Optimal weighting for linear inverse problems[*]

Jean-Pierre FLORENS[†] Senay SOKULLU[‡]

Toulouse School of Economics University of Bristol

September 24, 2019

Abstract

Linear equations in functional spaces where the solution is not continuous require regularization to estimate the unknown function of interest. Under this set-up, we derive the optimal weighting operator which minimizes the mean integrated square error (MISE). We then use this result to construct the optimal feasible estimator. We illustrate our theoretical findings and the small sample properties of the proposed optimal estimator by means of simulations.

Keywords: Nonparametric IV Regression, Inverse problems, Tikhonov Regularization, Regularization Parameter, GMM

JEL Classification: C13, C14, C30

[*] We thank Geert Dhaene, Irene Gijbels, Chris Muris, Whitney Newey, David Pacini, James Powell, Demian Pouzo, Olivier Scaillet and Sami Stouli for valuable discussions and suggestions.

[†] 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, the optimal variance of the estimator is obtained if the empirical moments are weighted by the inverse of the square root of the variance matrix of the moments. In this paper we investigate Hansen's result for linear inverse problems such as nonparametric instrumental variables regression and show that "*the optimal*" weighting of Hansen (1982), which holds in many special cases of GMM such as minimum distance estimation or IV estimation, is no longer optimal once we have an infinite dimensional parameter of interest and in such a case the optimal weighting should take into account the regularity of the parameter of interest.

Consider the linear GMM problem corresponding to the following model:

$$y_i = z_i'\beta + u_i, \qquad \mathbb{E}(u_i|z_i) \neq 0, \qquad i = 1, .., n.$$

Assume that we have a vector of instruments w_i satisfying:

$$Cov(z_i, w_i) \neq 0, \quad \mathbb{E}(u_i|w_i) = 0 \quad \text{and} \quad Var(u_i|w_i) = \sigma^2.$$

Then the GMM estimator $\hat{\beta}$ of β is given by $\hat{\beta} = \text{argmin}_\beta \, \|w_i(y_i - z_i\beta)\|_{\Omega_n}$ for any symmetric and positive definite weighting matrix $\Omega_n \xrightarrow{p} \Omega$. Given this structure, it is straightforward to show that $\hat{\beta}$ is asymptotically normally distributed:

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, V),$$

where

$$V = \sigma^2 \left(\mathbb{E}(z_i w_i')\Omega\mathbb{E}(w_i z_i')\right)^{-1} \mathbb{E}(z_i w_i')\Omega\mathbb{E}(w_i w_i')\Omega\mathbb{E}(w_i z_i') \left(\mathbb{E}(z_i w_i')\Omega\mathbb{E}(w_i z_i')\right)^{-1}.$$

Hansen (1982) shows that the optimal GMM estimator is obtained for $\Omega = [\mathbb{E}(w_i w_i')]^{-1}$ with asymptotic variance given by $V = \sigma^2 \left(\mathbb{E}(z_i w_i')\Omega\mathbb{E}(w_i z_i')\right)^{-1}$. In this paper we investigate if this optimality result still holds when the dimensions of z_i and w_i are large or infinite. More generally, we investigate the question of optimal weighting in linear inverse problems. This question may also be viewed as an extension of the usual Generalized Least Squares (GLS) method to the infinite-dimensional case. In the GLS approach the optimal estimator is 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 in the case of linear inverse problems, i.e., if the solution requires a penalty.

Optimal weighting in linear inverse problems can be motivated by the estimation of simultaneous equations systems. Suppose $y_{1i}, y_{2i}, z_{1i}, z_{2i} \in \mathbb{R}$ satisfy the following system:

$$y_{1i} = \phi(z_{1i}) + \epsilon_{1i}$$

$$y_{2i} = \psi(z_{2i}) + \epsilon_{2i},$$

where z_{1i} and z_{2i} are endogenous and there is a vector of valid and relevant instruments w_i such that $\mathbb{E}[(\epsilon_{1i}, \epsilon_{2i})|w_i] = 0$. Moreover, ϵ_{1i} and ϵ_{2i} are potentially correlated. Note that in this nonparametric set-up, if the system is estimated equation by equation, it will lead to the usual nonparametric instrumental variables (NPIV) case. In this paper, we aim to answer the question of how to include the information coming from the error terms in the estimation procedure. We ask the question: Is it optimal to estimate the equations jointly by including the information coming from $Var(\epsilon|\mathbf{w})$?

Nonparametric instrumental variables regression is an ill-posed inverse problem, see Darolles, Fan, Florens, and Renault (2011); Newey and Powell (2003); Ai and Chen (2003) and Horowitz (2011) among others. As a result, the solution of this problem requires regularization. Among many solutions, Tikhonov Regularization provides a good solution to this problem where the minimization is modified by an L^2 penalty. However, in return, this penalty introduces a regularization bias which vanishes under certain conditions. We show that in the presence of regularization bias, the optimal weighting matrix derived for parametric problems is no longer optimal due to the contribution of the bias to the MISE. We then derive the optimal weighting operator for a general class of linear inverse problems including nonparametric instrumental variable regression.

From a mathematical viewpoint, the weighting problem can be considered as follows: The ill-posed inverse problem we consider is an integral equation. If the weighting operator is an integral operator, then we end up with a larger degree of ill-posedness. In such a case, an intuitive approach would be to select the weighting operator as a differential operator (such as inverse of variance operator) in order to reduce the degree of ill-posedness. If it is defined, weighting means differentiation of the equation before its resolution. However, the impact of weighting 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 a weighting operator which is an integral operator but the optimal values of regularization parameter is smaller and hence the effect is ambiguous. In this paper, we first derive the MISE of a weighted linear inverse problem, then minimize the MISE for a fixed regularization parameter with respect to the weighting operator. We find that the optimal weighting depends on both the regularity of the function of interest and the rate of decay of eigenvalues of the variance of the noise. Given our result, we come up with unfeasible and feasible optimal estimators and show that they are both consistent under the general set-up. Finally, we investigate our theoretical findings by means of simulations.

This paper can be related to three strands of literature. First of all, as already stated, one example of linear inverse problems in econometrics is nonparametric IV estimation. Hence, this paper is closely related to nonparametric instrumental variables literature, see Darolles et al. (2011); Newey and Powell (2003); Ai and Chen (2003) and Horowitz (2011) among others. Darolles et al. (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 papers show that the estimators they use are consistent however none of them considers optimality of the estimator of the infinite dimensional parameter.

With the growth of nonparametric IV literature in the recent years, attention has also been given to models that are semiparametric, where the parameter of interest includes both an infinite-dimensional function and a finite-dimensional vector. Florens, Johannes, and Van Bellegem (2012); Ai and Chen (2003); Chen and Pouzo (2009) consider the 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 the semiparametric efficiency bound when the weighting matrix is equal to the inverse of the 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 regularized nonparametric IV estimation. The main contribution of Gagliardini and Scaillet (2012) is the computation of an explicit asymptotic MISE for a Sobolev regularized estimator. However, they do not investigate the optimality of their estimator with respect to the choice of the weighting matrix.

This paper is also related to the literature on GMM with a 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 when 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. The problem we investigate in this paper is different since in our case the ill-posedness of the inverse problem is not caused by the choice of the weighting matrix; but we analyze the optimal choice of weighting operator for a problem which is ill-posed by construction.

Finally, the use of nonparametric techniques in structural models has increased the interest in nonparametric estimation of simultaneous equations, see Matzkin (2015) and Berry and Haile (2018). 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.

The paper proceeds as follows. In *Section 2* we introduce our model. In *Section 3* we examine the optimization of the MISE and present our result on optimal weighting. We then introduce the optimal unfeasible and feasible estimators and present the example of NPIV. In *Section 4*, we present simulation results which demonstrate our theoretical findings as well as small sample properties of the optimal feasible estimator. Finally, in *Section 5* we conclude.

2 The Set-up

Consider a linear inverse problem of the form:

$$\hat{r} = K\varphi + U, \tag{1}$$

$\varphi \in \mathcal{E}$ 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 and $\Sigma : \mathcal{F} \mapsto \mathcal{F}$ is a trace-class (nuclear) variance operator.[1] 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.[2]

Let L be a differential operator defined on $\mathcal{E}$ such that L is densely defined, self adjoint and L^{-1} is a compact operator from $\mathcal{E} \mapsto \mathcal{E}$. Moreover consider a weighting operator $A : \mathcal{F} \mapsto \mathcal{F}$. Assume that $\hat{r} \in \mathcal{D}(A)$ where $\mathcal{D}(A) \subset \mathcal{R}(K)$ and $\varphi \in \mathcal{D}(L)$.[3]

In case of a well-posed inverse problem, to solve for φ, the strategy would be to minimize $\|A\hat{r} - AK\varphi\|^2$ and in order to minimize the variance of the estimator an optimal choice would be $A = \Sigma^{-\frac{1}{2}}$. Consider the general ill-posed inverse problems. The Tikhonov regularized estimator using a Hilbert scale penalty is defined as the solution of:

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

and it is equal to:

$$\hat{\varphi}_\alpha = (\alpha L^*L + K^*A^*AK)^{-1}K^*A^*A\hat{r}. \tag{3}$$

[1] We assume that the variance of U is given by $\frac{1}{n}\Sigma$, however $\frac{1}{n}$ is not essential and it is assumed for the sake of exposition. One can replace $\frac{1}{n}$ by δ_n which should approach to 0 as n tends to infinity.

[2] K and Σ are assumed to be given for simplicity and this assumption can be relaxed.

[3] Note that in this section we introduce the problem of optimal weighting under a general setting. In Section 3.2, we show that one can define a NPIV problem under this setting. This setting can also be shown to fit to cases such as deconvolution problems (Carrasco, Florens, and Renault, 2007) or functional instrumental variables regression (Florens and Van Bellegem, 2015).

If L is invertible, equation (3) can be rewritten as:

$$\hat{\varphi}_\alpha = L^{-1}(\alpha I + L^{-1}K^*A^*AKL^{-1})^{-1}L^{-1}K^*A^*A\hat{r}. \tag{4}$$

Here we consider Tikhonov regularization with Hilbert Scale penalty. This approach leads to regularization with a smooth norm as well as it gives higher convergence rates than with Tikhonov regularization with L^2 penalty if the true function is smooth enough, see Neubauer (1988). Note that, Krein and Petunin (1966) show that the Sobolev Spaces $H^s(\mathbb{R}^n)$ build a Hilbert scale. Hence a Hilbert scale penalty is equivalent to penalization in Sobolev norm, which needs the assumption that the function of interest belongs to a Sobolev space, i.e. has square integrable derivatives up to a finite order. Gagliardini and Scaillet (2012) show with Monte Carlo simulations that Tikhonov regularization with Sobolev penalty increases the performance of the estimator compared to the one which is obtained with Tikhonov regularization with L^2 penalty.

In the sequel, we work with the spectral representation of the model. For the ease of exposition we assume the following:

Assumption 1 *There exist ϕ_j and ψ_j for $j = 1, 2, ..., \infty$ such that ϕ_j is an orthonormal base of $\mathcal{E}$ and ψ_j is an orthonormal base of $\mathcal{F}$. There also exist λ_{Kj}, λ_{Aj} and $\lambda_{L^{-1}j}$ which satisfy the following properties:*

(i) $(\phi_j)_{j=1}^\infty$'s are the eigenvectors of K^*A^*AK with eigenvalues $\lambda_{Kj}^2\lambda_{Aj}^2$ and:

$$A^*AK\phi_j = \lambda_{Aj}^2\lambda_{Kj}\psi_j.$$

(ii) $(\phi_j)_{j=1}^\infty$'s are the eigenvectors of $L^{-1*}L^{-1}$ with eigenvalues $\lambda_{L^{-1}j}^2$

The first part of Assumption 1 can be rephrased in the following way: K^*A^*AK has a discrete spectrum characterized by the eigenvectors ϕ_j and the eigenvalues μ_j^2. This assumption is essentially a regularity assumption which may be extended to the case of continuous spectrum. Indeed, the main assumption is that $A^*AK\phi_j = \tilde{\psi}_j$ constitute an orthogonal family in $\mathcal{F}$. In this case, one can normalize the $\tilde{\psi}_j$ in ψ_j and there exists positive numbers ρ_j such that $A^*AK\phi_j = \rho_j\psi_j$ where ψ_j is an orthonormal family of $\mathcal{F}$. Finally λ_{Kj} and λ_{Aj} can be defined by the following relations:

$$\mu_j = \lambda_{Kj}^2\lambda_{Aj}^2 \qquad \text{and} \qquad \rho_j = \lambda_{Kj}\lambda_{Aj}^2.$$

This assumption can be satisfied by defining ϕ_j, ψ_j and λ_{Kj}^2 as the singular value decomposition of K and by choosing A such that the eigenvectors of A^*A are ψ_j. Then ψ_j are also the eigenvectors of AA^* and λ_{Aj}^2 are the eigenvalues of A^*A. The second part of Assumption 1 limits the possible choices for L by imposing previously defined ϕ_j to be the eigenvectors of $L^{-1*}L^{-1}$.

Under Assumption 1, the spectral representation of the model in equation 1 can be written as:

$$\langle \hat{r}, \psi_j \rangle = \langle K\varphi, \psi_j \rangle + \langle u, \psi_j \rangle, \tag{5}$$

$$\langle \hat{r}, \psi_j \rangle = \lambda_{Kj} \langle \varphi, \phi_j \rangle + \frac{1}{\sqrt{n}} \langle \Sigma\psi_j, \psi_j \rangle^{1/2} \epsilon_j, \tag{6}$$

$$\langle \hat{\varphi}_\alpha, \phi_j \rangle = \frac{\lambda_{L^{-1}j}^2 \lambda_{Aj}^2 \lambda_{Kj}}{\alpha + \lambda_{L^{-1}j}^2 \lambda_{Aj}^2 \lambda_{Kj}^2} \langle \hat{r}, \psi_j \rangle, \tag{7}$$

where $E(\epsilon_j) = 0$, $Var(\epsilon_j) = 1$. The representation given in Equation 6 is standard in the literature of inverse problems. Especially, in statistical models, the noise U is assumed to be random contrary to deterministic which is, in general, the case in ill-posed inverse problem literature. Hence, this notation captures the fact that the model in Equation 1 can be written as Gaussian white noise model when $\Sigma = I$, see Cavalier (2008). In econometric applications the model is not a white noise model because the variance of the noise, $1/n\langle \Sigma\psi_j, \psi_j\rangle$ also declines with j, see Knapik, van der Vaart, van Zanten et al. (2011). Moreover, it can be seen from Equation 7 that the ill-posedness is coming from the decay of λ_{Kj}, i.e., $\lambda_{Kj} \to 0$ as $j \to \infty$ which then implies that small changes in $\hat{r}$ may explode the solution of $\hat{\varphi}$ in the case of no regularization (when $\alpha = 0$).

Given the spectral representation of the model introduced in equations 5 to 7 above, Proposition 1 states the mean integrated square error of the regularized estimate $\hat{\varphi}_\alpha$:

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

$$\mathbb{E}\|\hat{\varphi}_\alpha - \varphi\|^2 = \frac{1}{n}\sum_{j=1}^{\infty} \frac{\langle \Sigma\psi_j, \psi_j\rangle \lambda_{Kj}^2 \lambda_{Aj}^4 \lambda_{L^{-1}j}^4}{(\alpha + \lambda_{Kj}^2\lambda_{Aj}^2\lambda_{L^{-1}j}^2)^2} + \alpha^2 \sum_{j=1}^{\infty} \frac{\langle \varphi, \phi_j\rangle^2}{(\alpha + \lambda_{Kj}^2\lambda_{Aj}^2\lambda_{L^{-1}j}^2)^2}. \tag{8}$$

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 $tr(\Omega) = \sum_{j=1}^{\infty}\langle \Omega\delta_j, \delta_j\rangle$, we get the result. ■

As can be seen from the MISE expression in (8), L^{-1} plays the same role as A. Then the same value can be obtained either by weighting by A or by penalizing by LA^{-1}. Hence, in the following sections, we just consider weighting by A but our result may be reinterpreted in terms of Hilbert scale penalization.

Going back to the discussion of Assumption 1, it is an important assumption and it limits our presentation. In particular, in the general case, choosing $A = \Sigma^{-\frac{1}{2}}$ does not necessarily satisfy this assumption. However, for some important cases such as NPIV, this assumption allows $\Sigma^{-\frac{1}{2}}$ as a possible choice for A.[4] The importance of Assumption 1 may be underlined by the following remark: Consider the MISE expression given in Proposition 1 and consider a case where $\alpha = 0$ is possible, for example a finite dimensional case. In such a case, under Assumption 1, λ_{Aj}^2 disappears and the choice of A has no impact on the MISE of the estimator. It should be noted that in our framework, the possibility of choosing an optimal weighting operator is due to the trade-off between the variance and bias; it is not due to the minimization of the variance only as in the GMM literature. In the GMM case, a higher order asymptotic expansion of the estimator is necessary to introduce such a trade-off and it leads to an optimality result, see Newey and Smith (2004). In other words, we can say that Assumption 1 is relevant only in the ill-posed case, as the weighting would cancel out in the usual parametric case once we impose Assumption 1.

The estimation strategy which minimizes the risk measured by the MISE consists of the choice of a regularization parameter α and a weighting operator A which minimize $\mathbb{E}\|\hat{\varphi}_\alpha - \varphi\|^2$ at n, K and Σ fixed. The related result is presented in the next section.

3 MISE Optimisation

Consider the case where the regularization parameter α is fixed so are the ϕ_j and ψ_j families and the eigenvalues λ_{Kj} . Given Assumption 1, the optimization is not

[4] We discuss this assumption in the case of NPIV in Section 4.

on the full space of the operator A. The weighting operator A is constrained by the eigenvectors ϕ_j and ψ_j and the optimization is done over its eigenvalues λ_{Aj}. The MISE expression in Proposition 1 leads to the following result:

Proposition 2 *1. The optimal value for the sequence λ_{Aj}^2 is given by:*

$$\lambda_{Aj}^2 = \frac{\langle \varphi, \phi_j \rangle^2}{\langle \Sigma\psi_j, \psi_j \rangle} \alpha n.$$

2. This choice leads to the optimal (unfeasible) estimator:

$$\hat{\varphi}_u = \sum_{j=1}^{\infty} \frac{\langle \varphi, \phi_j \rangle^2 \lambda_{Kj} \langle \hat{r}, \psi_j \rangle}{\frac{1}{n}\langle \Sigma\psi_j, \psi_j \rangle + \langle \varphi, \phi_j \rangle^2 \lambda_{kj}^2} \phi_j,$$

$$\hat{\varphi}_u = (\frac{1}{n} Q + K^* K)^{-1} K^* \hat{r},$$

where Q is the operator:

$$Q : \mathcal{E} \mapsto \mathcal{E} : g \mapsto Qg = \sum_{j=1}^{\infty} \frac{\langle \Sigma\psi_j, \psi_j \rangle}{\langle \varphi, \phi_j \rangle^2} \langle g, \phi_j \rangle \phi_j \quad for \quad g \in \mathcal{E}.$$

3. Then the MISE of the optimal estimator is given by:

$$\frac{1}{n} \sum_{j=1}^{\infty} \frac{\lambda_{Kj}^2 \langle \Sigma\psi_j, \psi_j \rangle}{\left(\frac{1}{n} \frac{\langle \Sigma\psi_j, \psi_j \rangle}{\langle \varphi, \phi_j \rangle^2} + \lambda_{Kj}^2 \right)^2} + \sum_{j=1}^{\infty} \frac{\langle \Sigma\psi_j, \psi_j \rangle^2}{\langle \varphi, \phi_j \rangle^2 \left(\frac{1}{n} \frac{\langle \Sigma\psi_j, \psi_j \rangle}{\langle \varphi, \phi_j \rangle^2} + \lambda_{Kj}^2 \right)^2}.$$

The proof is presented in Appendix A.

This result differs from the standard result for GMM. In the usual case the optimal λ_{Aj}^2 is proportional to $\frac{1}{\langle \Sigma\psi_j, \psi_j \rangle}$. In the infinite-dimensional case with penalty, the optimal choice incorporates the smoothness of φ through the Fourier coefficients $\langle \varphi, \phi_j \rangle^2$. The optimal choice for A is then unfeasible because it depends on the unknown function φ. The estimator $\hat{\varphi}_u$ may be viewed as an oracle estimator and it does not depend on α. Equivalently, one can say that α is replaced by $1/n$. Note that the value of the MISE does not depend on α either. In some sense, the introduction of the $\langle \varphi, \phi_j \rangle^2$ replaces the choice of α.

The estimator $\hat{\varphi}_u$ can be interpreted as a Hilbert scale type extension of Tikhonov estimation. Indeed, $\hat{\varphi}_u$ is the argument φ that minimizes the following:

$$\hat{\varphi}_u = \operatorname*{argmin}_{\varphi} \|\hat{r} - K\varphi\|^2 + \frac{1}{n}\|Q^{1/2}\varphi\|^2.$$

Note that, the operator A may be a differential or an integral operator depending on the relative rate of decline of the Fourier coefficients of φ and of the $\langle \Sigma\psi_j, \psi_j \rangle$.If $\sum_j \frac{\langle \Sigma\psi_j, \psi_j \rangle}{\langle \varphi, \phi_j \rangle^2} < \infty$, A^{-1} becomes an integral operator and A is then a differential operator (as $\Sigma^{-1/2}$ in the parametric case). If, on the other hand, $\sum_j \frac{\langle \varphi, \phi_j \rangle^2}{\langle \Sigma\psi_j, \psi_j \rangle} < \infty$, A is a Hilbert-Schmidt integral operator. In other words, if φ is sufficiently regular, A becomes an integral operator. Or if we reconsider Hilbert Scale penalization, it means L becomes a differential operator. This result is very intuitive: if φ is sufficiently smooth regarding to Σ, a penalization by the norm of the derivative is optimal. Note that this idea was supported before by Gagliardini and Scaillet (2012). They suggest to penalize the derivatives of the unknown function to prevent oscillations in the estimated function. This result is also in line with Newey and Powell (2003)'s restriction of the parameter space. Tikhonov regularization with Hilbert scale penalty can be interpreted as minimization of $\|K\varphi - r\|$ subject to the constraint $\|L\varphi\| < \rho$ for some ρ, see Carrasco et al. (2007). In other words, it is equivalent to looking for a solution in a space where the norm of the derivatives of the functional parameter is bounded as in Newey and Powell (2003). Moreover, this case where φ is sufficiently smooth, optimal weighting can be interpreted as optimal norm. More precisely, given a regularization parameter α, our result suggests that it is optimal to use a Sobolev penalty.[5]

Regarding the consistency of $\hat{\varphi}_u$, it is intuitive to think that it is consistent as it has a smaller MISE than the MISE of $\hat{\varphi}_\alpha$ given in Equation 1 which converges to zero as $n \to \infty$, $n\alpha \to \infty$ and $\alpha \to 0$. The assumption below is needed for the formal proof of consistency of the optimal unfeasible estimator as well as for calculation of its rate of convergence.

Assumption 2

$$\sum_{j=1}^{\infty} \frac{\langle \varphi, \phi_j \rangle^{2(1-\beta)} \langle \Sigma\psi_j, \psi_j \rangle^{\beta}}{\lambda_{Kj}^{2\beta}} < \infty \quad \forall \quad \beta \in [0,1).$$

One can note the similarity of Assumption 2 and the source condition which has already been stated in papers such as Darolles et al. (2011) and Florens et al. (2012). Assumption 2 does not only state the regularity space which the function φ belongs to but it states a regularity space for both the function φ and the variance of the noise. Hence Assumption 2 can be seen as an extended source condition. The next theorem states the rate of convergence of $\hat{\varphi}_u$ under this extended source condition.

Theorem 1 *Assume that Assumptions 1 and 2 hold. Then:*

$$E\|\hat{\varphi}_u - \varphi\|^2 = O_p(n^{-\beta}).$$

Theorem 1 shows that the unfeasible estimator is consistent and it converges with a rate of $n^{-\beta}$ which is slower than the usual parametric rate. Though, this is not surprising. The optimal unfeasible estimator is also a nonparametric estimator, and by weighting we optimize its small sample MISE, not its asymptotic MISE. Moreover,

[5] We thank Damien Pouzo for making this point.

this oracle estimator has a faster rate of convergence than the minimax rate of the unweighted estimator which is $O_p(n^{-\frac{\beta}{\beta+1}})$.

3.1 The Optimal Feasible Estimator

Although Proposition 2 provides the optimal estimator, it is not feasible as it depends on the smoothness of the unknown function, φ. In this section, we construct optimal feasible estimator. A natural idea is to construct a two-step estimator. In a first step, φ is estimated using Tikhonov regularization with a regularization parameter α and in a second step, we replace $\langle \varphi, \phi_j \rangle^2$ by its estimator in the optimal weighting operator.

The first-step regularized estimate of φ is given by:

$$\hat{\varphi}_\alpha = \sum_j \frac{\lambda_{Kj}}{\alpha + \lambda_{Kj}^2} \langle \hat{r}, \psi_j \rangle \phi_j.$$

Then, $\langle \varphi, \phi_j \rangle$ can be replaced by:[6]

$$\langle \hat{\varphi}, \phi_j \rangle = \frac{\lambda_{Kj}}{\alpha + \lambda_{Kj}^2} \langle \hat{r}, \psi_j \rangle.$$

Hence the feasible estimator is equal to:

$$\hat{\varphi}_f = \sum_j \frac{\lambda_{Kj}^3 \langle \hat{r}, \psi_j \rangle^3}{\frac{1}{n}(\alpha + \lambda_{Kj}^2)^2 \langle \Sigma \psi_j, \psi_j \rangle + \lambda_{Kj}^4 \langle \hat{r}, \psi_j \rangle^2} \phi_j. \qquad (9)$$

As can be seen from Equation 9, the feasible estimator does depend on α through its dependence on first stage estimator, $\hat{\varphi}_\alpha$. Also, note that $\varphi = \sum_j \frac{1}{\lambda_{Kj}} \langle r, \psi_j \rangle \phi_j$ so the usual Tikhonov regularized estimator is obtained by replacing $\frac{1}{\lambda_{Kj}}$ by $\frac{\lambda_{Kj}}{\alpha + \lambda_{Kj}}$. Hence,

[6] As $\lambda_{Kj} \to 0$ very fast, this prevents us estimating φ by $\langle \hat{\varphi}, \phi_j \rangle = \frac{1}{\lambda_{Kj}} \langle \hat{r}, \psi_j \rangle \phi_j$ even though r can be estimated with a $\sqrt{n}$-rate.

the feasible estimator $\hat{\varphi}_f$ is also regularized where $\frac{1}{\lambda_{Kj}}$ is replaced by:

$$\frac{\lambda_{Kj}}{\frac{1}{n}(\alpha+\lambda_{Kj}^2)^2\frac{\langle\Sigma\psi_j,\psi_j\rangle}{\lambda_{Kj}^2\langle\hat{r},\psi_j\rangle^2}+\lambda_{Kj}^2}. \tag{10}$$

Equation (10) can also be written as $\frac{\lambda_{Kj}}{\alpha_j+\lambda_{Kj}^2}$ i.e., once the first step estimation is done, the second step can be seen as regularization with a sequence of α_j. The next theorem states the consistency of the optimal feasible estimator. The proof follows from Engl, Hanke, and Neubauer (1996) and it is presented in Appendix A.

Theorem 2 *Consider the feasible estimator given in Equation 9. Assume that α is fixed. Then as $n \to \infty$:*

$$\|\hat{\varphi}_f-\varphi\| \xrightarrow{p} 0$$

Theorem 2 shows that the optimal feasible estimator is consistent and the consistency can be achieved with a fixed regularization parameter, in other words, we do not need $\alpha \to 0$. As it is shown in the proof in Appendix A3, in this case, the role of α is replaced by $\frac{1}{n}$. This result is very important as it does not only show the consistency of the optimal feasible estimator but it also eliminates the problem of selection of optimal regularization parameter.[7] Next remark discusses very briefly the case with unknown K and Σ.

Remark 3 *If K and Σ are unknown, one needs to use estimators for K and Σ which will give estimates for λ_{Kj}, ϕ_j and ψ_j. In such a case, λ_{Kj}, ϕ_j and ψ_j in Equation 9 could be replaced by their estimates $\hat{\lambda}_{Kj}, \hat{\phi}_j$ and $\hat{\psi}_j$ and the proof of consistency should take into account these estimated values.*

[7] This result is supported by our simulation study in Appendix B.2

3.2 Example: Nonparametric IV Regression

In this section we consider optimal weighting in non-parametric instrumental variable regression setting. NPIV regression has been well studied in many papers; see Carrasco et al. (2007); Darolles et al. (2011); Hall and Horowitz (2005) among others. However, to the best of our knowledge, none of these papers has considered the optimality of the infinite dimensional parameter before. Below, we present optimal unfeasible and feasible estimators under this setup.

Consider a vector of random elements (Y, Z, W) such that:

$$Y = \varphi(Z) + U \quad and \quad \mathbb{E}(U|W) = 0. \tag{11}$$

The model then generates a linear inverse problem:

$$\begin{aligned} \mathbb{E}(\mathbb{E}(Y|W)|Z) &= \mathbb{E}(\mathbb{E}(\varphi(Z)|W)|Z), && (12) \\ r &= K\varphi, && (13) \end{aligned}$$

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, $\hat{r}$, and we assume that K is given, then one can write:

$$\hat{r} = K\varphi + U. \tag{14}$$

We assume that $\mathbb{E}(U) = 0$ and $\mathbb{V}(U) = \frac{\sigma^2}{n}K$ where σ^2 is known. The operator K is a self-adjoint trace class operator which has been discussed in many previous papers, see Carrasco et al. (2007); Darolles et al. (2011).

Note that this model has a particular feature as Σ and K are equal up to a multiplicative order. This means that the choice of $A = \Sigma^{-\frac{1}{2}}$ is possible in this set-up and it would result in $K^*A^*AK = K$, the $\phi_j (= \psi_j)$ being the eigenvectors of K and $A^*AK = I$. More precisely, in this case $\lambda_{Aj} = \lambda_{Kj}^{1/2}$. Although $A = \Sigma^{-1/2}$ is a possible choice for the weighting operator, it is not optimal due to regularization.

Given this setup, using the result (ii) in Proposition 2, the unfeasible estimator can be written as:

$$\hat{\varphi}_{u,IV} = \sum_{j=1}^{\infty} \frac{\langle \varphi, \phi_j \rangle^2 \lambda_{Kj} \langle \hat{r}, \psi_j \rangle}{\frac{\sigma^2}{n}\lambda_{Kj} + \langle \varphi, \phi_j \rangle^2 \lambda_{Kj}^2} \phi_j. \tag{15}$$

Define the operator R such that $R : \mathcal{E} \mapsto \mathcal{E} : g \mapsto Rg = \sum_{j=1}^{\infty} \langle \varphi, \phi_j \rangle^2 \langle g, \phi_j \rangle \phi_j$. Then the unfeasible estimator can be rewritten as:

$$\hat{\varphi}_{u,IV} = \left(\frac{\sigma^2}{n} K + RK^*K \right)^{-1} RK^*\hat{r}. \tag{16}$$

To obtain the feasible estimator, $\langle \varphi, \phi_j \rangle$ in Equation 15 is replaced by $\frac{\lambda_{Kj}}{\alpha + \lambda_{Kj}^2} \langle \hat{r}, \psi_j \rangle$:

$$\hat{\varphi}_{f,IV} = \sum_{j=1}^{\infty} \frac{\langle \hat{r}, \psi_j \rangle^2}{\frac{\sigma^2}{n}(\alpha + \lambda_{Kj}^2)^2 + \lambda_{Kj}^3 \langle \hat{r}, \psi_j \rangle^2} \lambda_{Kj}^2 \langle \hat{r}, \psi_j \rangle \phi_j. \tag{17}$$

Define operator T such that $T : \mathcal{E} \mapsto \mathcal{E} : g \mapsto Tg : \sum_{j=1}^{\infty} \langle \hat{r}, \psi_j \rangle^2 \langle g, \phi_j \rangle \phi_j$. Then the feasible estimator can be rewritten as:

$$\hat{\varphi}_{f,IV} = \left(\frac{\sigma^2}{n} (\alpha I + K^2)^2 + K^3 T \right)^{-1} KTK^*\hat{r}. \tag{18}$$

4 Simulations

In this section we present simulations to show the performance of the unfeasible and the feasible optimal estimators compared to that of the unweighted estimator. We conduct two different types of simulations. The first group, numerical illustrations, for which we generated the eigenvalues and compute the MISE for the unweighted, the weighted unfeasible and weighted feasible estimators. In the second group of simulations, we conduct a Monte Carlo experiment. We generate data from an NPIV model and estimate the unknown function for a given K using weighted feasible estimator and unweighted estimator and we compute their MISE. We then extend this IV simulation to the semiparametric case in which we assume that the operator K is known to be coming from a normal family with an unknown variance. Numerical illustration and results of Monte Carlo simulations show that the feasible optimal estimator performs better than unweighted estimator.

4.1 Numerical Illustration

In this set of simulations, using the spectral representation given in Equations 5 to 7, we generate eigenvalues of the model and compute the MISE for the unweighted estimator ($\hat{\varphi}_\alpha$), the optimal unfeasible estimator ($\hat{\varphi}_u$) and the optimal feasible estimator ($\hat{\varphi}_f$).

Given Equations 5 to 7, we generate λ_{Kj}, $\langle \varphi, \phi_j \rangle$ and $\langle \Sigma\psi_j, \psi_j \rangle$ under the assumptions of geometric and exponential spectra. In the geometric spectrum case, we posit that:

$$\lambda_{Kj} = \frac{1}{j^a}, \qquad \langle \varphi, \phi_j \rangle = \frac{1}{j^b}, \qquad \langle \Sigma\psi_j, \psi_j \rangle = \frac{1}{j^{2c}},$$

for $j = 1, ..., 1000$, sample size $n = 100$ and $a = 4$, $b = 3$ and $c = 1$. Then the MISE is calculated by using the following formula:

$$MISE = E\left[\sum_j \left(\langle \Phi, \phi_j \rangle - \frac{1}{j^b}\right)^2\right]$$

where $\Phi = \{\hat{\varphi}_\alpha, \hat{\varphi}_u, \hat{\varphi}_f\}$.

In the case with the exponential spectrum, we posit the following:

$$\lambda_{Kj} = \rho^j, \qquad \langle \varphi, \phi_j \rangle = \rho^{j\beta}, \qquad \langle \Sigma\phi_j, \phi_j \rangle = \rho^{2j\mu},$$

for $j = 1, .., 100$, $\rho = 0.6$, $\mu = 1$ and $\beta = 0.5$. Then for $\Phi = \{\hat{\varphi}_\alpha, \hat{\varphi}_u, \hat{\varphi}_f\}$, the MISE is given by:[8]

$$MISE = E\left[\sum_j \left(\langle \Phi, \phi_j \rangle - \rho^{j\beta}\right)^2\right].$$

In both set of simulations we use the optimal value of α given our design. To be more precise, given a grid of values for α, we select the one which minimizes the MISE of the $\hat{\varphi}_f$.[9] The results are presented in Table 1.

Table 1: **Numerical Illustration**

	MISE			
	$\hat{\varphi}_\alpha$	$\hat{\varphi}_u$	$\hat{\varphi}_f$	α_{opt}
Geometric	0.0710	0.0271	0.0593	0.0223
Exponential	0.5374	0.0973	0.3501	0.1220

Two main results can be reached from Table 1. First of all, unsurprisingly, the smallest MISE is obtained with unfeasible estimator. Hence, the weighting does decrease the MISE compared to the case where no weighting is used. Secondly, the feasible estimator, as well, performs better than the unweighted estimator. Hence, by using the two step feasible estimator one can improve the MISE. In the next section, we present Monte Carlo simulations to compare the performance of optimal feasible estimator and unweighted estimator in the NPIV setup.

[8] In the exponential spectrum case, eigenvalues decay to zero much faster than in the geometric spectrum case. For this reason, we take jmax=100 instead of 1000 in this set of simulations.

[9] We also conduct simulations where we select α which minimizes the MISE of the first step estimator $\hat{\varphi}_\alpha$. The results are qualitatively the same, see Table 4 in Appendix B.

4.2 Monte Carlo Experiment: Nonparametric IV Regression

4.2.1 Known K

We generate the data as the following: X and Z are drawn from a multivariate normal distribution with mean $(0 \quad 0)'$ and variance $\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ where we fix ρ to be equal to 0.6. Moreover, U is drawn from a univariate normal distribution with 0 mean and variance equal to 0.01.[10] We generate φ to be equal to $\varphi = \frac{Z^2-1}{\sqrt{2}}$. Then Y is given by:

$$Y = \frac{Z^2 - 1}{\sqrt{2}} + U.$$

We chose a geometric spectrum, so the eigenvalues are given by $\lambda_{Kj} = \rho^{2j}$ and the basis functions $\phi_j(Z)$ and $\psi_j(X)$ are generated using Hermite polynomials. Given this set-up, we estimate unweighted φ using the following:

$$\hat{\varphi}_\alpha(z) = \sum_j \frac{\lambda_{Kj}}{\alpha + \lambda_{Kj}^2} \left(\frac{1}{n} \sum_{i=1}^{n} y_i \phi_j(z_i) \lambda_{Kj}^{1/2} \right) \phi_j(z). \tag{19}$$

Then the scalar product can be written as:

$$\langle \hat{\varphi}_\alpha, \phi_j \rangle = \frac{\lambda_{Kj}}{\alpha + \lambda_{Kj}^2} \left(\frac{1}{n} \sum_{i=1}^{n} y_i \phi_j(z_i) \lambda_{Kj}^{1/2} \right). \tag{20}$$

We use first stage estimate $\hat{\varphi}_\alpha$ to obtain $\langle \hat{\varphi}_\alpha, \phi_j \rangle$ and $\hat{U}$ - which is then used to compute $\hat{\sigma}_u^2$ - to obtain the feasible estimator:

$$\hat{\varphi}_f(z) = \sum_j \frac{\langle \hat{\varphi}_\alpha, \phi_j \rangle^2 \lambda_{Kj} \left(\frac{1}{n} \sum_{i=1}^{n} y_i \phi_j(z_i) \lambda_{Kj}^{1/2} \right)}{1/n \hat{\sigma}_u^2 \lambda_{Kj} + \langle \hat{\varphi}_\alpha, \phi_j \rangle^2 \lambda_{Kj}^2} \phi_j(z). \tag{21}$$

[10] We also tried different values for the variance of U, our results stay the same qualitatively.

We replicate this exercise 100 times for a sample size equals to 200. We truncate the sum at $j = 25$. As for the regularization parameter, α, we select it in two different ways. In the first set of simulations, given a grid of values of α, we select the one which minimizes the MISE of the first stage estimator, and in the second set of simulations, we select the one which minimizes the MISE of the optimal feasible estimator. Table 2 shows the MISE of the unweighted and weighted optimal estimators under two different selection rules for α. Not surprisingly, the feasible estimator performs better when α is selected in the second stage, i.e., when the optimal α is selected as the minimizer of the MISE of the second step estimator. However, when α is selected at the first stage unweighted estimator has a smaller MISE. Moreover, α chosen with respect to first step estimator is larger than α chosen with respect to second step estimator showing that optimal estimator requires undersmoothing at the first stage. Figures 1 and 2 confirm these results. The unweighted estimator is closer to the true curve in Figure 1, whereas it is the optimal feasible estimator which is closer the true curve in Figure 2. Figure 3 shows the plots of $\hat{\varphi}_f$ for all draws over the true function where α is chosen with respect to the second step estimator.

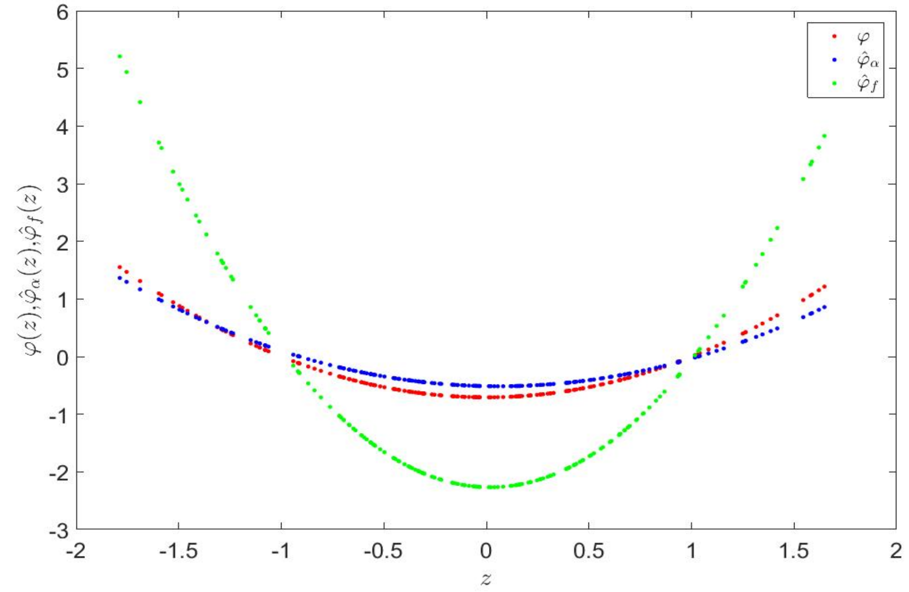
We can conclude that for a known K optimal weighted estimator performs better than unweighted estimator. However, in an empirical application, it is unlikely that one knows K. In the next section, we analyze the case when K is partially known.[11]

Table 2: **IV simulation results - K known**

	MISE		
	$\hat{\varphi}_\alpha$	$\hat{\varphi}_f$	α_{opt}
α_{opt}^{1s} first stage	0.3745	2.2020	0.0541
α_{opt}^{2s} second stage	0.5828	0.2795	0.0310

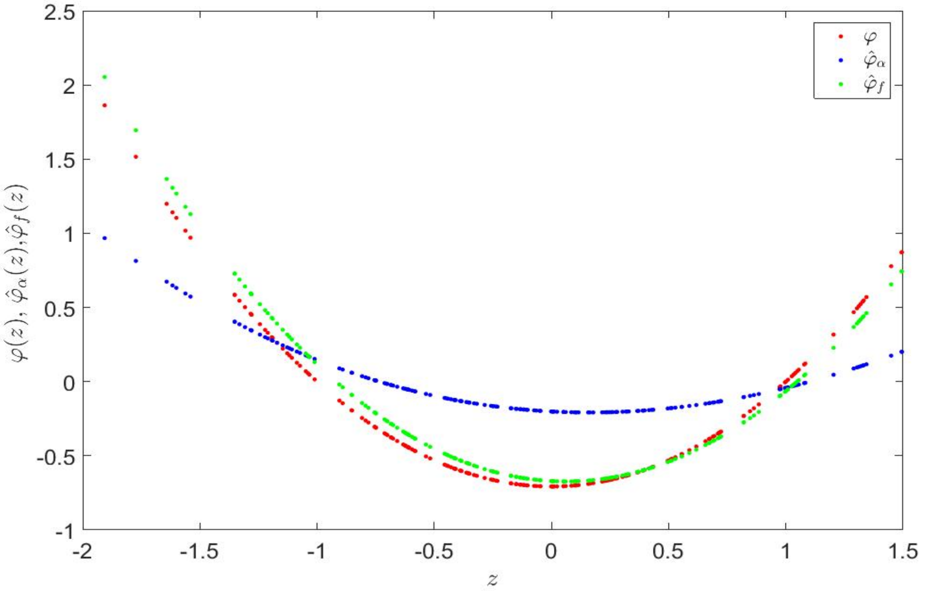
[11] Although we developed the theory for known K throughout the paper, in Appendix B we present Monte Carlo simulation results in the case of NPIV when K is unknown.Theoretical analysis with unknown K is left for future work.

Figure 1: **Simulation result with one draw for α_{opt}^{1s}**



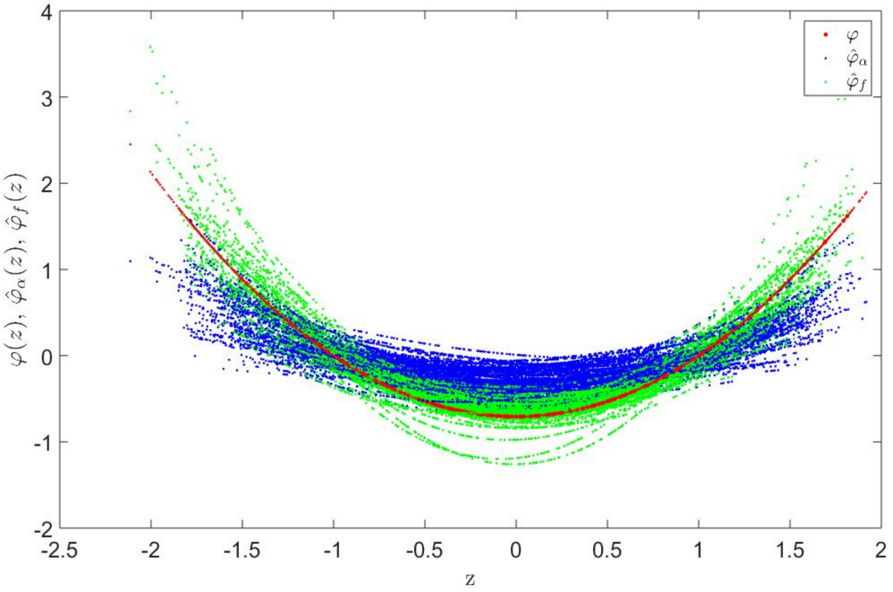
Note: α is selected in order to minimize the MISE of the first step estimator.

Figure 2: **Simulation result with one draw for α_{opt}^{2s}**



Note: α is selected in order to minimize the MISE of the second step estimator.

Figure 3: **Simulation result with 100 draws**



Note: α is selected in order to minimize the MISE of the second step estimator. Green dots are the estimated curve by using optimal feasible estimator at each draw while the blue dots are unweighted estimates.

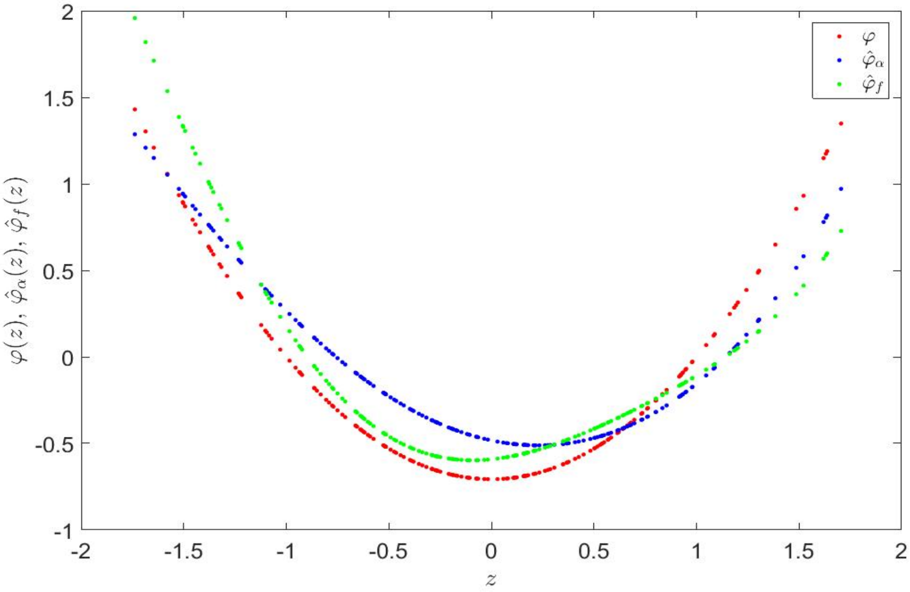
4.2.2 Partially Known K

In this set of simulations, the data is generated in the same way as in Section 5.2.1. The only difference is that we no longer assume that the operator K is fully known but the relation $\Sigma = \sigma^2 K$ still holds. We instead assume that the operator K is from a normal family with an unknown variance. Hence, the eigenvalues are given by $\lambda_{Kj} = \hat{\rho}^{2j}$ where $\hat{\rho}$ is an estimator for the covariance of X and Z. $\hat{\varphi}_\alpha$ and $\hat{\varphi}_f$ can then be obtained using Equations 19 to 21. We replicate the model and the estimation 100 times for samples of size 200. The regularization parameter α is selected as in the case with known K. The results are presented in Table 3 and Figures 4 to 6. Optimal feasible estimator again performs better when α is selected to minimize MISE at the second stage. Moreover, contrary to the case with known K, to get better fit for $\hat{\varphi}_f$, one needs to oversmooth at the first stage as α_{opt}^{2s} is larger than α_{opt}^{1s}.

Table 3: **IV simulation results - K partially known**

	MISE		
	$\hat{\varphi}_\alpha$	$\hat{\varphi}_f$	α_{opt}
α_{opt}^{1s} first stage	0.3501	0.6894	0.0111
α_{opt}^{2s} second stage	0.5925	0.3058	0.0351

Figure 4: **Simulation result with one draw for α_{opt}^{1s}**



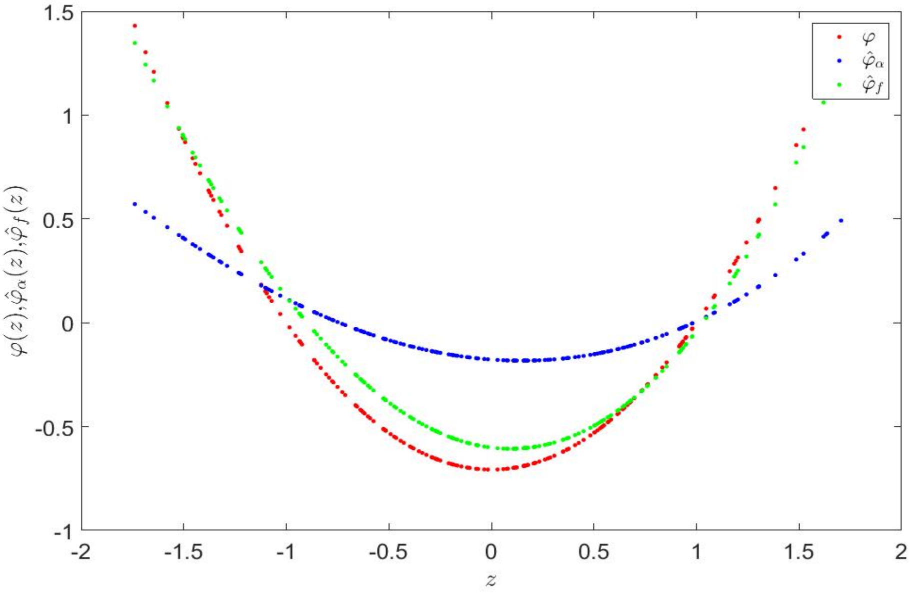
Note: α is selected in order to minimize the MISE of the first step estimator.

5 Conclusion

In this paper we examine the question of optimal weighting in linear inverse problems. We have several results. First of all, under a very general set-up, we have shown that weighting and Hilbert scale penalty play the same role. Hence, one may fix one of these operators to identity. Secondly, we have found that the optimal weighting depends on the regularity of the function of interest and on the variance of the noise. A conjecture would be to equalize the regularity of φ and the sum of the degree of ill-posedness of A and Σ. Third, given our results on optimal weighting we have come up with the optimal feasible estimator and shown that it is consistent and its consistency does not depend on the regularization parameter. Finally, we have supported our theoretical findings by means of Monte Carlo simulations.

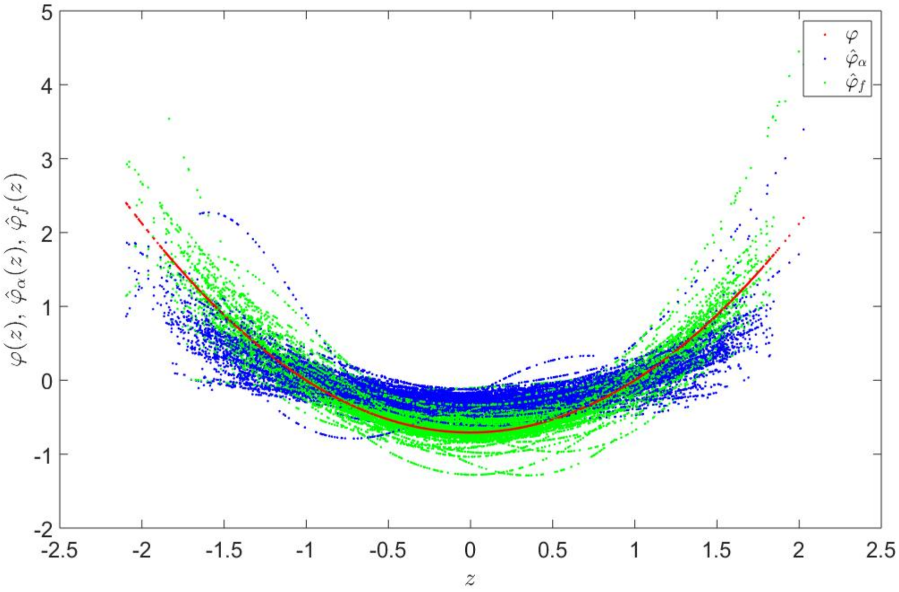
This paper can be considered as the first of a series of papers on this topic. We leave establishing the oracle properties for the optimal feasible estimator and the treatment of the problem when K and Σ are unknown for future work. We believe that our results will contribute to the literature on the nonparametric estimation of simultaneous equations. Hence, the development of nonparametric three stage least squares estimator using this optimal weighting matrix is also left for future work.

Figure 5: **Simulation result with one draw for α_{opt}^{2s}**



Note: α is selected in order to minimize the MISE of the second step estimator.

Figure 6: **Simulation result with 100 draws**



Note: α is selected in order to minimize the MISE of the second step estimator. Green dots are the estimated curve by using optimal feasible estimator at each draw while the blue dots are unweighted estimates.

References

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

Berry, S. T. and P. A. Haile (2018): "Identification of nonparametric simultaneous equations models with a residual index structure," *Econometrica*, 86, 289–315.

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

——— (2014): "On The Asymptotic Efficiency of GMM," *Econometric Theory*, 30, 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, Elsevier, vol. 6 of *Handbook of Econometrics*, chap. 77.

Cavalier, L. (2008): "Nonparametric statistical inverse problems," *Inverse Problems*, 24, 034004.

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

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

Engl, H. W., M. Hanke, and A. Neubauer (1996): *Regularization of inverse problems*, vol. 375, Springer Science & Business Media.

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

Florens, J.-P. and S. Van Bellegem (2015): "Instrumental variable estimation in functional linear models," *Journal of Econometrics*, 186, 465–476.

Gagliardini, P. and O. Scaillet (2012): "Tikhonov regularization for nonparametric instrumental variable estimators," *Journal of Econometrics*, 167, 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, 1029–54.

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

Knapik, B. T., A. W. van der Vaart, J. H. van Zanten, et al. (2011): "Bayesian inverse problems with Gaussian priors," *The Annals of Statistics*, 39, 2626–2657.

Krein, S. G. and Y. I. Petunin (1966): "Scales of Banach spaces," *Russian Mathematical Surveys*, 21, 85.

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

NEUBAUER, A. (1988): "When do Sobolev spaces form a Hilbert scale?" *Proceedings of the American Mathematical Society*, 103, 557–562.

NEWEY, W. K. AND J. L. POWELL (2003): "Instrumental Variable Estimation of Nonparametric Models," *Econometrica*, 71, 1565–1578.

NEWEY, W. K. AND R. J. SMITH (2004): "Higher order properties of GMM and generalized empirical likelihood estimators," *Econometrica*, 72, 219–255.

Appendices

A Proofs

A.1 Proof of Proposition 2

Proof. We first minimize the MISE given in Proposition 1 with respect to λ_{Aj}^2 which gives the first result. Note that $\hat{\varphi}_\alpha$ is given by:

$$\hat{\varphi}_\alpha = \sum_j \frac{\lambda_{Kj}\lambda_{Aj}^2}{\alpha + \lambda_{Kj}^2\lambda_{Aj}^2}\langle \hat{r}, \psi_j\rangle \varphi_j.$$

Then the second result is obtained by replacing optimal λ_{Aj}^2 in the above equation. Finally, the third result is obtained by substituting λ_{Aj}^2 by $\frac{\langle \varphi, \phi_j\rangle^2}{\langle \Sigma\psi_j, \psi_j\rangle}\alpha n$ in the MISE formula. ■

A.2 Proof of Theorem 1

Proof. If we replace the λ_{Aj}^2 by $\frac{\langle \varphi, \phi_j\rangle^2}{\langle \Sigma\psi_j, \psi_j\rangle}\alpha n$ in the MISE formula given in Proposition 1, we obtain the MISE of the optimal unfeasible estimator:

$$E\|\hat{\varphi}_u - \varphi\|^2 = \frac{1}{n}\sum_{j=1}^{\infty} \frac{\langle \Sigma\psi_j, \psi_j\rangle \lambda_{Kj}^2}{\left(\frac{1}{n}\frac{\langle \Sigma\psi_j, \psi_j\rangle}{\langle \varphi, \phi_j\rangle^2} + \lambda_{Kj}^2\right)^2} + \frac{1}{n^2}\sum_{j=1}^{\infty} \frac{\langle \Sigma\psi_j, \psi_j\rangle^2}{\langle \varphi, \phi_j\rangle^2 \left(\frac{1}{n}\frac{\langle \Sigma\psi_j, \psi_j\rangle}{\langle \varphi, \phi_j\rangle^2} + \lambda_{Kj}^2\right)^2}.$$

In the above MISE, the first term is the variance whereas the second term is the bias square. We will analyze them separately. Let us first consider the bias term. If we divide and multiply it by $\frac{\langle \varphi, \phi_j\rangle^2}{\langle \Sigma\psi_j, \psi_j\rangle^2}$, we obtain the following after some manipulation:

$$\frac{1}{n^2}\sum_{j=1}^{\infty} \frac{\langle \varphi, \phi_j\rangle^2}{\left(\frac{1}{n} + \frac{\langle \varphi, \phi_j\rangle^2}{\langle \Sigma\psi_j, \psi_j\rangle}\lambda_{Kj}^2\right)^2}. \tag{22}$$

Denote $x = \frac{\langle \varphi, \phi_j \rangle^2}{\langle \Sigma \psi_j, \psi_j \rangle} \lambda_{Kj}^2$. If we divide and multiply Equation 22 by x^β:

$$\frac{1}{n^2} \sum_{j=1}^{\infty} \frac{\langle \varphi, \phi_j \rangle^2}{x^\beta} \frac{x^\beta}{(1/n + x)^2},$$

where $\frac{x^\beta}{(1/n+x)^2}$ is $O_p(n^{2-\beta})$. Then the whole bias term is $O_p(n^{-\beta})$ and under assumption 2, bias goes to 0 as $n \to \infty$. We now examine the variance term. As before, after some manipulation the variance term can be rewritten as:

$$\frac{1}{n} \sum_{j=1}^{\infty} \frac{\langle \varphi, \phi_j \rangle^2 \lambda_{Kj}^2 \frac{\langle \varphi, \phi_j \rangle^2}{\langle \Sigma \psi_j, \psi_j \rangle}}{\left(\frac{1}{n} + \lambda_{Kj}^2 \frac{\langle \varphi, \phi_j \rangle^2}{\langle \Sigma \psi_j, \psi_j \rangle} \right)^2}. \tag{23}$$

As is done with the bias term, denote $x = \frac{\langle \varphi, \phi_j \rangle^2}{\langle \Sigma \psi_j, \psi_j \rangle} \lambda_{Kj}^2$, and divide and multiply Equation 23 by x^β, one obtains:

$$\frac{1}{n} \sum_{j=1}^{\infty} \frac{\langle \varphi, \phi_j \rangle^2}{x^\beta} \frac{x^{\beta+1}}{(1/n + x)^2}.$$

The term $\frac{x^{\beta+1}}{(1/n+x)^2}$ is $O_p(n^{-1})$ and the whole variance term is $O_p(n^{-\beta})$. Thus under our extended source condition, the variance term as well vanishes as $n \to \infty$. ■

A.3 Proof of Theorem 2

Proof. The proof follows by Theorem 4.1 and Theorem 4.2 in Engl et al. (1996). One can decompose $\|\hat{\varphi}_f - \varphi\|$ as the following:

$$\|\hat{\varphi}_f - \varphi\| = \underbrace{\|\hat{\varphi}_f - \varphi_f\|}_{A} + \underbrace{\|\varphi_f - \varphi\|}_{B}$$

We now show that both A and B converge to zero. Let us start with B. Note that B captures the regularization bias and it can be shown to converge to zero by using Theorem 4.1 in Engl et al. (1996). The theorem states that for $g_\rho(x)$ such that

$$(1) \quad |xg_\rho(x)| < C \quad and \quad (2) \quad \lim_{\rho \to 0} g_\rho(x) = \frac{1}{x} \quad \text{for all} \quad x \in [0, \|K\|^2]$$

then

$$\lim_{\rho \to 0} g_\rho(K^*K)K\phi = r$$

If one can verify (1) and (2) in the case of feasible estimation, then one can conclude $\|\varphi_f - \varphi\| \to 0$. Using Equation 10, one can write $g_\rho(x)$:

$$g_\rho(x) = \frac{x \left\langle \hat{r}, \psi_j \right\rangle^2}{\rho(\alpha + x)^2 \left\langle \Sigma\psi_j, \psi_j \right\rangle + x^2 \left\langle \hat{r}, \psi_j \right\rangle^2}$$

where $\rho = 1/n$. Then:

$$\lim_{\rho \to 0} g_\rho(x) = \frac{x \left\langle \hat{r}, \psi_j \right\rangle^2}{x^2 \left\langle \hat{r}, \psi_j \right\rangle^2} = \frac{1}{x}$$

It is straightforward to show the first condition as well, as:

$$|xg_\rho(x)| = \left| \frac{x^2 \left\langle \hat{r}, \psi_j \right\rangle^2}{\rho(\alpha + x)^2 \left\langle \Sigma\psi_j, \psi_j \right\rangle + x^2 \left\langle \hat{r}, \psi_j \right\rangle^2} \right| < 1$$

and it is bounded.

We now show that the term A, $\|\hat{\varphi}_f - \varphi_f\|$ converges to zero in probability. The result will follow from Theorem 4.2 in Engl et al. (1996). Define $G_\rho := \sup\{|g_\rho(x)||x \in [0, \|T\|^2]\}$. Then the theorem shows that:

$$\|\hat{\varphi}_f - \varphi_f\| \leq \frac{1}{\sqrt{n}}\sqrt{CG_\rho}$$

$\sup g_\rho(x)$ is given when $x = \sqrt{x^*}$ where

$$x^* = \frac{\frac{1}{n}\alpha^2 \langle \Sigma\psi_j, \psi_j \rangle}{\langle \hat{r}, \psi_j \rangle^2 + \frac{1}{n} \langle \Sigma\psi_j, \psi_j \rangle}$$

Then

$$\frac{1}{n} \sup g_\rho(x) = \frac{x^* \langle \hat{r}, \psi_j \rangle^2}{(\alpha + x^*)^2 \langle \Sigma\psi_j, \psi_j \rangle + \frac{\alpha^2 \langle \Sigma\psi_j, \psi_j \rangle}{\langle \hat{r}, \psi_j \rangle^2 + \frac{1}{n} \langle \Sigma\psi_j, \psi_j \rangle} \langle \hat{r}, \psi_j \rangle^2} \tag{24}$$

Let us first examine x^*. After some manipulations, it can be rewritten as:

$$\frac{\alpha^2}{\frac{n\langle \hat{r}, \psi_j \rangle^2}{\langle \Sigma\psi_j, \psi_j \rangle} + 1}$$

The term in denominator is bounded which makes x^* is of order $O(1)$ for α fixed.Then we can conclude that the numerator of Equation 24 is of order $O(1)$. As for the denominator of 24, the second term dominates so we can examine just that term. One can write:

$$\frac{\alpha^2 \langle \Sigma\psi_j, \psi_j \rangle}{\frac{1}{n} \langle \tilde{r}, \psi_j \rangle^2 + \frac{1}{n} \langle \Sigma\psi_j, \psi_j \rangle}$$

where $\tilde{r} = n\hat{r}$. Again for α fixed, this term is $O_p(n)$ and hence we can conclude that 24 is $O_p(1/n)$:

$$\text{As} \quad n \to \infty, \|\hat{\varphi}_f - \varphi_f\|^2 \xrightarrow{p} 0.$$

■

B Simulation results

B.1 Numerical Illustration

Table 4: Numerical illustration results, α optimized with respect to first stage estimator

	MISE			
	$\hat{\varphi}_\alpha$	$\hat{\varphi}_u$	$\hat{\varphi}_f$	α_{opt}
Geometric	0.0948	0.0273	0.0710	0.0236
Exponential	0.5105	0.0967	0.3379	0.1072

If we compare Tables 1 and 4, we see that the exponential spectrum requires oversmoothing in the first step to improve the performance of the weighted feasible estimator. Moreover, optimal feasible estimator performs always better than unweighted estimator either one selects α at the first or second stage.

B.2 Monte Carlo Experiment: NPIV with unknown K

In this paper, we develop theory for optimal weighting in inverse problems assuming that the operator K is known. In this section, we present some Monte Carlo evidence on the small sample performance of the optimal feasible estimator when K is unknown.

The data is generated exactly the same way as in Section 5.2.1, however, we do not assume that K is known, hence we cannot use Hermite polynomials to generate the basis functions. In this case, one way to estimate the model is to estimate the operator K, then obtain its eigenvalues and eigenvectors to estimate the function of interest φ. The conditional expectation operator can be estimated following Carrasco et al. (2007). For a function $f(t)$ and $Kf(t) = E[f(t)|W = w]$, the kernel estimation of K for a bandwidth h_w is given by:

$$\hat{K}_n f(t) = \frac{\sum_{i=1}^{n} f(t_i) K\left(\frac{w-w_i}{h_w}\right)}{\sum_{i=1}^{n} K\left(\frac{w-w_i}{h_w}\right)} = \sum a_i(f)\varepsilon_i,$$

where

$$a_i(f) = f(t_i) \quad and \quad \varepsilon_i = \left[\frac{K\left(\frac{w-w_i}{h_w}\right)}{\sum_{i=1}^{n} K\left(\frac{w-w_i}{h_w}\right)}\right].$$

Note that in our problem K is given by $K\varphi(Z) = E[E[\varphi(Z)|X]|Z]$. Hence $\hat{K}$ is given by $K_Z K_X$ where the K_Z and K_X matrices are the ones with the following $(i, j)th$ elements:

$$K_z(i, j) = \frac{K_z\left(\frac{z_i - z_j}{h_z}\right)}{\sum_j K_z\left(\frac{z_i - z_j}{h_z}\right)},$$

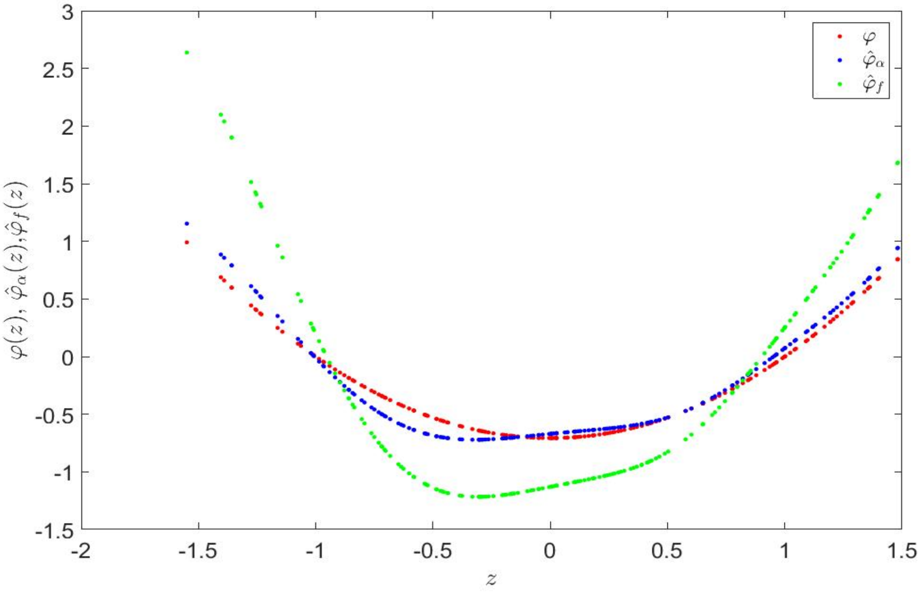
$$K_x(i, j) = \frac{K_x\left(\frac{x_i - x_j}{h_x}\right)}{\sum_j K_x\left(\frac{x_i - x_j}{h_x}\right)}.$$

Given this $\hat{K}$, the estimated eigenvalues $\hat{\lambda}_j^2$ and eigenvectors $\hat{\phi}_j$ are given by the eigenvalues and eigenvectors of $\hat{K}'\hat{K}$. Given these values, $\hat{\varphi}_\alpha$ and $\hat{\varphi}_f$ can be estimated using Equations 19 to 21. As in the previous Monte Carlo exercises, the model is replicated for 100 times for samples of size 200 and the regularization parameter is selected in the same two ways. The results are presented in Table 5 and Figures 7 to 9. Optimal feasible estimator performs better when α is selected to minimize MISE at the second stage and to get better fit for $\hat{\varphi}_f$, one needs to slightly undersmooth at the first stage as α_{opt}^{2s} is smaller than α_{opt}^{1s}. One final thing that can be noticed from Table 5 is that the optimal feasible estimator looks less sensitive to different values of α. Figure 10 below shows how MISE of the unweighted estimator and optimal feasible estimator change with different values of α. As can be seen from Figure 10, optimal feasible estimator is less sensitive to regularization parameter. The MISE of the optimal feasible estimator ranges between 0.5 and 1 whereas the range is much broader for the unweighted estimator. Hence, with the weighting we do not only improve the MISE of the estimator but we also make it more robust to different values of regularization parameter. Given the importance of the selection of smoothing parameters in nonparametric approaches, this result is very important. We can say that optimal weighting makes the estimator more robust to different values of smoothing parameter and hence decreases the need of finding the best selection rule for α.

Table 5: **IV simulation results - K unknown**

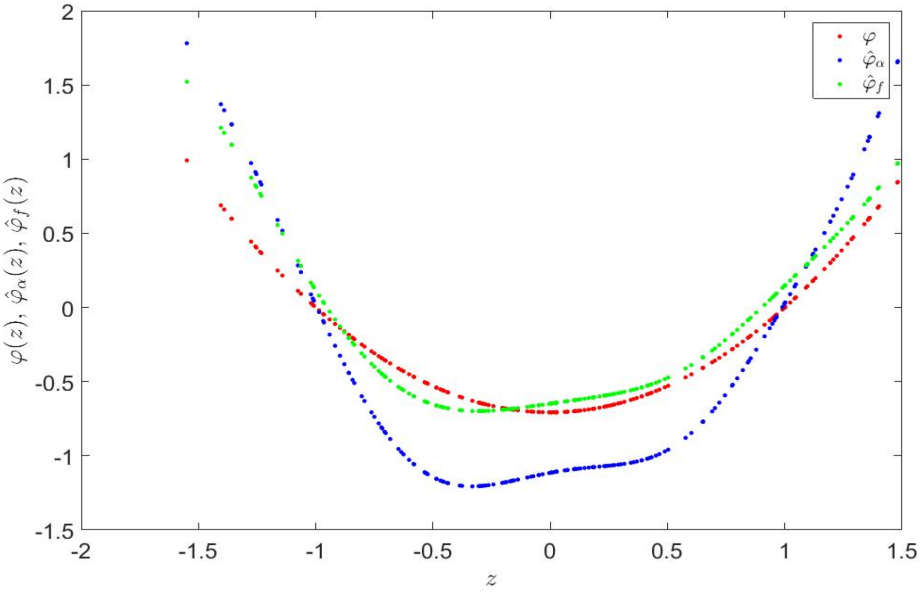
	MISE		
	$\hat{\varphi}_\alpha$	$\hat{\varphi}_f$	α_{opt}
α_{opt}^{1s} first stage	0.2949	0.5332	0.0544
α_{opt}^{2s} second stage	0.5954	0.3929	0.0287

Figure 7: **Simulation result with one draw for** α_{opt}^{1s}



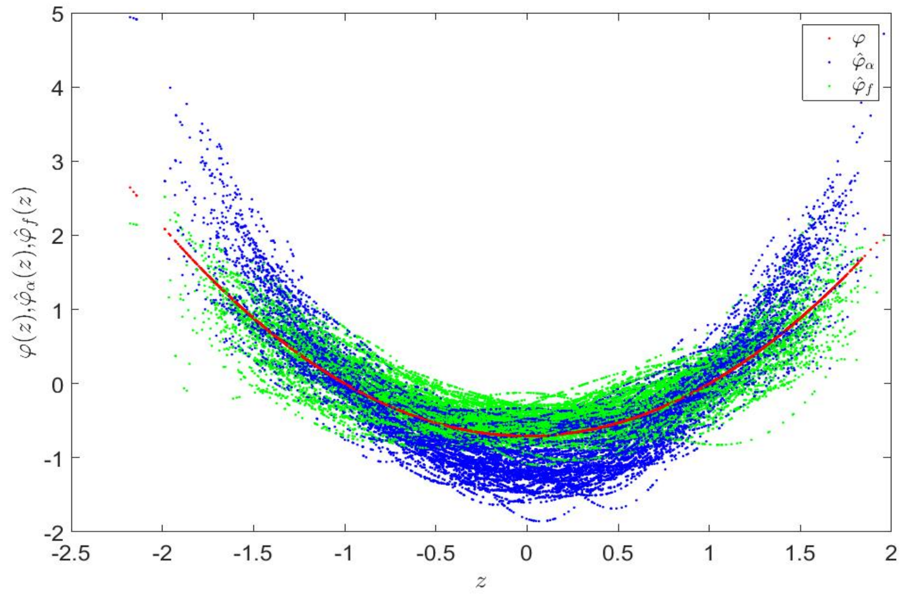
Note: α is selected in order to minimize the MISE of the first step estimator.

Figure 8: **Simulation result with one draw for** α_{opt}^{2s}



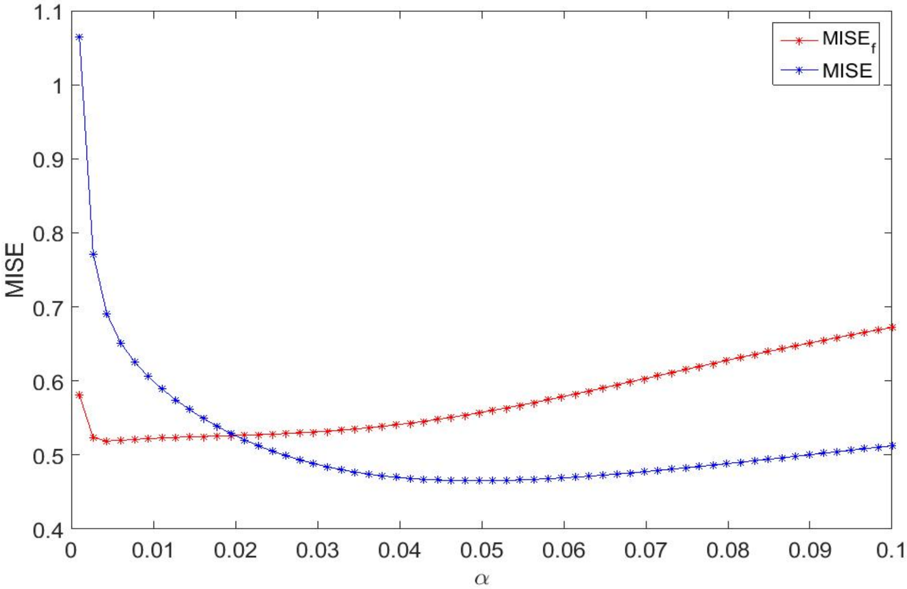
Note: α is selected in order to minimize the MISE of the second step estimator.

Figure 9: **Simulation result with 100 draws**



Note: α is selected in order to minimize the MISE of the second step estimator. Green dots are the estimated curve by using optimal feasible estimator at each draw while the blue dots are unweighted estimates.

Figure 10: α **vs. MISE**


MISE_f
MISE
MISE
α