R_2 = \left[\frac{2(a-c)+1}{2b-1}\right]^{\frac{2b-1}{2(a-c)+2b}}$. Note that μ, R_1 and R_2 do not depend on d. Moreover the speed of convergence will be captured by the term $\left(\frac{1}{n}\right)^{\frac{2b-1}{2(a-c)+2b}}$ which does not depend on d either. So, we are left with the following term:

$$\frac{1}{2(a+d)}\left[\int_0^{\infty}\frac{y^{\frac{-2d-2c+1}{2(a+d)}}}{(y+1)^2}dy\right]^{\mu} \times \left[\int_0^{\infty}\frac{y^{\frac{2(a+d)-2b+1}{2(a+d)}}}{(y+1)^2}dy\right]^{1-\mu} \tag{8}$$

We do not use this approximation for our simulation and we go back to equation (5) to optimize with respect to α and d.

4 MISE Optimisation in the case of exponential spectrum

Let us now consider MISE formula given in equation (4) if all the elements of the series decline exponentially. We assume that for a given $\rho \in]0,1[$ we have:

$$\lambda_{Kj} = \rho^j, \quad \lambda_{Aj} = \rho^{j\delta}, \quad \langle \Sigma\varphi_j, \varphi_j \rangle = \rho^{2j\mu}, \quad \langle \varphi, \varphi_j \rangle = \rho^{j\beta}$$

Then the MISE formula becomes:

$$\mathbb{E}\|\hat{\varphi}_\alpha - \varphi\|^2 = \frac{1}{n}\sum_{j\geq 1}^{\infty} \frac{\rho^{2(1+\mu+2\delta)j}}{(\alpha + \rho^{2(1+\delta)j})^2} + \alpha^2 \sum_{j\geq 1}^{\infty} \frac{\rho^{2\beta j}}{(\alpha + \rho^{2(1+\delta)j})^2} \tag{9}$$

Let us consider the rate of convergence of the MISE. Using the same integral approximation as before we get:

$$MISE = \frac{1}{n}\int_0^\infty \frac{\rho^{2(1+\mu+2\delta)x}}{(\alpha + \rho^{2(1+\delta)x})^2}dx + \alpha^2 \int_0^\infty \frac{\rho^{2\beta x}}{(\alpha + \rho^{2(1+\delta)x})^2}dx$$

If we make the following change of variable:

$$y = \frac{\rho^{2(1+\delta)x}}{\alpha}$$

The integral approximation of MISE will be given by:

$$= \frac{1}{\alpha^2}\frac{1}{(1+\delta)ln\rho^2}\int_0^\infty \frac{1}{y(1+y)^2}\rho^{\frac{2(1+\mu+2\delta)}{(1+\delta)ln\rho}(ln\alpha + lny)}dy + \frac{1}{(1+\delta)ln\rho^2}\int_0^\infty \frac{1}{y(1+y)^2}\rho^{\frac{2\beta}{(1+\delta)ln\rho^2}(ln\alpha + lny)}dy$$

After some manipulation we can write

$$\frac{1}{n\alpha^2}\alpha^{\frac{1+\mu+2\delta}{1+\delta}}V + \alpha^{\frac{\beta}{1+\delta}}B$$

where V and B are suitable constants coming from the bias and variance term, respectively. We assume that $\frac{1+\mu+2\delta}{1+\delta} < 2$ and $\frac{\beta}{1+\delta} < 2$. Then the optimal value of α is then proportional to $n^{-\frac{1+\delta}{\beta+1-\mu}}$ and the rate of convergence is proportional to $n^{-\frac{\beta}{\beta+1-\mu}}$. As in the geometric case, the rate of convergence does not depend on δ, the degree of ill-posedness of A.

5 Simulations

In this section we explain and present our numerical simulations. In the geometric case we minimize the MISE given in equations (5) numerically by first picking the optimal value of α. Then the optimal weighting will be characterised by parameter d which minimizes the $MISE(\alpha_{opt})$. In the exponential case we minimize the MISE given in equation (9) again by with respect to α and then with respect to δ to get the optimal weighting.

5.1 Simulations in Geometric Case

To optimize the MISE numerically, we generate $j = 1, ..., 10^6$ and we compute the MISE given in equation (5) where we fix the sample size n to 100. Then for each simulation we pick different values of a, b and c and compute the MISE given a grid of α and d. In other words, for fixed values of n,a,b and c, we take a grid of α and a grid of d and compute MISE for each value of α and d on these grids. Then the optimal value of regularization parameter and the optimal weighting are given by those α and d which lead to the minimum value of MISE.

First of all, Figure (1) shows how MISE changes with α for given values of d, whereas the left panel of figure (2) shows how bias square changes with α for given values of d and the right panel shows the same for the variance. Bias increases and variance decreases with

α as expected. A stronger regularization, i.e., a larger α increases the regularization bias while making the estimation more efficient.

Figure (3) shows the MISE vs. d given that α is optimal. For a simulation design where a, b and c are set to be equal to 5,1 and 1.5 respectively and d is a grid of 40 equidistant points between -2.5 and 2, figure (3) shows that the MISE is minimized at $d = -0.4231$. Moreover, variance has a decreasing trend with d whilst the bias square has an increasing trend, see figures 4 and 5. Given these results we see that the MISE is not minimized for $d = -c$. So the optimality condition for the finite dimensional case does not hold in the infinite dimensional case. However, to see if we can find a rule for the optimal value of d, we try several designs for different values of a, b and c; d is always obtained to be equal to $b - c$ and the optimal value of d does not depend on the value of a since we get exactly the same results whatever is the value of a. Depending on the values of b and c, d can be positive or negative. In other words, depending on the regularity of the function and the degree of ill-posedness of the variance operator, the optimal weighting can be given by a differential or an integral operator. Tables 1 and 2 present the values of optimal values of α and d obtained under different designs.

Table 1: Simulation results in geometric case

$a = 5$, $c = 1.5$, $n = 100$				
b	0.5	1	2	3
d	-1	-0.4872	0.5385	1.5641
α	0.01	0.0098	0.0096	0.0094

Table 2: Simulation results in geometric case

$a = 5$, $c = 2$, $n = 100$				
b	0.5	1	2	3
d	-1.4872	-0.9744	0.05	1.0769
α	0.0098	0.0096	0.0094	0.0092

Figures 6-8 show the 3 dimensional surfaces for α and d and MISE, variance and bias terms respectively. We pick 40 equidistant points in $[-3, 2]$ for the grid of d and 50 equidistant points in $[10^{-4}, 10^{-2}]$ for the grid of α. Moreover we set the following values for the parameters: $a = 5$, $b = 1$ and $c = 1.5$. The optimal value of d is obtained to be -0.4872 while the optimal value of α is obtained to be 0.0098.

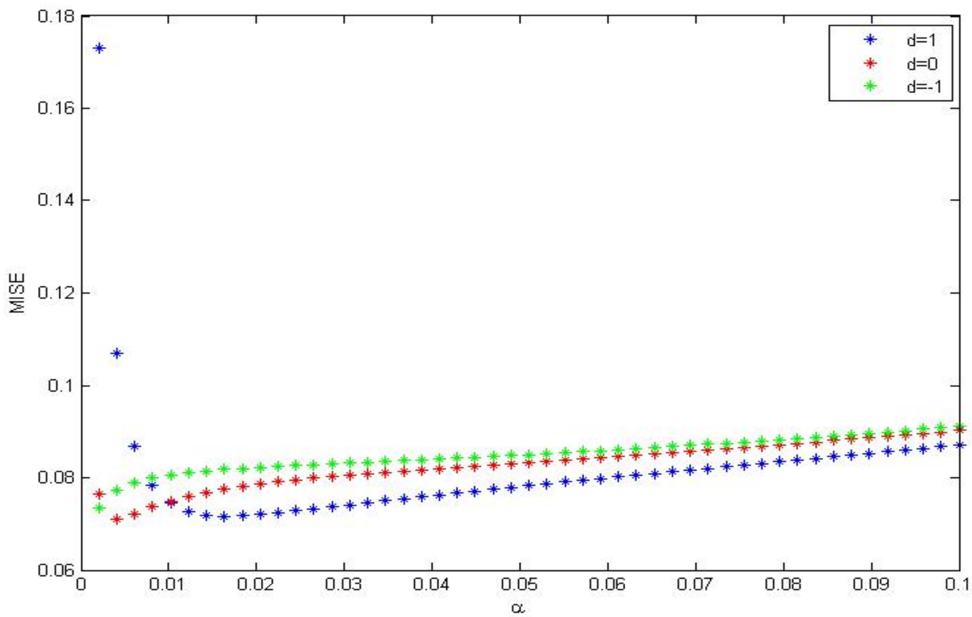

Figure 1: *MISE vs. α for different values of d*

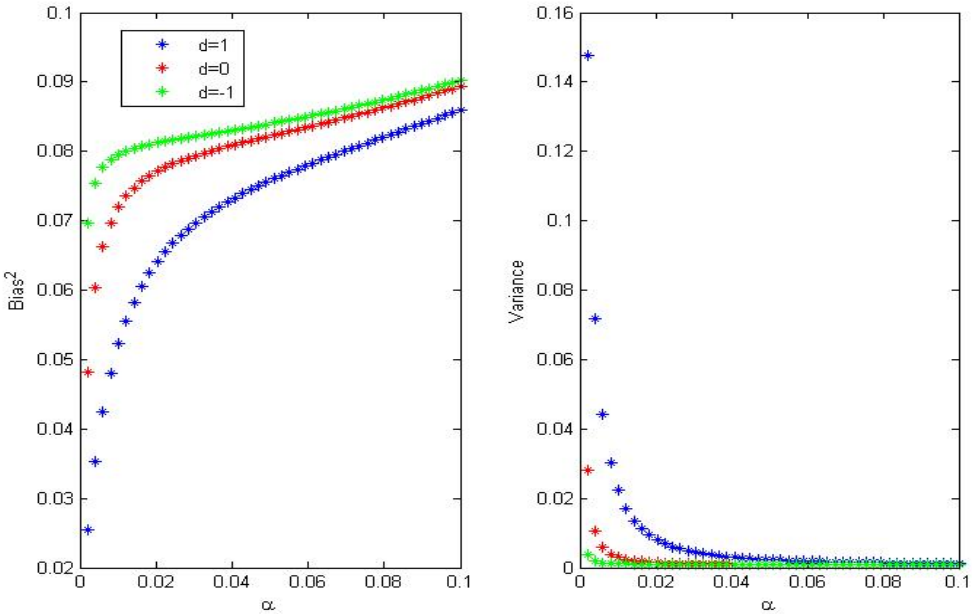

Figure 2: *Left panel: Bias^2 vs. α for different values of d. Right panel: Variance vs. α for different values of d*

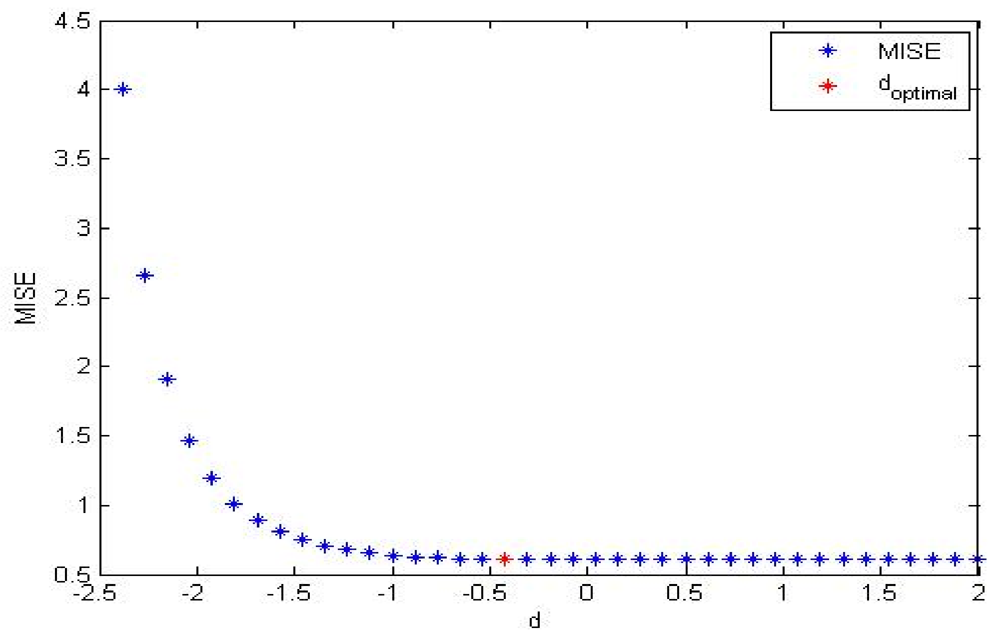

Figure 3: *MISE vs. d given that $\alpha = \alpha_{opt}$, $b = 1$, $c = 1.5$*

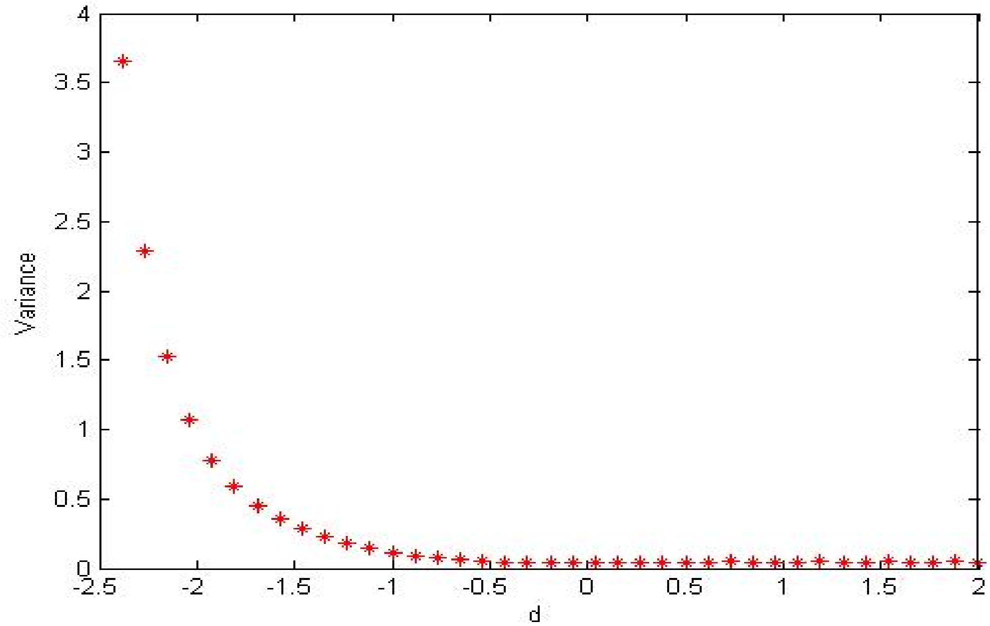

Figure 4: *Variance vs. d given that* $\alpha = \alpha_{opt}$, $b = 1$, $c = 1.5$

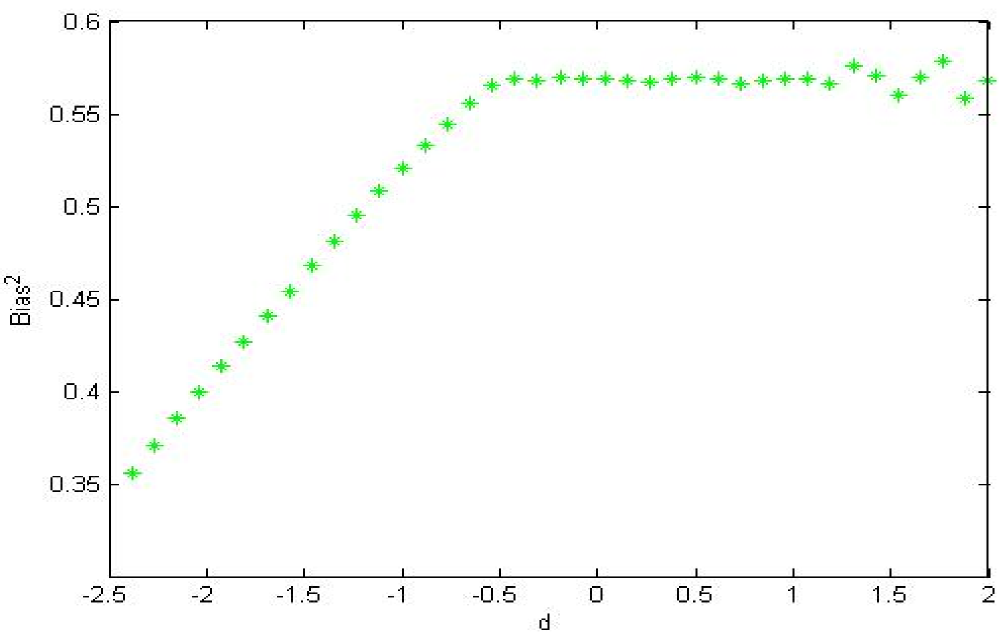

Figure 5: $Bias^2$ *vs. d given that* $\alpha = \alpha_{opt}$, $b = 1$, $c = 1.5$

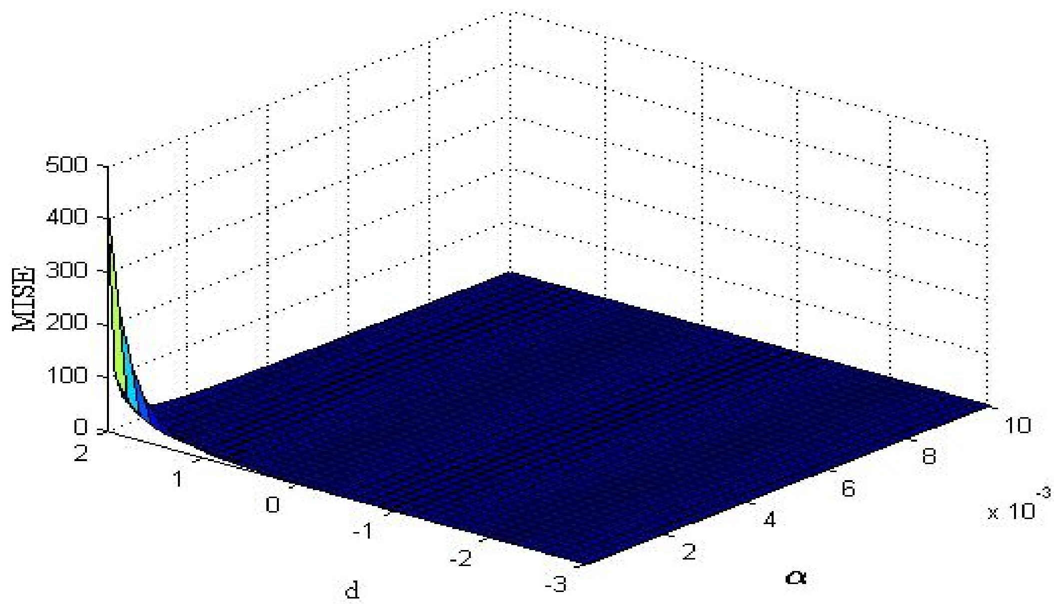

Figure 6: *MISE vs. d and* α, $b = 1$, $c = 1.5$, $\alpha \in [10^{-4}, 10^{-2}]$

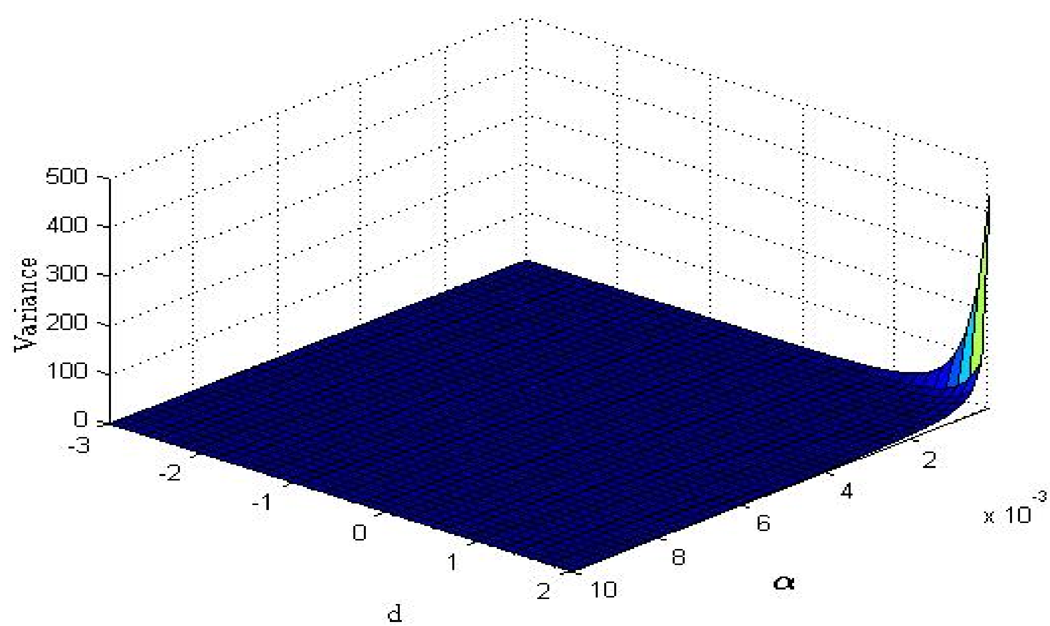

Figure 7: *Variance vs. d and* α, $b = 1$, $c = 1.5$, $\alpha \in [10^{-4}, 10^{-2}]$

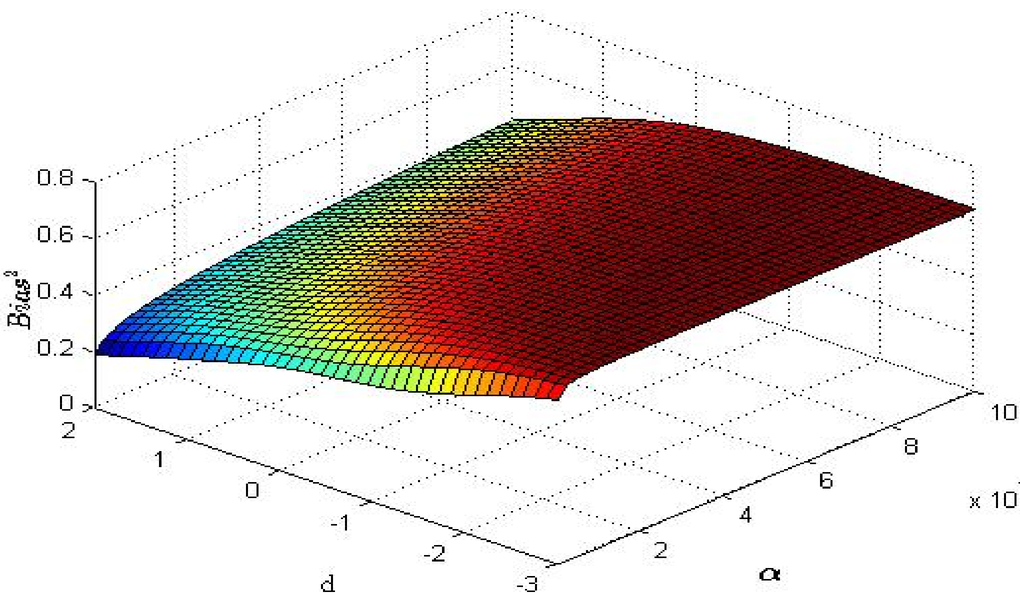

Figure 8: $Bias^2$ *vs.* d *and* α, $b = 1$, $c = 1.5$, $\alpha \in [10^{-4}, 10^{-2}]$

5.2 Simulations in Exponential Case

We conduct numerical simulations also in the exponential case to see if we can still get the same result, i.e., the optimal weighting depends on the regularity of the function and the degree of ill-posedness of the variance operator.

We minimize equation (9) with respect to α and δ. We generate $j = 1, ..., 500$ and set $\rho = 0.6$. As in the geometric case we then optimize MISE for different values of β and μ. In this case too, we find that the optimal weighting depends on both β and μ. More precisely, the optimal delta is equal to $\beta - \mu$. Our results are presented in Tables 3 and 4.

Table 3: Simulation results in exponential case

$\rho = 0.6$, $\mu = 1.5$, $n = 100$				
b	0.5	1	2	3
d	-0.6735	-0.5102	0.5510	1.5714
α	2.14×10^{-4}	0.01	0.009	0.009

Figures 9-11 show the 3 dimensional surfaces for α and δ and MISE, variance and bias

Table 4: Simulation results in exponential case

$\rho = 0.6,\ \mu = 1,\ n = 100$				
b	0.5	1	2	3
d	-0.4694	0.0408	1.0612	1.9796
α	0.008	0.0086	0.009	0.01

terms respectively. We pick 50 equidistant points in $[-2, 3]$ for the grid of δ and 50 equidistant points in $[10^{-5}, 10^{-2}]$ for the grid of α. Moreover we set the following values for the parameters:$\beta = 2$ and $\mu = 1$. The optimal value of δ is obtained to be 1.0612 while the optimal value of α is obtained to be 0.009.

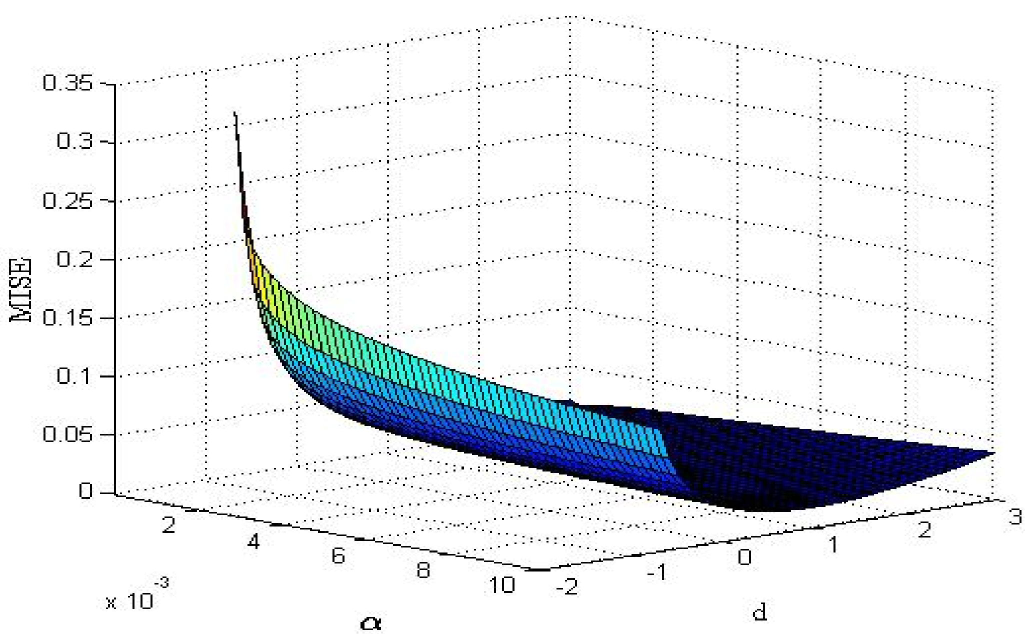

Figure 9: *MISE vs. d and* α, $\beta = 2$, $\mu = 1$, $\alpha \in [10^{-5}, 10^{-2}]$

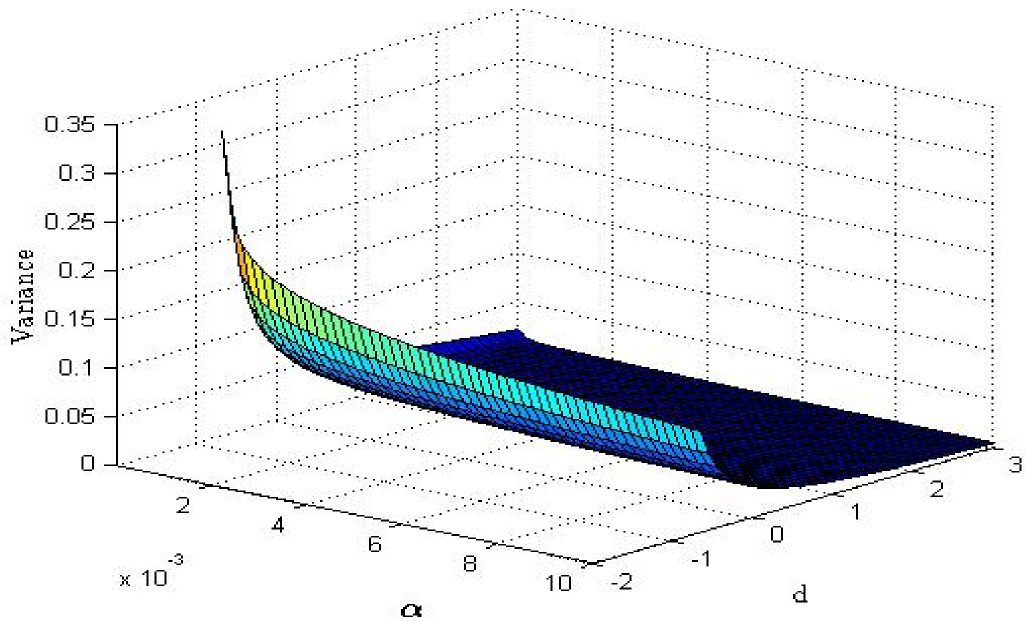

Figure 10: *Variance vs. d and α, $\beta = 2$, $\mu = 1$, $\alpha \in [10^{-5}, 10^{-2}]$*

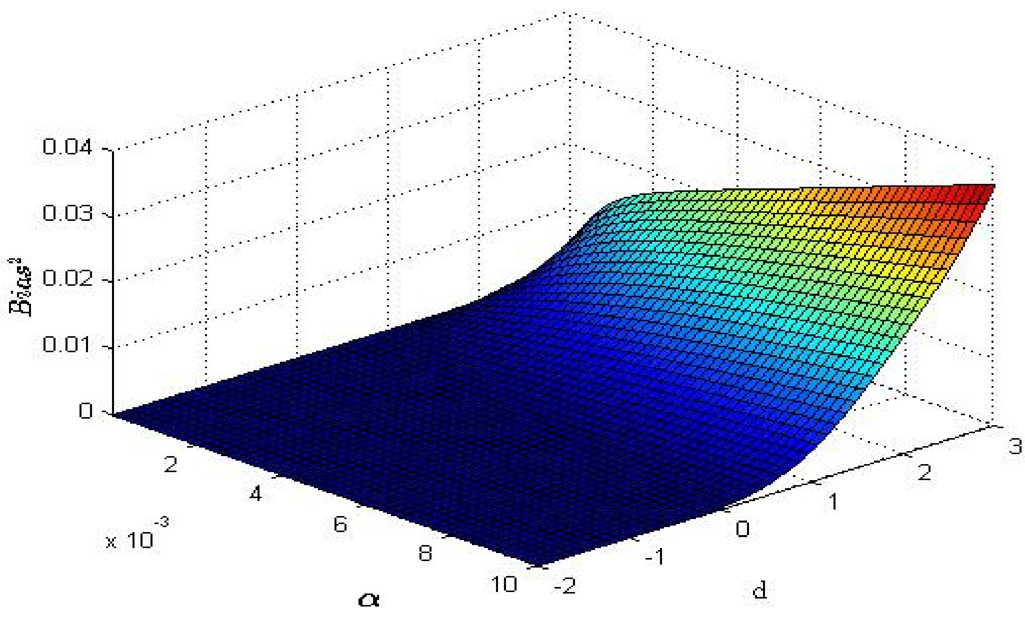

Figure 11: *$Bias^2$ vs. d and α, $\beta = 2$, $\mu = 1$, $\alpha \in [10^{-5}, 10^{-2}]$*

6 The Nonparametric IV Case

We consider the usual NPIV case as developed now in many papers; see Carrasco, Florens, and Renault (2007); Darolles, Fan, Florens, and Renault (2011); Hall and Horowitz (2005) among others. We consider a vector of random elements (Y, Z, W) such that:

$$Y = \varphi(Z) + U \quad and \quad \mathbb{E}(U|W) = 0 \tag{10}$$

The model then generates a linear inverse problem:

$$\begin{aligned} \mathbb{E}(\mathbb{E}(Y|W)|Z) &= \mathbb{E}(\mathbb{E}(\varphi(Z)|W)|Z) \quad (11) \\ r &= K\varphi \quad (12) \end{aligned}$$

where $r \in L^2_Z$, $\varphi \in L^2_Z$ and $K : L^2_Z \mapsto L^2_Z$. We assume that all the L^2 spaces are related to the true distribution. We have a noisy observation of r, $\hat{r}$ but we assume that K is given, then we write:

$$\hat{r} = K\varphi + U \tag{13}$$

We may assume that $\mathbb{E}(U) = 0$ and $\mathbb{V}(U) = \sigma^2 K$. The operator K is a self-adjoint trace class operator. This situation is discussed in many previous papers, see Carrasco, Florens, and Renault (2007); Darolles, Fan, Florens, and Renault (2011).

Let us discuss first the rate of convergence of the Tikhonov estimator of equation (13). In this case Σ is proportional to K. If the spectrum is geometrical, this means that $c = \frac{a}{2}$ and the rate is $n^{-\frac{2b-1}{2b+a}}$. If the spectrum is exponential as in *Section 5*, $\mu = \frac{1}{2}$ and the rate is $n^{-\frac{\beta}{\beta+1/2}}$. In order to compare these rates with the ones discussed by Chen and Reiss (2011), let us consider for example the geometric case. In their approach, if φ has the regularity ν, i.e., if φ is ν times differentiable, $\mathbb{E}(Y|W)$ has the regularity $\nu+p$ where p is the degree of ill-posedness of $\mathbb{E}(\varphi(Z)|W)$. Then if the dimension of W is equal to 1, the rate of convergence is given by $\left(n^{-\frac{2(\nu+p)}{2(\nu+p)+1}}\right)^{\frac{\nu}{\nu+p}} = n^{-\frac{2\nu}{2(\nu+p)+1}}$. In our framework, $\nu = b - 1$ and $p = a$, then we can rewrite their rate as $n^{-\frac{2b-2}{2b+2a+1}}$. Hence, our rate is faster than the rate obtained by Chen and Reiss (2011). Note that, this result does not contradict with the minimax result because it incorporates a restriction on the estimation of $\mathbb{E}(\mathbb{E}(Y|W)|Z)$ which is to have a variance equal to $\sigma^2 K$. Even if the spectrum is exponential, if the Fourier coefficients of φ also decline exponentially, the variance property of U implies a very good rate for the MISE ($n^{-\frac{2\beta}{2\beta+1}}$).

Let us now consider our problem of optimal weighting. We consider the class of estimators:

$$\hat{\varphi}_{\alpha,d} = (\alpha I + K^{2(d+1)})^2 K^{2d+1} \hat{r}$$

Note that the case $d = \frac{1}{2}$ (weighting by $\Sigma^{-1/2}$) corresponds to the usual estimator of NPIV, solution of:

$$\alpha\varphi + \mathbb{E}(\mathbb{E}(\varphi|W)|Z) = \mathbb{E}(\mathbb{E}(Y|W)|Z)$$

We will show by simulation that the MISE may be improved by choosing a larger value of d, if α is taken at its optimal value.

7 Conclusion

This paper presents some analytical computations and numerical simulations for the mean squared error of the estimation of a function φ deduced from the program:

$$\min \|A\hat{r} - AK\varphi\|^2 + \alpha\|L\varphi\|^2$$

The function $\hat{r}$ is observed and the operator K and the variance of $\hat{r} - K\varphi$ are given. We discuss the choice of the tuning parameters α, L and A. Our main conclusions are the following:

- By analytical computations, we show that A and L^{-1} play the same role and that the product AL^{-1} only matters. We may then fix one of them equal to the identity.

- From the same analytical results, we are also able to compute to rate of decline of MISE taking into account the knowledge of Σ. This assumption allows us to improve the usual minimax rates of convergence (obtained without any constraint on Σ).

- The choice of the optimal weighting operator is only obtained numerically. Intuitively, the optimal choice depends on the regularity of φ and on Σ. A conjecture would be to equalize the regularity of φ and the sum of the degree of ill-posedness of A and Σ. Then the choice $A = \Sigma^{-1/2}$ (or $L = \Sigma^{1/2}$ an integral operator) corresponds to a completely irregular φ. If φ is sufficiently regular, A becomes an integral operator or L a differential operator. This last result is relatively intuitive: if φ is sufficiently smooth regarding to Σ, a penalization by the norm of the derivative is optimal.

References

Ai, C., and X. Chen (2003): "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions," *Econometrica*, 71(6), 1795–1843.

Berry, S. T., and P. A. Haile (2014): "Identification of Nonparametric Simultaneous Equations Models with a Residual Index Structure," *mimeo*.

Carrasco, M., and J.-P. Florens (2000): "Generalization Of GMM To A Continuum Of Moment Conditions," *Econometric Theory*, 16(06), 797–834.

——— (2014): "On The Asymptotic Efficiency Of GMM," *Econometric Theory*, 30(02), 372–406.

Carrasco, M., J.-P. Florens, and E. Renault (2007): "Linear Inverse Problems in Structural Econometrics Estimation Based on Spectral Decomposition and Regularization," in *Handbook of Econometrics*, ed. by J. Heckman, and E. Leamer, vol. 6 of *Handbook of Econometrics*, chap. 77. Elsevier.

Chen, X., and D. Pouzo (2009): "Efficient Estimation of Semiparametric Conditional Moment Models with Possibly Nonsmooth Residuals," *Journal of Econometrics*, 152(1), 46–60.

Chen, X., and M. Reiss (2011): "On Rate Optimality For Ill-Posed Inverse Problems In Econometrics," *Econometric Theory*, 27(03), 497–521.

Darolles, S., Y. Fan, J.-P. Florens, and E. Renault (2011): "Nonparametric Instrumental Regression," *Econometrica*, 79(5), 1541–1565.

Florens, J.-P., J. Johannes, and S. V. Bellegem (2012): "Instrumental regression in partially linear models," *Econometrics Journal*, 15(2), 304–324.

Florens, J.-P., and A. Simoni (2014): "Regularising Priors for Linear Inverse Problems," *Econometric Theory*, p. to appear.

Florens, J.-P., and S. Van Bellegem (2015): "Instrumental variable estimation in functional linear models," *Journal of Econometrics*, p. to appear.

Gagliardini, P., and O. Scaillet (2012): "Tikhonov regularization for nonparametric instrumental variable estimators," *Journal of Econometrics*, 167(1), 61–75.

Hall, P., and J. L. Horowitz (2005): "Nonparametric Methods for Inference in the Presence of Instrumental Variables," *Annals of Statistics*, 32, 2904–2929.

Hansen, L. P. (1982): "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50(4), 1029–54.

Horowitz, J. L. (2011): "Applied Nonparametric Instrumental Variables Estimation," *Econometrica*, 79, 347–394.

Matzkin, R. L. (2015): "Estimation of Nonparametric Models with Simultaneity," *Econometrica*, 83(1), 1–66.

Newey, W. K., and J. L. Powell (2003): "Instrumental Variable Estimation of Nonparametric Models," *Econometrica*, 71(5), 1565–1578.