Electronic Journal of Statistics
Vol. 18 (2024) 775–802
ISSN: 1935-7524
https://doi.org/10.1214/23-EJS2197

Optimal weighting for linear inverse problems*

Jean-Pierre Florens[†1] **and Senay Sokullu**[‡2]

[1] *1 Esplanade de L'Universite, 31080 Toulouse Cedex 06, France*
e-mail: jean-pierre.florens@tse-fr.eu

[2] *University of Bristol, Priory Road Complex, Priory Road, Bristol BS8 1TU, UK*
e-mail: senay.sokullu@bristol.ac.uk

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.

MSC2020 subject classifications: Primary 45Q05, 62Q05; secondary 62P20.
Keywords and phrases: Nonparametric IV regression, inverse problems, Tikhonov Regularization.

Received March 2022.

1. Introduction

In this paper we consider linear inverse problems 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 linear 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. 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

*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.

[†] Toulouse School of Economics.

[‡] University of Bristol.

are given. We derive an optimal estimator which minimizes the mean integrated square error (MISE) for linear inverse problems given in Equation (1). The model given in (1) covers a general class of linear ill-posed problems, such as X-ray tomography, denoising and deblurring in imaging, functional linear instrumental variable (IV) models and non-parametric instrumental variable models (NPIV). For instance, in denoising the operator K is identity, in the case of deconvolution, the operator K is known and the variance Σ can be estimated consistently whereas in the case of functional linear IV and NPIV, both K and Σ can be estimated consistently. In this paper, we will focus on linear inverse problems with known K and Σ.

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 the model given in (1), 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 related to the models with moment restrictions where the parameter of interest is finite-dimensional. There is a large literature on the optimality of the estimator in this class of models. For instance, Hansen (1982) shows that the optimal Generalized Method of Moments (GMM) estimator is obtained by setting the weighting matrix to the inverse of variance covariance matrix of the moment conditions and Chamberlain (1987) shows that one can reach the efficiency of Generalized Method of Moments estimator in the nonparametric models with conditional mean restrictions by using instruments given by the power series of the exogenous variable. Consider the linear GMM problem corresponding to the following model:

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

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

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

In this paper, following Hansen (1982), we obtain an optimal weighting which minimizes the 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.

Solving an ill-posed inverse problem requires regularization, see for instance Carrasco, Florens, and Renault (2007). 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 MISE for a general class of linear inverse problems.

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 MISE of a weighted linear inverse problem with given K and Σ, 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, φ and the rate of decay of eigenvalues of the variance of the noise, Σ. Given our result, we propose optimal infeasible and feasible estimators and show that they are both consistent under a general set-up. We show that the MISE of the optimal infeasible estimator no longer depends on the regularization parameter and the MISE of the feasible estimator has the same order as that of the optimal infeasible estimator for a fixed value of regularization parameter. This latter result implies that the use of the optimal weighting operator leads to a feasible estimator without the need for the choice of regularization parameter. In other words, it provides data driven way of regularization.[2]

This paper is mostly related to the literature on linear inverse problems. Cavalier (2008) reviews statistical inverse problems and explains some theoretical issues such as convergence rates, regularization, adaptation and oracle inequalities. Carrasco et al. (2007) on the other hand, studies inverse problems in structural econometrics. In contrast to most statistical inverse problems where K and Σ are given, NPIV models in econometrics require the estimation of these

[2] We thank the anonymous referee for pointing this out.

operators from the data, see Darolles, Fan, Florens, and Renault (2011); Newey and Powell (2003); Ai and Chen (2003); Horowitz (2011). Rate optimality of the NPIV estimator has been well studied (Hall and Horowitz, 2005; Chen and Reiss, 2011; Chen and Christensen, 2018), however, 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 NPIV estimation. The main contribution of Gagliardini and Scaillet (2012) lies in their computation of an explicit asymptotic mean integrated square error (MISE) for a Sobolev regularized estimator. However, their study does not delve into the optimality of their estimator concerning the choice of the weighting matrix. In contrast, our paper investigates the optimality of our estimator by considering different choices of the weighting matrix. While our paper focuses on linear inverse problems with given K and Σ, we conduct Monte Carlo simulations using a NPIV model. This allows us to compare the performance of our proposed estimator under various setups, including situations where both operators K and Σ are known and where they are unknown. By examining different scenarios, we gain insights into performance of our estimator in diverse settings.

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. However, we limit our approach to L^2-estimation under Sobolev restrictions of the function of interest where the impossibility of adaptive honest confidence sets is analysed, see Giné and Nickl (2021); Chen, Christensen, and Kankanala (2021); Babii (2020). Moreover, we focus on the case where the variance of the error is not identity as in the Gaussian white noise models. More recently in statistics, the problem of optimal regularization has been studied together with machine learning techniques, where the optimal regularization functional is obtained using a training set, see Lunz, Öktem, and Schönlieb (2018); Alberti, De Vito, Lassas, Ratti, and Santacesaria (2021).

The paper proceeds as follows. In *Section 2* we introduce our set-up. 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. In *Section 4*, we present simulation results which demonstrate our theoretical findings as well as the small sample properties of the 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. This general setting can be shown to fit with cases such as computed tomography (Alberti et al., 2021), magnetic resonance imaging (Alberti et al., 2021), deconvolution problems (Carrasco et al., 2007), functional linear regression (Hall and Horowitz, 2007), functional instrumental variables regression (Florens and Van Bellegem, 2015) or nonparametric instrumental variable regression (NPIV) (Carrasco et al., 2007).[3]

Consider the linear inverse problem given in (1). 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 $A = \Sigma^{-\frac{1}{2}}$. In the case of 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:

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

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

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

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 oscillating component of the estimated function. They further show with Monte Carlo simulations that Tikhonov regularization with

[3] In this paper we obtain our results for linear inverse problems where K and Σ are known. For functional IV and NPIV, these operators need to be estimated. We leave the treatment of estimated K and Σ for future work.

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 $j = 1, 2, \dots, \infty$ 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:

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

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

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

where $E(\epsilon_j) = 0$, $Var(\epsilon_j) = 1$. The representation given in Equation (6) is standard in the literature of inverse problems. 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:*

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

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}. Moreover, if we include $\alpha > 0$ in the definition of the operator L, one can also say that weighting by A or penalizing by αLA^{-1} are equivalent which means that the choice of A will also regularize the problem. We indeed show in Theorems 1 and 2 that the optimally weighted infeasible and feasible estimators are consistent for a fixed α. In the following sections, we only consider weighting by A and L^2 penalty with a regularization parameter α, but our results may be reinterpreted in terms of Hilbert scale penalization. One advantage of this is that the use of Hilbert scale penalty without weighting would mean that the derivatives of the function of interest is being penalised which would be optimal for smooth functions. However, we do not place any restrictions on A and we show in Proposition 1 that the optimal weighting operator A might be a differential or integral operator depending on the smoothness of φ which will then provide a data driven selection of optimal norm for the penalty.

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. 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 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.

In what follows, we derive our main results under the assumption of known K and Σ and we provide empirical evidence via Monte Carlo simulations that our results might hold when K and Σ are estimated. Note that most of the literature on inverse problems is limited to the case of known K and Σ, see Cavalier, Golubev, Picard, and Tsybakov (2002). For example, in image treatment models such as tomography, denoising, deblurring (Alberti et al., 2021), 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).

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:*

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

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

where Q is the operator:

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

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 if $\langle\varphi,\phi_j\rangle=0$, $\langle r,\psi_j\rangle=0$. 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 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 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.[4]

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, Johannes, and Van Bellegem (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 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 MISE for n fixed, not its asymptotic MISE.

3.1. The 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 a feasible estimator. A natural idea is to construct a two-step estimator.

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

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 feasible estimator and Theorem 3 shows that the MISE of the feasible estimator has the same order as 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 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 feasible estimator but it also eliminates the problem of selection of the optimal regularization parameter. In fact, selection of the optimal regularization parameter is replaced by a data-driven way of regularization via the use of optimal weighting operator, A.[5]

Assumption 3. *Source Condition: There exists $\gamma > 0$ such that:*

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

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 $\hat{\varphi}_\alpha$. If $\frac{1}{n^2\alpha} + \frac{\alpha^\gamma}{n} < n^{1-\beta}$ and Assumptions 1, 2 and 3 hold, then:*

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

Four points related to Theorem 3 are worth discussing. First, it should be noted that as the feasible estimator $(\hat{\varphi}_f)$ requires the first step estimator $(\hat{\varphi}_\alpha)$ to be plugged in, so 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 split the sample 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 feasible estimator and that of the oracle estimator have the same order asymptotically. Third, Giné and Nickl (2021) introduces *self similar functions*, whose smoothness can be estimated and this imposes some restrictions on the function of interest. Even though we indirectly estimate the smoothness of φ via the estimation of its Fourier coefficients, $\langle \varphi, \phi_j \rangle$, we do not explicitly restrict our analysis to self-similar functions. We do however impose restrictions on the rate of decay of $\langle \varphi, \phi_j \rangle$ relative to that of λ_{Kj} and $\langle \Sigma\psi_j, \psi_j \rangle$ in Assumptions 2 and 3. Finally, regarding econometric application NPIV, minimax rates have been obtained in papers such as Hall and Horowitz (2007); Chen and Reiss (2011); Chen and Christensen (2018); Chen et al. (2021). The rate in Theorem 3 is not a minimax rate as we do not define any class of estimators nor we consider a specific model. It is an oracle equality which shows that $MISE(\hat{\varphi}_f)$ is equal to $MISE(\hat{\varphi}_{if})$ plus a term which has exactly the same rate as $MISE(\hat{\varphi}_{if})$ for a general class of linear inverse problems.

4. Monte Carlo simulations

In this section we present Monte Carlo simulations to show the performance of the proposed feasible estimator compared to that of the unweighted estimator. We generate data from a NPIV model as we are interested in econometric applications of inverse-problems. Even though NPIV does not correspond to our

[5] In Theorem 3 below where we show the convergence of the MISE of the feasible estimator, we impose a bound on α given by $\frac{1}{n^2\alpha} + \frac{\alpha^\gamma}{n} < n^{1-\beta}$. Although it may look restrictive as it depends on the unknown constants γ and β, it also depends on the sample size, n, and when n increases its effect dominates and increases the bound on α.

model as K and Σ need to be estimated, NPIV design allows us to see performance of our proposed estimator in the cases of both known and unknown K and Σ. Below we first write NPIV model under our setting and then describe our simulation design.

4.1. Nonparametric IV regression

NPIV regression has been well studied in many papers; see Carrasco et al. (2007); Darolles et al. (2011); Hall and Horowitz (2005) among others and to the best of our knowledge, none of these papers has considered the optimality of the infinite dimensional parameter in terms of minimum MISE.[6] Consider a vector of random elements (Y, Z, X) such that:

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

The model then generates a linear inverse problem:

$$\begin{aligned} \mathbb{E}(\mathbb{E}(Y|X)|Z) &= \mathbb{E}(\mathbb{E}(\varphi(Z)|X)|Z), & (14)\\ r &= K\varphi, & (15)\end{aligned}$$

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 for the purposes of this paper, we assume that K is given. In this case, one can write:

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

We assume that $\mathbb{E}(U) = 0$.[7] 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 variance of V conditional on X, $\mathbb{V}(V|X) = \sigma^2$, see Carrasco et al. (2007).

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:

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

[6] As already mentioned, in this paper we consider optimality in terms of minimal MISE for fixed n, not rate-optimality.

[7] In the NPIV model described above, it is important to recognise that $E(U)$ is not strictly equal to zero and $Var(U)$ is not strictly equal to $\frac{\sigma^2}{n}K$ due to the presence of the nonparametric estimate $\hat{r}$. However, the additional terms in these moments can be controlled by appropriately selecting the bandwidth in the estimation of $\hat{r}$ and can be shown to be negligible, see Darolles et al. (2011).

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

$$\hat{\varphi}_{f,IV} = \sum_{j=1}^{\infty} \frac{\langle \hat{r}, \psi_j \rangle^2}{\frac{\sigma^2}{n}(\alpha + \lambda_{Kj}^2)^2 + \lambda_{Kj}^3 \langle \hat{r}, \psi_j \rangle^2} \lambda_{Kj}^2 \langle \hat{r}, \psi_j \rangle \phi_j. \tag{18}$$

4.2. Simulation design

We first generate data from the NPIV model introduced in previous section with given K and estimate the unknown function using feasible, $\hat{\varphi}_f$ and unweighted estimators, $\hat{\varphi}_\alpha$. Second, we simulate the design in Newey and Powell (2003) where K is known to be from a normal family but with an unknown variance. We call this case *partially known* K. Third, we use the design in the first set of simulations but this time we assume that K and Σ are unknown and we estimate them. The results of the Monte Carlo experiments show that the feasible estimator performs better than the unweighted estimator and this is also true for the case of unknown K and Σ.

Known *K*: We generate the data as the following: X, Z and V are drawn from a multivariate normal distribution with mean $(0 \quad 0 \quad 0)'$ and variance:

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

where we fix ρ to be equal to 0.6. 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, which is set to be equal to 0.5 in this design. Hence we estimate ρ from the data. Then the eigenvalues are given by $\lambda_{Kj} = \hat{\rho}^{2j}$.

Unknown *K*: The data is generated exactly the same way as in the first case, however, we assume that K is unknown.

In the simulations with known K and partially known K, 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. In the case of unknown K, 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:

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

where

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

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 the set-up above, for all three designs, 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 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:

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

We replicate this exercise 250 times for sample sizes equal to 100, 200 and 400 and estimate unweighted estimator, $\hat{\varphi}_\alpha$ and the feasible estimator, $\hat{\varphi}_f$. We truncate the sum at $j = 15$.[8] As for the regularization parameter, α, we select it in two different ways: 1) Given a grid of values of α, we select the one which minimizes the MISE of the unweighted (first stage) estimator; 2) we select the one which minimizes the MISE of the feasible (second stage) estimator.

Results: Table 1 shows the Root Mean Integrated Square Error (RMISE) of $\hat{\varphi}_\alpha$ and $\hat{\varphi}_f$ under the two different selection rules for α, for known, partially known and unknown K. In the cases of known and unknown K, the feasible estimator performs better than the unweighted estimator when α is selected in the second stage i.e. such that it minimizes the MISE of $\hat{\varphi}_f$. In the case of partially known K, the feasible estimator always performs better. As expected, in all cases RMISE of both $\hat{\varphi}_\alpha$ and $\hat{\varphi}_f$ decreases when the sample size increases.

TABLE 1
Simulation results.

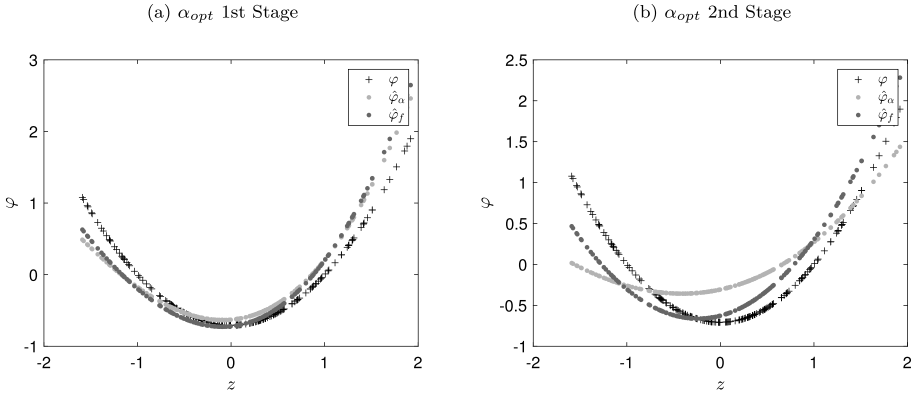
	α_{opt} first stage			α_{opt} second stage		
	RMISE			RMISE		
	$\hat{\varphi}_\alpha$	$\hat{\varphi}_f$	α_{opt}	$\hat{\varphi}_\alpha$	$\hat{\varphi}_f$	α_{opt}
K **known**						
$n = 100$	0.4511	0.6275	0.0110	0.7013	0.4327	0.0237
$n = 200$	0.3936	0.4917	0.0060	0.5219	0.3335	0.0211
$n = 400$	0.3115	0.4137	0.0048	0.5193	0.2269	0.0256
K **partially known**						
$n = 100$	0.7372	0.5687	0.0768	0.9059	0.5446	0.0719
$n = 200$	0.6255	0.4892	0.0745	0.7071	0.4663	0.0801
$n = 400$	0.8171	0.4548	0.0684	0.9302	0.4390	0.0866
K **unknown**						
$n = 100$	0.6361	0.8050	0.1062	1.1051	0.6118	0.0077
$n = 200$	0.6240	0.7785	0.1062	1.0637	0.5588	0.0111
$n = 400$	0.6029	0.7583	0.0536	1.0057	0.5353	0.0044

Two things are worth discussion in detail. First, when selected in the first stage, optimal α decreases with the sample size as one might expect. However, this is not the case for optimal α selected so as to minimize the MISE of the second stage estimator, $\hat{\varphi}_f$. Although this might seem counterintuitive, it is actually in line with our theoretical result. In Theorem 3, we show that the MISE of the feasible estimator has the same order as that of the oracle estimator for a fixed α which satisfies the condition $\frac{1}{n\alpha^2} + \frac{\alpha^\gamma}{n} < n^{1-\beta}$. Hence, when the sample size n increases, the bound on α relaxes allowing $\hat{\varphi}_f$ to have smaller RMISE for larger values of α. Second, Theorem 3 also implies that α needs to be sufficiently small depending on the values of n, γ and β. In the case of partially known K, our results show that the MISE of $\hat{\varphi}_f$ does not depend much

[8] We chose the truncation point based on values of λ_{Kj}. More precisely, we truncate at the point where λ_{Kj}'s get very close to zero.

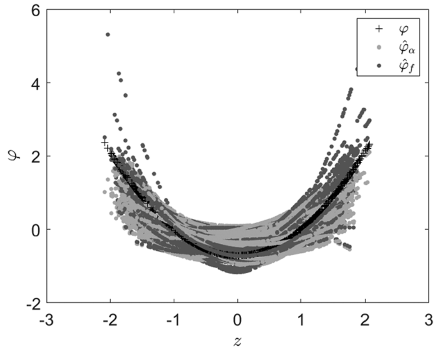
on α and this is potentially because given the function of interest, we use a grid which contains the values of α that are sufficiently small. This is not true for the design we used for the cases of known and unknown K as our results suggests that the MISE of $\hat{\varphi}_f$ changes with α. This could be fixed by using a grid which contains smaller values of α. In Figure 7, we plot the RMISE of $\hat{\varphi}_\alpha$ (RMISE1) and $\hat{\varphi}_f$ (RMISE2) against the values of α. In Panel (a), the grid of α is given by $\left[10^{-6}, 10^{-3}\right]$ and it can be seen that as the RMISE of $\hat{\varphi}_f$ does not change much. Whereas in Panels (b) and (c), the grid is larger, $\left[10^{-3}, 10^{-1}\right]$, and the RMISE of $\hat{\varphi}_f$ varies more compared to Panel (a).

Figures 1 to 6 show $\hat{\varphi}_\alpha$, $\hat{\varphi}_f$ and the true curve in all three cases. While Figues 1, 3 and 5 show the results from a single draw under for two different selection rules of the regularization parameter, Figures 2, 4 and 6 show estimated curves from 250 draws where α is chosen to minimize the MISE of $\hat{\varphi}_f$. It can also be seen in Figures 1, 3 and 5 that $\hat{\varphi}_f$ is much less dependent on α compared to $\hat{\varphi}_\alpha$.


(a) α_{opt} 1st Stage

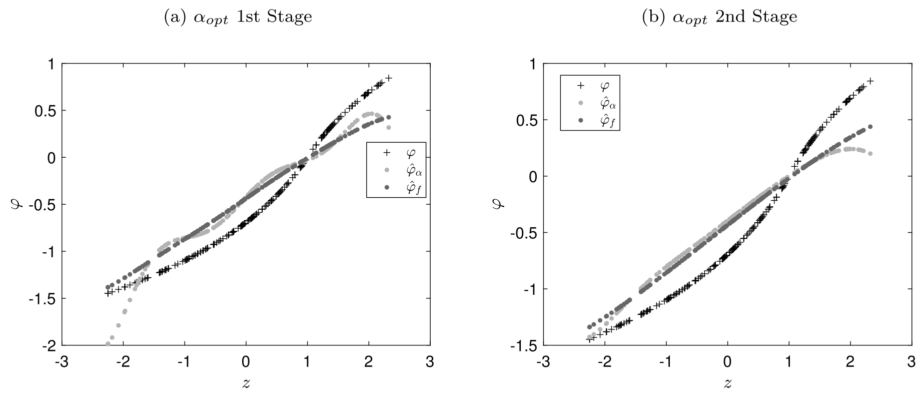
(b) α_{opt} 2nd Stage

FIG 1. *Simulation result with one draw - K known.*



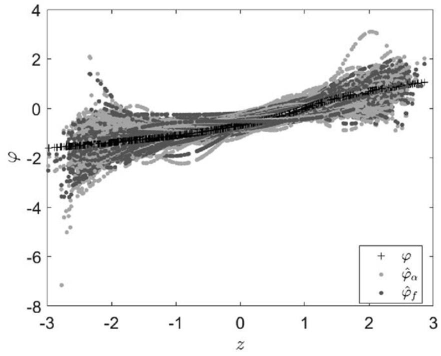
Note: α is selected in order to minimize the MISE of the second step estimator. Black pluses are the true values of the φ function. Dark gray dots are the estimated curve using the feasible estimator at each draw while the light gray dots are unweighted estimates.

FIG 2. *Simulation result with 250 draws.*


(a) α_{opt} 1st Stage

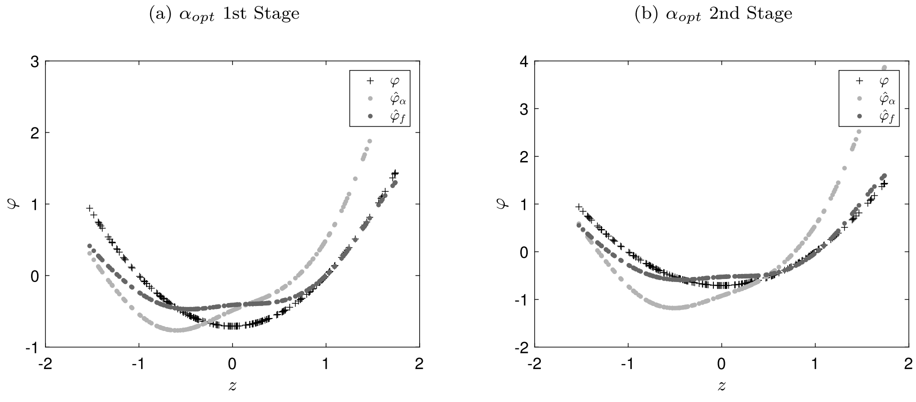
(b) α_{opt} 2nd Stage

FIG 3. *Simulation result with one draw - K partially known.*



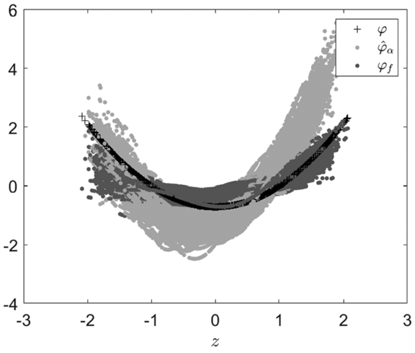
Note: α is selected in order to minimize the MISE of the second step estimator. Black pluses are the true values of the φ function. Dark gray dots are the estimated curve using the feasible estimator at each draw while the light gray dots are unweighted estimates.

FIG 4. *Simulation result with 250 draws.*


(a) α_{opt} 1st Stage

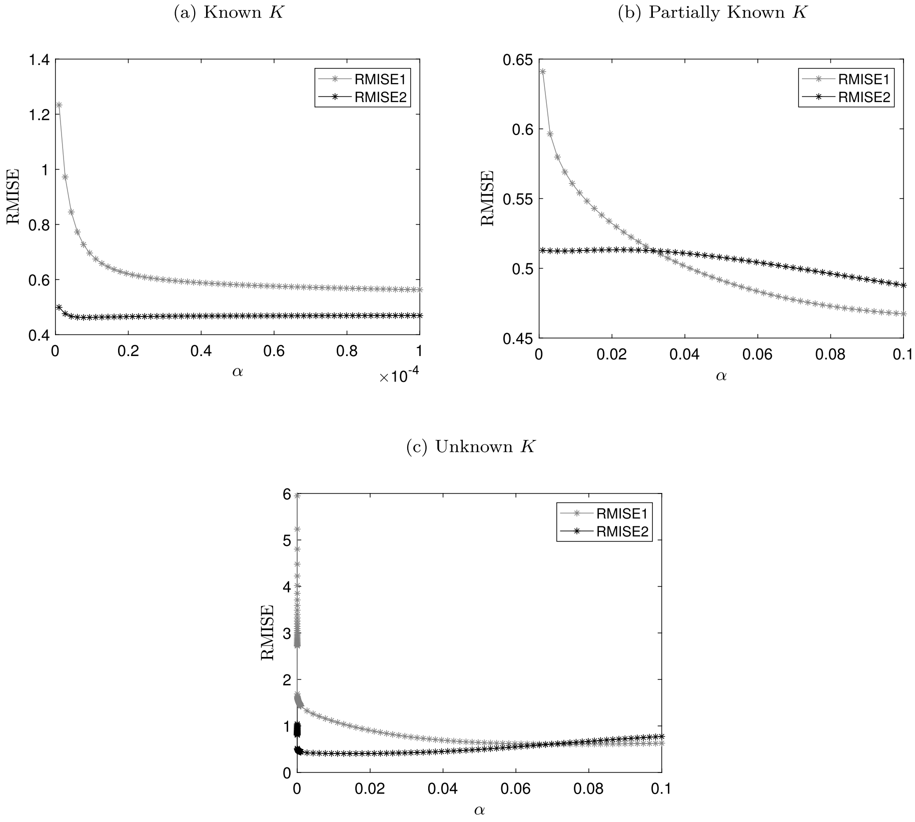
(b) α_{opt} 2nd Stage

FIG 5. *Simulation result with one draw - K unknown.*



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 feasible estimator at each draw while the light gray dots are unweighted estimates.

FIG 6. *Simulation result with 250 draws.*


(a) Known K

(b) Partially Known K

(c) Unknown K

FIG 7. *RMISE vs. α.*

5. Conclusion

In this paper we study the 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 a feasible estimator. Fourth, we study the asymptotic properties of our proposed feasible estimator and show that for a fixed value of regularization parameter, α, (i) it is consistent; (ii) its MISE has the same order as that of the oracle estimator. 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. Extension of our results to the case with estimated K and Σ is the first thing in our agenda. We believe that with this extension, 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. Second, optimal weighting in inverse problems can be studied in Bayesian context. Third, machine learning techniques have been used in the solution of inverse problems recently and extension of our results to these settings can also be studied.

Appendix 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:

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

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

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

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:

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

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

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

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 $O_p(1/n)$:

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

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:

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

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

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

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:

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

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:

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

Under Assumption 3, 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:

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

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:

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

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. □

References

Ai, C. and X. Chen (2003, November). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica 71*(6), 1795–1843. MR2015420

Alberti, G. S., E. De Vito, M. Lassas, L. Ratti, and M. Santacesaria (2021). Learning the optimal Tikhonov regularizer for inverse problems. *Advances in Neural Information Processing Systems 34*, 25205–25216.

Babii, A. (2020). Honest confidence sets in nonparametric IV regression and other ill-posed models. *Econometric Theory 36*(4), 658–706. MR4125475

Benatia, D., M. Carrasco, and J.-P. Florens (2017). Functional linear regression with functional response. *Journal of Econometrics 201*(2), 269–291. MR3717564

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. MR2421941

Cavalier, L., G. Golubev, D. Picard, and A. Tsybakov (2002). Oracle inequalities for inverse problems. *The Annals of Statistics 30*(3), 843–874. MR1922543

Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics 34*(3), 305–334. MR0888070

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

Chen, X., T. Christensen, and S. Kankanala (2021). Adaptive estimation and uniform confidence bands for nonparametric IV. *arXiv preprint arXiv:2107.11869*.

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. MR3789729

Chen, X. and M. Reiss (2011). On rate optimality for ill-posed inverse problems in econometrics. *Econometric Theory 27*(3), 497–521. MR2806258

Darolles, S., Y. Fan, J.-P. Florens, and E. Renault (2011). Nonparametric instrumental regression. *Econometrica 79*(5), 1541–1565. MR2883763

Engl, H. W., M. Hanke, and A. Neubauer (1996). *Regularization of Inverse Problems*, Volume 375. Springer Science & Business Media. MR1408680

Florens, J.-P., J. Johannes, and S. Van Bellegem (2012, 06). Instrumental regression in partially linear models. *Econometrics Journal 15*(2), 304–324. MR2951059

Florens, J.-P. and S. Van Bellegem (2015, 06). Instrumental variable estimation in functional linear models. *Journal of Econometrics 186*(2), 465–476. MR3343797

Gagliardini, P. and O. Scaillet (2012). Tikhonov regularization for nonparametric instrumental variable estimators. *Journal of Econometrics 167*(1), 61–75. MR2885439

Giné, E. and R. Nickl (2021). *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge University Press. MR3588285

Hall, P. and J. L. Horowitz (2005). Nonparametric methods for inference in the presence of instrumental variables. *Annals of Statistics 32*, 2904–2929. MR2253107

Hall, P. and J. L. Horowitz (2007). Methodology and convergence rates for functional linear regression. *The Annals of Statistics 35*(1), 70–91. MR2332269

Hansen, L. P. (1982, July). Large sample properties of generalized method of moments estimators. *Econometrica 50*(4), 1029–54. MR0666123

Horowitz, J. L. (2011, March). Applied nonparametric instrumental variables estimation. *Econometrica 79*, 347–394. MR2809374

Horowitz, J. L. (2014). Adaptive nonparametric instrumental variables estimation: Empirical choice of the regularization parameter. *Journal of Economet-*

rics 180(2), 158–173. MR3197791

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. MR2906881

Krein, S. G. and Y. I. Petunin (1966). Scales of banach spaces. *Russian Mathematical Surveys 21*(2), 85. MR0193499

Lunz, S., O. Öktem, and C.-B. Schönlieb (2018). Adversarial regularizers in inverse problems. *Advances in Neural Information Processing Systems 31*. MR4119649

Neubauer, A. (1988). When do Sobolev spaces form a Hilbert scale? *Proceedings of the American Mathematical Society 103*(2), 557–562. MR0943084

Newey, W. K. and J. L. Powell (2003, 09). Instrumental variable estimation of nonparametric models. *Econometrica 71*(5), 1565–1578. MR2000257

Newey, W. K. and R. J. Smith (2004). Higher order properties of gmm and generalized empirical likelihood estimators. *Econometrica 72*(1), 219–255. MR2031017