Optimal weighting for linear inverse problems*

Jean-Pierre FLORENS† Toulouse School of Economics

Senay SOKULLU‡ University of Bristol

August 5, 2021

Abstract

Linear equations in functional spaces where the solution is not continuous require regularization to estimate the unknown function of interest. In this paper we consider the estimation of an infinite dimensional parameter φ by solving a linear equation $\hat{r} = K\varphi + U$, where the random noise U has a variance Σ. Under this set-up, we derive the optimal weighting operator which minimizes the mean integrated square error (MISE). In the finite dimensional case the minimum variance estimator is obtained by weighting the equation by $\Sigma^{-1/2}$. However in the infinite dimensional case that we consider, regularization introduces a bias to the estimator. We show that in the infinite dimensional case, the optimal estimator in terms of the MISE should involve Σ and the unknown smoothness of φ. We then use this result to propose a new feasible two-step estimator. We illustrate our theoretical findings and the small sample properties of the proposed optimal estimator by means of simulations.

Keywords: Ill-posed inverse problems, Nonparametric Methods, Nonparametric IV Regression, Tikhonov Regularization, Regularization Parameter

JEL Classification: C13, C14, C26

*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.

†Email:jean-pierre.florens@tse-fr.eu; Address: 1 Esplenade de L'Universite, 31080 Toulouse Cedex 06, FRANCE

‡Email: senay.sokullu@bristol.ac.uk, Address: University of Bristol, Priory Road Complex, Priory Road, Bristol BS8 1TU, UK

1 Introduction

In this paper we derive an optimal estimator for linear inverse problems which minimizes the small sample mean integrated square error (MISE). Under the Generalized Least Squares (GLS) approach, the minimum variance (optimal) estimator is obtained by weighting the sum of squares by the inverse of the variance of the residuals. We show that for linear inverse problems such as nonparametric instrumental variables regression, weighting by the inverse of the variance of the residuals is no longer optimal due to the bias-variance trade-off and in such a case the optimal weighting should take into account the regularity of the functional parameter.

The linear inverse problems which are considered in this paper can be viewed as an extension of the Generalized Method of Moments, see Hansen (1982). 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 $\hat{\beta} = argmin_\beta \|w_i(y_i - z_i\beta)\|^2_{\Omega_n}$ for any symmetric and positive definite weighting matrix $\Omega_n \xrightarrow{p} \Omega$.[1] 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

[1] $\|.\|_\Omega$ denotes Euclidean norm, i.e., $\|x\|^2_\Omega = x'\Omega x$.

paper, following Hansen (1982), we obtain an optimal weighting which minimizes the small sample MISE when the dimensions of z_i and w_i are large or infinite. Given our result, we propose optimal infeasible and feasible estimators and study their asymptotic properties. Our optimality result is in terms of small sample MISE and it is not rate optimality.

Solving an ill-posed inverse problem requires regularization, for instance in the case of NPIV, see Darolles, Fan, Florens, and Renault (2011); Newey and Powell (2003); Ai and Chen (2003) and Horowitz (2011) among others. 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, in other words, the bias-variance trade-off. We then derive the optimal weighting operator which leads to minimum small sample MISE 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 the 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 value of regularization parameter is smaller and hence the effect is ambiguous. In this paper, we first derive the small sample 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 propose infeasible and feasible optimal estimators and show that they are both consistent under a general set-up. Finally, we investigate our theoretical findings by means of simulations.

This paper can be related to several strands of the literature. First, as al-

ready stated, one example of linear inverse problems is nonparametric IV estimation. Hence, this paper is related to the nonparametric instrumental variables literature, see Darolles et al. (2011); Newey and Powell (2003); Ai and Chen (2003) and Horowitz (2011) among others. All these papers show that the estimators they use are consistent, however none of them considers the small sample MISE optimality of the estimator of the infinite dimensional parameter. In this paper, we consider the "optimality" in terms of the small sample MISE of the infinite dimensional parameter and it is not a minimax or asymptotic argument. Rate optimality of the functional parameter has been well studied in papers such as Hall and Horowitz (2005); Chen and Reiss (2011) and Chen and Christensen (2018).

With the growth of the nonparametric IV literature in 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 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, as we do in this paper. Gagliardini and Scaillet (2017) use a weighting operator given by the inverse of the conditional variance of the moment conditions in a NPIV setting, however they do mention that this choice of the weighting function is not motivated by the notion of efficiency as in the parametric case.

The choice of regularization parameter is crucial in ill-posed inverse problems, and as a result, adaptive estimation has been studied extensively both in the econometrics and statistics literature. For instance, Horowitz (2014) and Chen and Christensen (2015) examine the selection of a regularization parameter in a NPIV model where the infinite dimensional parameter is estimated using sieve methods and show that the adaptive estimator could reach near-optimal (Horowitz, 2014) and uniform-optimal

(Chen and Christensen, 2015) convergence rates. This paper is related to the adaptive estimation literature as we show that selection of an optimal weighting operator replaces the selection of an optimal regularization parameter, and also we derive the convergence rates of the proposed estimator. Although this paper is closer to the statistics literature on adaptive estimation, since we assume that the operator we work with is known and not estimated ,contrary to the econometrics literature, see for instance Bauer and Hohage (2005); Spokoiny, Vial, et al. (2009).

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 infeasible and feasible estimators and present the example of NPIV. In *Section 4*, we present simulation results which demonstrate our theoretical findings as well as the small sample properties of the optimal feasible estimator. Finally, in *Section 5* we conclude. All proofs are presented in Appendix A.

2 The Set-up

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 general setting can also be shown to fit with cases such as deconvolution problems (Carrasco, Florens, and Renault, 2007), functional linear regression (Hall and Horowitz, 2007) or functional instrumental variables regression (Florens and Van Bellegem, 2015).

Consider a linear inverse problem of the form:

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

such that $\varphi \in \mathcal{E}$; $\hat{r}$ 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.[2] 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] $\frac{1}{n}$ 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.

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 $\varphi \in \mathcal{D}(L)$.

In the 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:

$$\min_{\varphi \in \mathcal{D}(L)} \|A\hat{r} - AK\varphi\|^2 + \alpha \|L\varphi\|^2 \tag{2}$$

and it is equal to:

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

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

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

Here we consider Tikhonov regularization with a Hilbert scale penalty. This approach leads to regularization with a smooth norm as well as giving higher convergence rates than with Tikhonov regularization with an 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 the 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.

The introduction of a differential operator L in the penalty term is a common practice in the resolution of ill-posed inverse problems, see Engl, Hanke, and Neubauer (1996). This approach has several advantages: i) If we assume some smoothness property for the solution (for example, $\varphi \in \mathbb{R}(L^{-1})$) this method guarantees that the estimator satisfies the same smoothness. ii) One could estimate jointly both φ and its derivative, $L\varphi$. iii) The qualification of the method, which can be defined as the maximum order of regularity which controls the rate of the regularization bias (Carrasco et al., 2007), increases by the introduction of L. Gagliardini and Scaillet (2012) advocate penalisation in the Sobolev norm to suppress the highly osciallting component of the estimated function. They further show with Monte

Carlo simulations that Tikhonov regularization with a Sobolev penalty increases the performance of the estimator compared to that which is obtained with Tikhonov regularization with an L^2 penalty.

In what follows, we work with the spectral representation of the model. For 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 basis of $\mathcal{E}$ and ψ_j is an orthonormal basis 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 a continuous spectrum. Indeed, the main assumption is that $A^*AK\phi_j = \tilde{\psi}_j$ constitutes an orthogonal family in $\mathcal{F}$. In this case, one can normalize the $\tilde{\psi}_j$ in ψ_j and there exist 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 the 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:

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

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

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

where $E(\epsilon_j) = 0$, $Var(\epsilon_j) = 1$. The representation given in Equation 6 is standard in the literature of inverse problems. In particular, in statistical models, the noise U is assumed to be random rather than deterministic which is, in general, the case in the ill-posed inverse problem literature. Hence, this notation captures the fact that the model in Equation 1 can be written as a 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, and van Zanten (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:*

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

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 only consider weighting by A, but our results 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, see Section 3.2. 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^2_{Aj} 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 only due to the minimization of the variance, 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.

Throughout the paper, we assume that K and Σ are known. This is the case for numerous inverse problems and most of the literature in this field is limited to this case of known K and Σ, see Cavalier, Golubev, Picard, and Tsybakov (2002). For example, in image treatment models such as tomography, the operator K is given. In some statistical applications of the inverse problems, for instance density estimation ($K\varphi = \int_0^x \varphi(u)du$) or deconvolution models where the distribution of the error term is given ($K\varphi = \int \varphi(s)f(t-s)ds$), the operator K is naturally given. This is also the case in functional linear regression where K depends on the sample size, see Benatia, Carrasco, and Florens (2017).

In some econometric applications K and Σ are unknown but are estimated using another sample or part of the sample. For example, in the Nonparametric Instrumental Variables Regression example in Section 3.2, one observes (y_i, z_i, x_i) where y_i and z_i are endogenous and x_i are the instruments. The estimation of K and Σ is done by using only x_i and z_i whereas the estimation of φ uses only y_i given the estimate of K, $\hat{K}$. All the results developed in this paper would hold if $E(\hat{r} - \hat{K}\varphi|\hat{\Sigma}, \hat{K}) = 0$ and $Var(\hat{r} - \hat{K}\varphi|\hat{\Sigma}, \hat{K}) = \frac{1}{n}\hat{\Sigma}$. In some cases these above conditions hold only up to a "small" element, for example, $E(\hat{r} - \hat{K}\varphi|\hat{\Sigma}, \hat{K}) = 0$ depends on a bandwidth choice through the estimation of a conditional expectation. In this paper, we do not examine or develop the conditions under which these small elements are negligible. It should also be noted that Σ cannot be estimated by the variance of $\hat{r} - \hat{K}\hat{\varphi}_\alpha$, where $\hat{\varphi}_\alpha$ is a first step estimate. This is because we have a single observation of this residual. Some models put a structure on Σ, e.g. in the NPIV case $\Sigma = \sigma^2 K$, see Section 3.2. However even in this case, only a part of Σ could be obtained from a first step estimation (the variance σ^2 in the NPIV model).

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 Optimization

It is shown in Proposition 1 that weighting by A or penalising by LA^{-1} is equivalent. For the sake of exposition, in the rest of the paper we consider the case with weighting by A only, i.e., without Hilbert scale penalty. 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 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}. Dropping L, MISE in Proposition 1 can be written as:

$$\mathbb{E}\|\hat{\varphi}_\alpha - \varphi\|^2 = \frac{1}{n}\sum_{j=1}^{\infty} \frac{\langle \Sigma\psi_j, \psi_j\rangle \lambda_{Kj}^2 \lambda_{Aj}^4}{(\alpha + \lambda_{Kj}^2\lambda_{Aj}^2)^2} + \alpha^2 \sum_{j=1}^{\infty} \frac{\langle \varphi, \phi_j\rangle^2}{(\alpha + \lambda_{Kj}^2\lambda_{Aj}^2)^2}. \tag{9}$$

This MISE expression leads to the following result:

Proposition 2 *Consider the MISE expression given in (9) under Assumption 1. Then:*

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 (infeasible) estimator:*

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

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

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:

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

This result differs from the standard result for GMM. In the usual finite-dimensional 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 infeasible because it depends on the unknown function φ. The estimator $\hat{\varphi}_{if}$ 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 $\hat{\varphi}_{if}$ can be interpreted as a Hilbert scale type extension of Tikhonov estimation. Indeed, $\hat{\varphi}_{if}$ is the argument φ that minimizes the following:

$$\hat{\varphi}_{if}=argmin_{\varphi}\|\hat{r}-K\varphi\|^2+\frac{1}{n}\|Q^{1/2}\varphi\|^2.$$

Note that our result remains valid even if some $\langle\varphi,\phi_j\rangle=0$. To illustrate this, let us assume that $\langle\varphi,\phi_j\rangle\neq 0$ for $j\in J$ and $\langle\varphi,\phi_j\rangle=0$ for $j\in\bar{J}$, so one can say that φ belongs to $\mathcal{E}_J$, a subspace generated by the ϕ_j such that $\langle\varphi,\phi_j\rangle\neq 0$. In this case the λ_{Aj}'s cancel out for $j\in\bar{J}$ and the optimal infeasible estimator belongs to $\mathcal{E}_J$. Our approach constructs an (infeasible) estimator which satisfies the constraint $\varphi\in\mathcal{E}_J$. In Section 3.1 we define a feasible estimator (given in Equation 11) which includes the terms $\langle\hat{r},\psi_j\rangle$ and these scalar products converge to $\langle r,\psi_j\rangle$ and $\langle r,\psi_j\rangle=0$ if $\langle\varphi,\phi_j\rangle=0$. Our feasible estimator approximately satisfies the constraint $\varphi\in\mathcal{E}_J$ even if this constraint is unknown. In the case of infeasible estimator all the formulae of Proposition 2 remain valid if some $\langle\varphi,\phi_j\rangle=0$ and then the sum $\sum_{j=0}^{\infty}$ can be replaced by $\sum_{j\in J}$. It should be noted that the operator A is not injective but AK remains injective on the set $\mathcal{E}_J$ and all the theory can be developed replacing $\mathcal{E}$ by $\mathcal{E}_J$.

Moreover, 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 oper-

ator (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 penalizing 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, in this case where φ is sufficiently smooth, the optimal weighting can be interpreted as the optimal norm. More precisely, given a regularization parameter α, our result suggests that it is optimal to use a Sobolev penalty.[3]

Regarding the consistency of $\hat{\varphi}_{if}$, 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 infeasible estimator as well as for the 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). In this paper we are considering statistical inverse problems, where U is assumed to be random. Hence, the variance of U matters for the solution. Assumption 2 incorporates the variance of U in the source condition as it 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, Σ. Hence Assumption 2 can be seen

[3]We thank Demian Pouzo for pointing this out.

as an extended source condition. The next theorem states the rate of convergence of $\hat{\varphi}_{if}$ under this extended source condition.

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

$$E\|\hat{\varphi}_{if} - \varphi\|^2 = O(n^{-\beta}).$$

Theorem 1 shows that the infeasible estimator is consistent and it converges at a rate of $n^{-\beta}$ which is slower than the usual parametric rate. This result is not surprising because the optimal infeasible estimator is still a nonparametric estimator, and by weighting we optimize its small sample MISE, not its asymptotic MISE.

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 the 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 the 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:

$$\langle\hat{\varphi}, \phi_j\rangle = \frac{\lambda_{Kj}}{\alpha + \lambda_{Kj}^2}\langle\hat{r}, \psi_j\rangle. \tag{10}$$

Note that 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 if r could be estimated with a $\sqrt{n}$-rate. Using (10), the feasible estimator is equal to:

$$\hat{\varphi}_f = \sum_j \frac{\lambda_{Kj}^3\langle\hat{r}, \psi_j\rangle^3}{\frac{1}{n}(\alpha + \lambda_{Kj}^2)^2\langle\Sigma\psi_j, \psi_j\rangle + \lambda_{Kj}^4\langle\hat{r}, \psi_j\rangle^2}\phi_j. \tag{11}$$

As can be seen from Equation 11, the feasible estimator does depend on α through its dependence on the 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 $\frac{\lambda_{Kj}}{\alpha+\lambda_{Kj}^2}$. Hence, the feasible estimator $\hat{\varphi}_f$ is a regularized estimator where $\frac{1}{\lambda_{Kj}}$ is replaced by:

$$\frac{\lambda_{Kj}}{\frac{1}{n}(\alpha+\lambda_{Kj}^2)^2\frac{\langle\Sigma\psi_j,\psi_j\rangle}{\lambda_{Kj}^2\langle\hat{r},\psi_j\rangle^2}+\lambda_{Kj}^2}. \tag{12}$$

Equation (12) 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. Theorem 2 below states the consistency of the optimal feasible estimator and Theorem 3 shows that the MISE of the optimal feasible estimator converges to that of the optimal infeasible estimator.

Theorem 2 *Consider the feasible estimator given in Equation 11. Assume that α is fixed. Then under Assumption 1 as $n \to \infty$:*

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

Theorem 2 shows that the optimal feasible estimator is consistent and that consistency can be achieved with a fixed regularization parameter, in other words, we do not need $\alpha \to 0$. As 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 the optimal regularization parameter.

Theorem 3 *Consider the feasible estimator given in Equation 11. Assume that we have a sample of size $2n$. Assume moreover that we use the first half of the sample to estimate $\hat{\varphi}_\alpha$ and then use the second half of the sample to estimate $\hat{\varphi}_f$ where φ is replaced by $\hat{\varphi}_\alpha$. Then under Assumptions 1 and 2, the feasible estimator is optimal:*

$$MISE(\hat{\varphi}_f) - MISE(\hat{\varphi}_{if}) = O_p(n^{-\beta})$$

Three points related to Theorem 3 are worth discussing. First, it should be noted that as the optimal feasible estimator ($\hat{\varphi}_f$) requires the first step estimator ($\hat{\varphi}_\alpha$) to be plugged in, to be able to analyze the MISE we assume that $\hat{\varphi}_\alpha$ is obtained using a separate sample. Although in Theorem 3 it is stated that we divide the sample into two equally, the result will hold if we take any fraction of the sample ($c \times n$ where

$0 < c < 1$) to estimate $\hat{\varphi}_\alpha$. Second, Theorem 3 shows that the MISE of the optimal feasible estimator reaches that of the oracle estimator. In other words, even though the MISE of $\hat{\varphi}_f$ is larger than MISE of $\hat{\varphi}_{if}$, it has the same order ($n^{-\beta}$). Third, this result does not contradict the previous results on minimax rates obtained in papers such as Hall and Horowitz (2007); Chen and Reiss (2011); Chen and Christensen (2018); as it builds on a different set of assumptions (such as Assumption 2).

3.2 Example: Nonparametric IV Regression

In this section we consider optimal weighting in a 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 in terms of minimum MISE.[4] Below, we present optimal infeasible and feasible estimators under this setup.

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

$$Y = \varphi(Z) + V \quad and \quad \mathbb{E}(V|X) = 0. \tag{13}$$

The model then generates a linear inverse problem:

$$\mathbb{E}(\mathbb{E}(Y|X)|Z) = \mathbb{E}(\mathbb{E}(\varphi(Z)|X)|Z), \tag{14}$$

$$r = K\varphi, \tag{15}$$

where $r \in L_Z^2$, $\varphi \in L_Z^2$ and $K : L_Z^2 \mapsto L_Z^2$. 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. In this case, one can write:

$$\hat{r} = K\varphi + U. \tag{16}$$

We assume that $\mathbb{E}(U) = 0$. The operator K is a self-adjoint trace class operator. This NPIV model is studied in detail in Darolles et al. (2011) and it is shown that $\mathbb{V}(U) = \frac{\sigma^2}{n}K$ under some regularity conditions including homoskedasticity of the

[4] As already mentioned, in this paper we consider optimality in terms of minimal small sample MISE, not rate-optimality.

variance of V conditional on X, $\mathbb{V}(V|X) = \sigma^2$.

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 would result in $K^*A^*AK = K$, with 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 infeasible estimator can be written as:

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

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 infeasible estimator can be rewritten as:

$$\hat{\varphi}_{if,IV} = \left(\frac{\sigma^2}{n} K + RK^*K \right)^{-1} RK^* \hat{r}. \tag{18}$$

To obtain the feasible estimator, $\langle \varphi, \phi_j \rangle$ in Equation 17 is replaced by $\frac{\lambda_{Kj}}{\alpha + \lambda_{Kj}^2} \langle \hat{r}, \psi_j \rangle$:

$$\hat{\varphi}_{f,IV} = \sum_{j=1}^{\infty} \frac{\langle \hat{r}, \psi_j \rangle^2}{\frac{\sigma^2}{n} (\alpha + \lambda_{Kj}^2)^2 + \lambda_{Kj}^3 \langle \hat{r}, \psi_j \rangle^2} \lambda_{Kj}^2 \langle \hat{r}, \psi_j \rangle \phi_j. \tag{19}$$

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:

$$\hat{\varphi}_{f,IV} = \left(\frac{\sigma^2}{n} (\alpha I + K^2)^2 + K^3 T \right)^{-1} KTK^* \hat{r}. \tag{20}$$

4 Simulations

In this section we present Monte Carlo simulations to show the performance of the proposed optimal feasible estimator compared to that of the unweighted estimator. We first generate data from a NPIV model and estimate the unknown function for a given operator using weighted feasible and unweighted estimators. We then simulate the design in Newey and Powell (2003) where K is known to be from a normal family but with an unknown variance. This is the case where K is partially known. The results of the Monte Carlo experiment show that the optimal feasible estimator

performs better than the unweighted estimator.

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, V is drawn from a univariate normal distribution with 0 mean and variance equal to 0.1.[5] We set $\varphi(Z)$ to be equal to $\varphi(Z) = \frac{Z^2-1}{\sqrt{2}}$. Then Y is given by:

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

Partially Known K: We simulate the design in Newey and Powell (2003). V, η and X are drawn from a normal distribution with mean $\begin{pmatrix} 0 & 0 & 0 \end{pmatrix}'$ and variance:

$$\begin{pmatrix} 1 & 0.5 & 0 \\ 0.5 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Moreover Z is given by $Z = X + \eta$. Finally φ is set to be equal to:

$$\varphi(Z) = ln(|Z - 1| + 1)sign(z - 1).$$

So, in this design the operator K still comes from a normal family but with an unknown variance, i.e. the ρ coefficient in the known K design. Hence we instead estimate ρ from data. Then the eigenvalues are given by $\lambda_{Kj} = \hat{\rho}^{2j}$.

In both simulations, we choose a geometric spectrum, so the eigenvalues are given by $\lambda_{Kj} = \rho^{2j}$ (or by $\lambda_{Kj} = \hat{\rho}^{2j}$) 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:

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

Then the scalar product can be written as:

$$\langle \hat{\varphi}_\alpha, \phi_j \rangle = \frac{\lambda_{Kj}}{\alpha + \lambda_{Kj}^2} \left(\frac{1}{n} \sum_{i=1}^n y_i \phi_j(z_i) \lambda_{Kj}^{1/2} \right). \tag{22}$$

[5] We also tried different values for the variance of V and our results stay the same qualitatively.

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

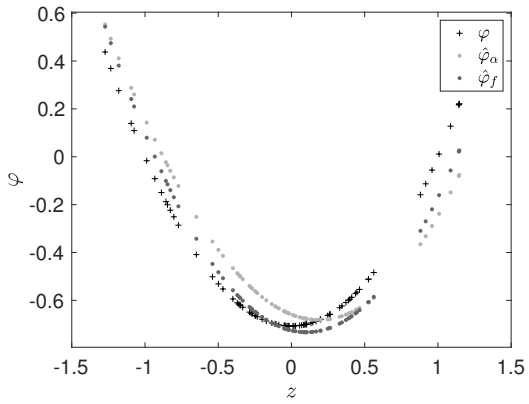
$$\hat{\varphi}_f(z) = \sum_j \frac{\langle \hat{\varphi}_\alpha, \phi_j \rangle^2 \lambda_{Kj} \left(\frac{1}{n} \sum_{i=1}^n y_i \phi_j(z_i) \lambda_{Kj}^{1/2} \right)}{1/n\hat{\sigma}_v^2 \lambda_{Kj} + \langle \hat{\varphi}_\alpha, \phi_j \rangle^2 \lambda_{Kj}^2} \phi_j(z). \tag{23}$$

We replicate this exercise 500 times for sample sizes equal to 100, 200 and 400. We truncate the sum at $j = 15$. 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 1 shows the Root Mean Integrated Square Error (RMISE) of the unweighted and weighted optimal estimators under the 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 squared residuals of the second step estimator. The RMISE of the optimal feasible estimator when the regularization parameter is selected in the second stage is smaller than that of the unweighted estimator even when the regularization parameter is selected so as to minimise the RMISE of the unweighted estimator. Moreover, when α is selected in the second stage, we see that the RMISE of the feasible estimator gets smaller as the sample size increases - as is the case for the RMISE of $\hat{\varphi}_\alpha$ when α is selected in the first stage. The average α chosen with respect to the first step estimator is larger than the average α chosen with respect to the second step estimator, showing that the feasible estimator requires oversmoothing in the first step. Figure 1 shows plots of the estimators for a single draw for the different selection rules of the regularization parameter. Figure 2 shows the plots of $\hat{\varphi}_f$ and $\hat{\varphi}_\alpha$ for all draws over the true function where α is chosen to minimise the MISE of $\hat{\varphi}_f$.

Figure 1: **Simulation result with one draw - K known**

(a) α_{opt} 1st Stage



(b) α_{opt} 2nd Stage

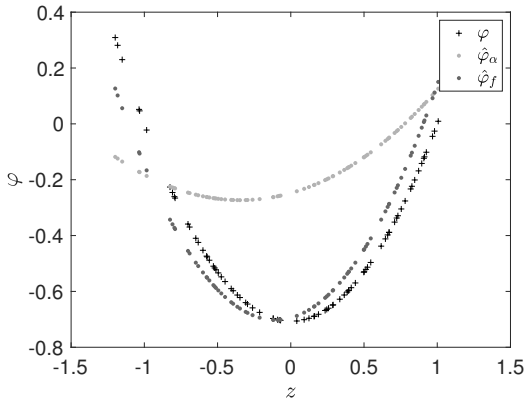
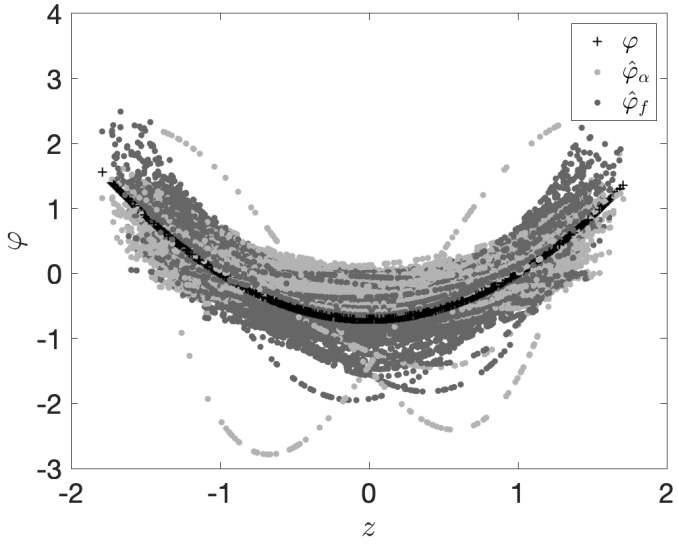


Figure 2: **Simulation result with 500 draws**



Note: α is selected in order to minimize the MISE of the second step estimator. Black pluses are the true values of the φ function. Dark gray dots are the estimated curve using the optimal feasible estimator at each draw while the light gray dots are unweighted estimates.

The results for the case where K is partially known are presented in Table 2 and Figures 3 and 4. The optimal feasible estimator performs better irrespective of how α is selected. Moreover, as in the case with known K, to get a better fit for $\hat{\varphi}_f$, one

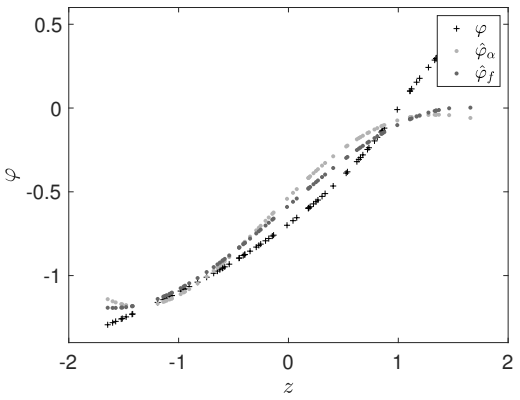
Table 1: **Simulation results - K known**

	α_{opt} first stage			α_{opt} second stage		
	MISE			MISE		
	$\hat{\varphi}_\alpha$	$\hat{\varphi}_f$	α_{opt}	$\hat{\varphi}_\alpha$	$\hat{\varphi}_f$	α_{opt}
$n = 100$	0.4171	0.6816	0.0073	0.5891	0.4153	0.0291
$n = 200$	0.3515	0.6851	0.0065	0.5721	0.2835	0.0317
$n = 400$	0.2891	0.5868	0.0047	0.5977	0.1958	0.0348

needs to oversmooth at the first stage as α_{opt}^{2s} is larger than α_{opt}^{1s}. Figures 3 and 4 show that our feasible estimator works well even when the sample size is equal to 100.

Figure 3: **Simulation result with one draw - K partially known**

(a) α_{opt} 1st Stage



(b) α_{opt} 2nd Stage

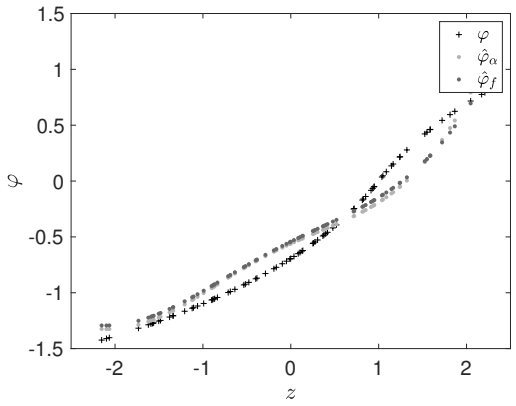
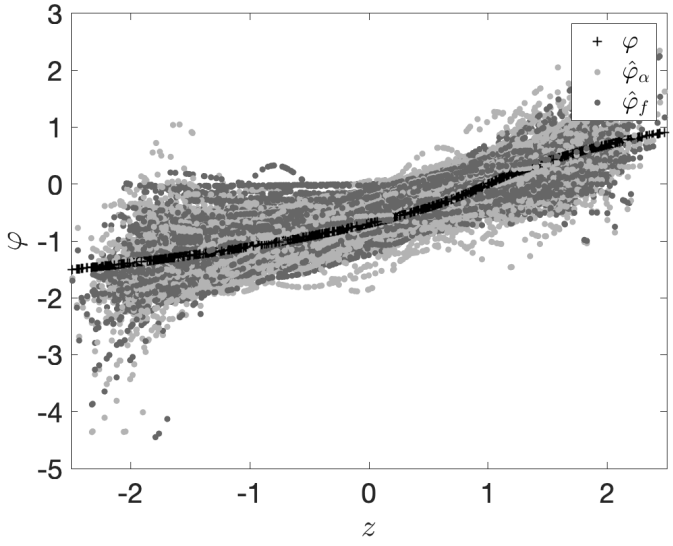


Figure 4: **Simulation result with 100 draws**



Note: α is selected in order to minimize the MISE of the second step estimator. Black pluses are the true values of the φ function. Dark gray dots are the estimated curve using the optimal feasible estimator at each draw while the light gray dots are unweighted estimates.

Table 2: **Simulation results - K partially known**

	α_{opt} first stage			α_{opt} second stage		
	MISE			MISE		
	$\hat{\varphi}_\alpha$	$\hat{\varphi}_f$	α_{opt}	$\hat{\varphi}_\alpha$	$\hat{\varphi}_f$	α_{opt}
$n = 100$	1.0183	0.5607	0.0764	0.9615	0.5314	0.0702
$n = 200$	2.3081	0.4687	0.0745	1.5665	0.4680	0.0849
$n = 400$	1.4939	0.4810	0.0741	2.1068	0.4744	0.0886

We can conclude that both for a known K and a partially known K, the optimal feasible estimator performs better than the unweighted estimator. However, in an empirical application, it is unlikely that one knows K. 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.

5 Conclusion

In this paper we study small sample MISE optimality in linear inverse problems and we derive the weighting operator which leads to minimal MISE. We have several important findings. First, under a very general set-up, we have shown that weighting and a Hilbert scale penalty play the same role. Hence, one may fix one of these operators to identity. Second, 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 proposed an optimal feasible estimator. Fourth, we study the asymptotic properties of our proposed feasible estimator and show that it is consistent. While doing this, we also introduce a new type of source condition which do not only take into account the smoothness of the functional parameter but also the variance of the noise. 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 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 a nonparametric three stage least squares estimator using this optimal weighting matrix is also left for future work.

References

Ai, C. and X. Chen (2003, November). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica 71*(6), 1795–1843.

Bauer, F. and T. Hohage (2005). A lepskij-type stopping rule for regularized newton methods. *Inverse Problems 21*(6), 1975.

Benatia, D., M. Carrasco, and J.-P. Florens (2017). Functional linear regression with functional response. *Journal of econometrics 201*(2), 269–291.

Carrasco, M., J.-P. Florens, and E. Renault (2007, June). Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization. In J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics*, Volume 6 of *Handbook of Econometrics*, Chapter 77. Elsevier.

Cavalier, L. (2008). Nonparametric statistical inverse problems. *Inverse Problems 24*(3), 034004.

Cavalier, L., G. Golubev, D. Picard, and A. Tsybakov (2002). Oracle inequalities for inverse problems. *The Annals of Statistics 30*(3), 843–874.

Chen, X. and T. Christensen (2015). Optimal sup-norm rates, adaptivity and inference in nonparametric instrumental variables estimation.

Chen, X. and T. M. Christensen (2018). Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric iv regression. *Quantitative Economics 9*(1), 39–84.

Chen, X. and D. Pouzo (2009, September). 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*(3), 497–521.

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

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

Florens, J.-P., J. Johannes, and S. Van Bellegem (2012, 06). Instrumental regression in partially linear models. *Econometrics Journal 15*(2), 304–324.

Florens, J.-P. and S. Van Bellegem (2015, 06). Instrumental variable estimation in functional linear models. *Journal of Econometrics 186*(2), 465–476.

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

Gagliardini, P. and O. Scaillet (2017). A specification test for nonparametric instrumental variable regression. *Annals of Economics and Statistics/Annales d'Économie et de Statistique* (128), 151–202.

Hall, P. and J. L. Horowitz (2005). Nonparametric methods for inference in the presence of instrumental variables. *Annals of Statistics 32*, 2904–2929.

Hall, P. and J. L. Horowitz (2007). Methodology and convergence rates for functional linear regression. *The Annals of Statistics 35*(1), 70–91.

Hansen, L. P. (1982, July). Large sample properties of generalized method of moments estimators. *Econometrica 50*(4), 1029–54.

Horowitz, J. L. (2011, March). Applied nonparametric instrumental variables estimation. *Econometrica 79*, 347–394.

Horowitz, J. L. (2014). Adaptive nonparametric instrumental variables estimation: Empirical choice of the regularization parameter. *Journal of Econometrics 180*(2), 158–173.

Knapik, B. T., A. W. van der Vaart, and J. H. van Zanten (2011). Bayesian inverse problems with gaussian priors. *The Annals of Statistics 39*(5), 2626–2657.

Krein, S. G. and Y. I. Petunin (1966). Scales of banach spaces. *Russian Mathematical Surveys 21*(2), 85.

Neubauer, A. (1988). When do sobolev spaces form a hilbert scale? *Proceedings of the American Mathematical Society 103*(2), 557–562.

Newey, W. K. and J. L. Powell (2003, 09). Instrumental variable estimation of nonparametric models. *Econometrica 71*(5), 1565–1578.

Newey, W. K. and R. J. Smith (2004). Higher order properties of gmm and generalized empirical likelihood estimators. *Econometrica 72*(1), 219–255.

Spokoiny, V., C. Vial, et al. (2009). Parameter tuning in pointwise adaptation using a propagation approach. *The Annals of Statistics 37*(5B), 2783–2807.

Appendices

A Proofs

A.1 Proof of Proposition 1

Proof.

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

where $\hat{\varphi}_\alpha = L^{-1}(\alpha I + L^{-1}K^*A^*AKL^{-1})^{-1}L^{-1}K^*A^*AK\varphi$ and $\mathbb{V}(.)$ denotes the variance. Using some elementary manipulations and the property that L^{-1} commutes with K^*A^*AK we get:

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

where $B = L^{-1}K^*A^*AKL^{-1}$. Given Assumption 1 and the fact that $tr(\Omega) = \sum_{j=1}^{\infty}\langle\Omega\delta_j, \delta_j\rangle$, the result follows from the definition of MISE above. ■

A.2 Proof of Proposition 2

Proof. The proof follows from the minimization of the MISE given in Equation 9 with respect to λ_{Aj}^2. The first order condition is given by:

$$\frac{\frac{2}{n}\langle\Sigma\psi_j, \psi_j\rangle(\alpha + \lambda_{Kj}^2\lambda_{Aj}^2)\lambda_{Kj}^2\lambda_{Aj}^2\alpha}{(\alpha + \lambda_{Kj}^2\lambda_{Aj}^2)^4} - \frac{2\alpha^2\langle\varphi, \phi_j\rangle^2(\alpha + \lambda_{Kj}^2\lambda_{Aj}^2)\lambda_{Kj}^2}{(\alpha + \lambda_{Kj}^2\lambda_{Aj}^2)^4} = 0.$$

After rearranging one can obtain:

$$\frac{1}{n}\langle\Sigma\psi_j, \psi_j\rangle\lambda_{Aj}^2 = \alpha\langle\varphi, \phi_j\rangle^2.$$

Then the result follows:

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

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 optimal 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.3 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 Equation 9, we obtain the MISE of the optimal infeasible estimator:

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

where the first term is the variance and the second term is the bias squared. The rest of the proof treats these two terms separately. Starting with the bias, if ones divides and multiplies it by $\frac{\langle\varphi,\phi_j\rangle^2}{\langle\Sigma\psi_j,\psi_j\rangle^2}$, the following can be obtained after some manipulation:

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

Denote $x_j=\frac{\langle\varphi,\phi_j\rangle^2}{\langle\Sigma\psi_j,\psi_j\rangle}\lambda_{Kj}^2$ and divide and multiply Equation A.1 by x_j^β:

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

where $\frac{x_j^\beta}{(1/n+x_j)^2}$ is $O(n^{2-\beta})$. Then the whole bias term is $O(n^{-\beta})$ and under assumption 2, bias term 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:

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

Replacing $\frac{\langle\varphi,\phi_j\rangle^2}{\langle\Sigma\psi_j,\psi_j\rangle}\lambda_{Kj}^2$ by x_j and dividing and multiplying Equation A.2 by x_j^β, one

obtains:

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

The term $\frac{x_j^{\beta+1}}{(1/n+x_j)^2}$ is $O(n^{-1})$ and the whole variance term is $O(n^{-\beta})$. Thus under Assumption 2 the variance term as well vanishes as $n\to\infty$. ■

A.4 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:

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

The proof follows showing both terms, A and B, converge to zero. Starting 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, $g_\rho(x)$ can be written as:

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

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:

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

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

$$\frac{1}{n} \sup g_\rho(x) = \frac{x^* \langle \hat{r}, \psi_j \rangle^2}{(\alpha + x^*)^2 \langle \Sigma\psi_j, \psi_j \rangle + \frac{\alpha^2 \langle \Sigma\psi_j, \psi_j \rangle}{\langle \hat{r}, \psi_j \rangle^2 + \frac{1}{n} \langle \Sigma\psi_j, \psi_j \rangle} \langle \hat{r}, \psi_j \rangle^2} \tag{A.3}$$

First note that x^* 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 a fixed α. Then we can conclude that the numerator of Equation A.3 is of order $O(1)$. As for the denominator of A.3, the second term dominates so one can examine just that term:

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

where $\tilde{r} = n\hat{r}$. For a fixed α, this term is $O_p(n)$. Hence it can be concluded that A.3 is $O_p(1/n)$:

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

■

A.5 Proof of Theorem 3

Proof. Using result two of Proposition 2, the infeasible estimator is given by

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

where Q is the operator:

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

Let us write the feasible estimator $\varphi_f = (\frac{1}{n}\hat{Q} + K^*K)^{-1}K^*\hat{r}$ and let us denote eigen values of the operator Q and $\hat{Q}$ by ν_j and $\hat{\nu}_j$, respectively. Then the MISE of the infeasible estimator can be written as:

$$MISE(\varphi_{if}) = \frac{1}{n}\sum_{j=1}^{\infty} \frac{1}{(\frac{1}{n}\nu_j + \lambda K j^2)^2}[\lambda_{Kj}^2\langle \Sigma\psi_j, \psi_j\rangle + \frac{1}{n}\nu_j^2\langle \varphi, \phi_j\rangle^2]$$

A second order Taylor expansion of MISE(φ_f) around MISE(φ_{if}) leads to:

$$\begin{aligned} MISE(\varphi_f) - MISE(\varphi_{if}) &= \frac{2}{n^2}\sum_{j=1}^{\infty} \frac{\lambda_{Kj}^2}{\left(\frac{1}{n}\nu_j + \lambda_{Kj}^2\right)^3}(\nu_j\langle \varphi, \phi_j\rangle^2 - \langle \Sigma\psi_j, \psi_j\rangle)(\hat{\nu}_j - \nu_j) \\ &\quad + \frac{1}{n^2}\sum_{j=1}^{\infty} \frac{\lambda_{Kj}^2}{(\frac{1}{n}\nu_j + \lambda_{Kj}^2)^3}\langle \varphi, \phi_j\rangle^2(\hat{\nu}_j - \nu_j)^2 + o\left((\hat{\nu}_j - \nu_j)^2\right) \end{aligned}$$

The term $(\nu_j\langle \varphi, \phi_j\rangle^2 - \langle \Sigma\psi_j, \psi_j\rangle)$ is equal to zero as $\nu_j = \frac{\langle \Sigma\psi_j, \psi_j\rangle}{\langle \varphi, \phi_j\rangle^2}$, so:

$$MISE(\varphi_f) - MISE(\varphi_{if}) = \frac{1}{n^2}\sum_{j=1}^{\infty} \underbrace{\frac{\lambda_{Kj}^2}{(\frac{1}{n}\nu_j + \lambda_{Kj}^2)^3}\langle \varphi, \phi_j\rangle^2(\hat{\nu}_j - \nu_j)^2}_{A} + \underbrace{o_p\left((\hat{\nu}_j - \nu_j)^2\right)}_{B}$$

First let us investigate the remainder term, B. Note that v_j is a function of φ and $\hat{v}_j$ is a function of $\hat{\varphi}_\alpha$. So, one can write:

$$\begin{aligned}
\frac{1}{n^2}\sum_{j=1}^{\infty} o_p\left((\hat{v}_j - v_j)^2\right) &= \frac{1}{n^2}\sum_{j=1}^{\infty} o_p(\langle\hat{\varphi}_\alpha - \varphi, \phi_j\rangle^2) \\
&= \frac{1}{n^2} o_p\left(\sum_{j=1}^{\infty}\langle\hat{\varphi}_\alpha - \varphi, \phi_j\rangle^2\right) \\
&= \frac{1}{n^2} o_p\left(\|\hat{\varphi}_\alpha - \varphi\|\right) \\
&= \frac{1}{n^2} o_p(1)
\end{aligned}$$

The final rate is obtained because of the following: Note that we do not use L or A during the first step estimation. Then the MISE of the first step estimator is given by:

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

If we assume that φ belongs to a regularity space characterized by parameter γ such that:

$$\sum_{j=1}^{\infty}\frac{\langle\varphi, \phi_j\rangle^2}{\lambda_{Kj}^{2\gamma}} < \infty.$$

Hence, for α and n fixed, we get:

$$\|\hat{\varphi}_\alpha - \varphi\|^2 = O_p\left(\frac{1}{n\alpha} + \alpha^\gamma\right),$$

given the above rate, one can always find an α such that:

$$\frac{1}{n^2}\left(\frac{1}{n\alpha} + \alpha^\gamma\right) < \frac{1}{n^\beta},$$

which will satisfy the final rate. Intuitively, it means that we need to select a small α for the first step estimation. In fact, if α is chosen optimally, the MISE would have a rate of $n^{-\gamma/\gamma+1}$ and one would need to verify that $\gamma/\gamma+1 > \beta$. As this is not possible, we choose α smaller than the optimal.

Let us now continue with term A. If we replace v_j and $\hat{v}_j$ to what they are equal

to in the second order term, we obtain:

$$\frac{1}{n^2}\sum_{j=1}^{\infty}\frac{\lambda_{Kj}^2\langle\varphi,\phi_j\rangle^2\langle\Sigma\psi_j,\psi_j\rangle^2}{\left(\frac{1}{n}\frac{\langle\Sigma\psi_j,\psi_j\rangle}{\langle\varphi,\phi_j\rangle^2}+\lambda_{Kj}^2\right)^3}\langle\varphi,\phi_j\rangle^2\left(\frac{1}{\langle\hat{\varphi}_\alpha,\phi_j\rangle^2}-\frac{1}{\langle\varphi,\phi_j\rangle^2}\right)^2$$

Another Taylor expansion of $\frac{1}{\langle\hat{\varphi}_\alpha,\phi_j\rangle^2}$ around φ gives:

$$\frac{1}{\langle\hat{\varphi}_\alpha,\phi_j\rangle^2}-\frac{1}{\langle\varphi,\phi_j\rangle^2}=-\frac{1}{\langle\varphi,\phi_j\rangle^3}\langle\hat{\varphi}_\alpha-\varphi,\phi_j\rangle+o(\hat{\varphi}_\alpha-\varphi)$$

Replacing the above equation back in term A and after some manipulations:

$$=\frac{1}{n^2}\sum_{j=1}^{\infty}\frac{\lambda_{Kj}^2\langle\Sigma\psi_j,\psi_j\rangle^2}{\left(\frac{1}{n}\frac{\langle\Sigma\psi_j,\psi_j\rangle}{\langle\varphi,\phi_j\rangle^2}+\lambda_{Kj}^2\right)^3}\frac{1}{\langle\varphi,\phi_j\rangle^4}\langle\hat{\varphi}_\alpha-\varphi,\phi_j\rangle^2$$

which has the expectation equal to:

$$=\underbrace{\frac{1}{n^3}\sum_{j=1}^{\infty}\frac{\langle\Sigma\psi_j,\psi_j\rangle^3\lambda_{Kj}^4}{\left(\frac{1}{n}\frac{\langle\Sigma\psi_j,\psi_j\rangle}{\langle\varphi,\phi_j\rangle^2}+\lambda_{Kj}^2\right)^3\langle\varphi,\phi_j\rangle^4(\alpha+\lambda_{Kj}^2)^2}}_{I}+\underbrace{\frac{1}{n^2}\sum_{j=1}^{\infty}\frac{\alpha^2\lambda_{Kj}^2\langle\Sigma\psi_j,\psi_j\rangle^2}{(\frac{1}{n}\frac{\langle\Sigma\psi_j,\psi_j\rangle}{\langle\varphi,\phi_j\rangle^2}+\lambda_{Kj}^2)^3\langle\varphi,\phi_j\rangle^4(\alpha+\lambda_{Kj}^2)^2}}_{II} \tag{A.4}$$

as

$$E[\langle\hat{\varphi}_\alpha-\varphi,\phi_j\rangle^2]=\frac{1}{n}\frac{\lambda_{Kj}^2\langle\Sigma\psi_j,\psi_j\rangle}{(\alpha+\lambda_{Kj}^2)^2}+\frac{\alpha^2\langle\varphi,\phi_j\rangle^2}{(\alpha+\lambda_{Kj}^2)^2}$$

Below, we investigate the term in A.4. Let us start with I. After some manipulation, I can be written as:

$$I=\frac{1}{n^3}\sum_{j=1}^{\infty}\frac{\lambda_{Kj}^4\langle\varphi,\phi_j\rangle^2}{\left(\frac{1}{n}+\lambda_{Kj}^2\frac{\langle\varphi,\phi_j\rangle^2}{\langle\Sigma\psi_j,\psi_j\rangle}\right)^3(\alpha+\lambda_{Kj}^2)^2}$$

Denote $x=\lambda_{Kj}^2\frac{\langle\varphi,\phi_j\rangle^2}{\langle\Sigma\psi_j,\psi_j\rangle}$ and divide and multiple the above equation by $\left(\lambda_{Kj}^2\frac{\langle\varphi,\phi_j\rangle^2}{\langle\Sigma\psi_j,\psi_j\rangle}\right)^\beta$. Then I can be written as:

$$I=\frac{1}{n^3}\sum_{j=1}^{\infty}(\frac{\lambda_{Kj}^2}{\alpha+\lambda_{Kj}^2})^2\frac{\langle\varphi,\phi_j\rangle^{2(1-\beta)}\langle\Sigma\psi_j,\psi_j\rangle^\beta}{\lambda_{Kj}^{2\beta}}\frac{x^\beta}{(\frac{1}{n}+x)^3}$$

The first term on the RHS is < 1 and the second term is finite by Assumption 2

and then I is $O(n^{-\beta})$. The order of II can be shown in similar way. After some manipulation, II can be rewritten:

$$II = \frac{1}{n^2}\sum_{j=1}^{\infty} \frac{\alpha^2 \lambda_{Kj}^2 \langle \varphi, \phi_j \rangle^4}{\left(\frac{1}{n} + \lambda_{Kj}^2 \frac{\langle \varphi, \phi_j \rangle^2}{\langle \Sigma\psi_j, \psi_j \rangle}\right)^3 \langle \Sigma\psi_j, \psi_j \rangle (\alpha + \lambda_{Kj}^2)^2}$$

If we divide and multiply II by $\left(\lambda_{Kj}^2 \frac{\langle \varphi, \phi_j \rangle^2}{\langle \Sigma\psi_j, \psi_j \rangle}\right)^{\beta}$:

$$II = \frac{1}{n^2}\sum_{j=1}^{\infty} \frac{\alpha^2}{(\alpha + \lambda_{Kj}^2)^2} \frac{\langle \varphi, \phi_j \rangle^{2(1-\beta)} \langle \Sigma\psi_j, \psi_j \rangle^{\beta}}{\lambda_{Kj}^{2\beta}} \frac{x^{1+\beta}}{(\frac{1}{n} + x)^3}$$

By making the similar arguments as in I, it can be shown that II is $O(n^{-\beta})$. Finally, $MISE(\varphi_f) - MISE(\varphi_{if}) = O_p(n^{-\beta})$ follows from Markov inequality. ■

B 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 4.1, 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, Florens, and Renault (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:

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

where

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

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

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

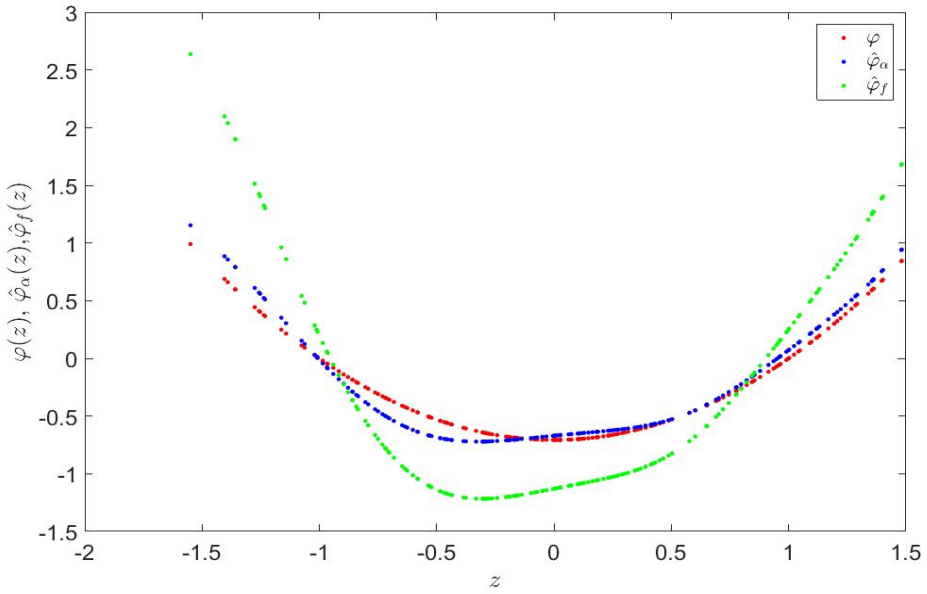
$$\mathcal{K}_x(i,j) = \frac{\mathcal{K}_x\left(\frac{x_i-x_j}{h_x}\right)}{\sum_j \mathcal{K}_x\left(\frac{x_i-x_j}{h_x}\right)}.$$

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 21 to 23. 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 B.1 and Figures B.1 to B.3. Optimal feasible estimator performs better when α is selected to minimize noise 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 B.1 is that the optimal feasible estimator looks less sensitive to different values of α. Figure B.4 below shows how MISE of the unweighted estimator and optimal feasible estimator change with different values of α. As can be seen from Figure B.4, 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 B.1: **IV simulation results - K unknown**

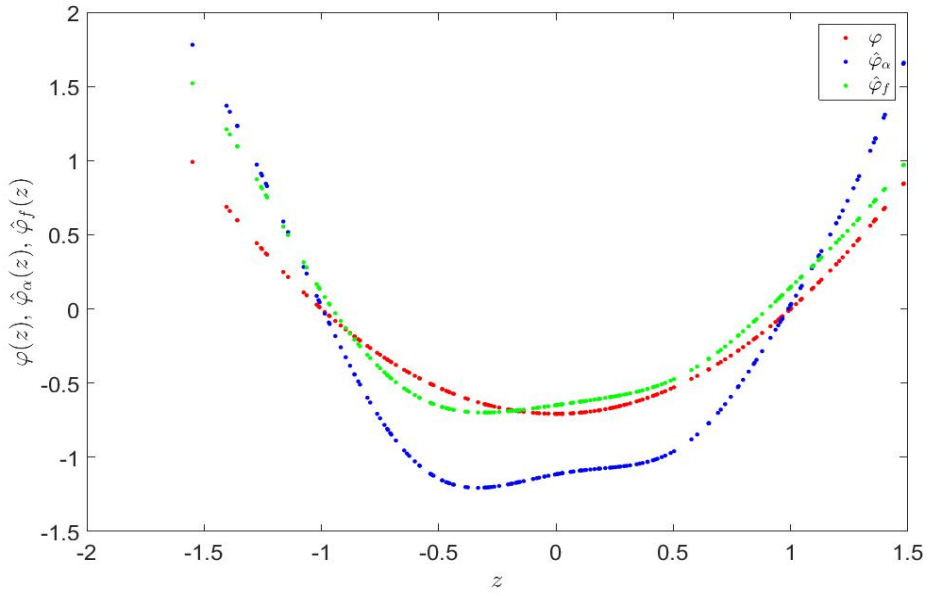
	MISE		
	$\hat{\varphi}_\alpha$	$\hat{\varphi}_f$	α_{opt}
α_{opt}^{1s} first stage	0.2949	0.5332	0.0544
α_{opt}^{2s} second stage	0.5954	0.3929	0.0287

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



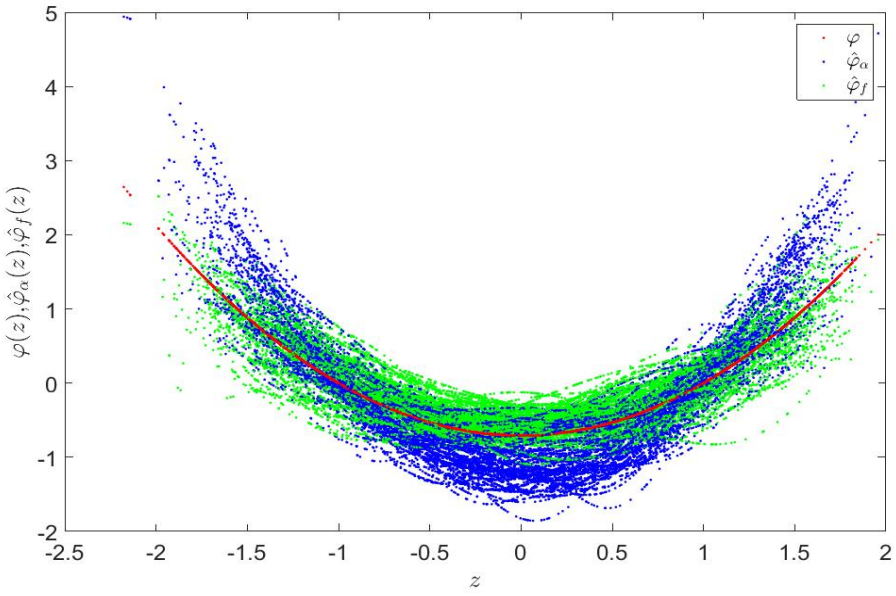
Note: α is selected in order to minimize the MISE of the first step estimator.

Figure B.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 B.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.

Figure B.4: α **vs. MISE**