TIKHONOV REGULARISATION FOR FUNCTIONAL
MINIMUM DISTANCE ESTIMATORS

P. Gagliardini* and O. Scaillet†

This version: May 2006

(first version: May 2006)

*University of Lugano.
†HEC Genève and Swiss Finance Institute.
Tikhonov Regularisation for Functional Minimum Distance Estimators

Abstract

We study the asymptotic properties of a Tikhonov regularised (TiR) estimator of a functional parameter based on a minimum distance principle for nonparametric conditional moment restrictions. The estimator is computationally tractable and even takes a closed form in the linear case. We derive its Mean Integrated Squared Error (MISE), its rate of convergence and its pointwise asymptotic normality under a regularisation parameter depending on the sample size. The optimal value of the regularisation parameter is characterised. We illustrate our theoretical findings and the small sample properties with simulation results for two numerical examples. We also discuss two data driven selection procedures of the regularisation parameter via a spectral representation and a subsampling approximation of the MISE.

Keywords and phrases: Minimum Distance, Nonparametric Estimation, Ill-posed Inverse Problems, Endogeneity, Generalized Method of Moments, Subsampling, Tikhonov Regularisation.

JEL classification: C13, C14.

1 Introduction

Minimum distance and extremum estimators have received a lot of attention in the literature to exploit conditional moment restrictions assumed to hold true on the data generating process [see e.g Newey and McFadden (1994) for a review]. In a parametric setting, leading examples are the Ordinary Least Squares estimator, which takes a closed form, and the Nonlinear Least Squares estimator, which is computed through numerical optimization. Correction for endogeneity are provided by the Instrumental Variable estimator in the linear case and by the Generalised Method of Moments estimator in the nonlinear case.

In a functional setting, regression curves are inferred by local polynomial estimators and sieve estimators. A well known example is the Rosenblatt-Parzen kernel estimator. Recently, several suggestions have been made to correct for endogeneity in the nonparametric context as well, mainly motivated by the interest for non-parametric IV estimation of structural equations. Newey and Powell (NP, 2003) consider the problem of estimating non-parametrically a regression function, which is the conditional expectation of the dependent variable given a set of instruments. They propose a consistent minimum distance estimator, which is a non-parametric analog of the Two-Stage Least Square estimator. The NP methodology extends to the case of general, nonlinear conditional moment restrictions. Ai and Chen (AC, 2003) follow a similar approach to estimate an unknown function contained in a conditional moment. Although their focus is more on the efficient estimation of the parametric component in a semi-parametric conditional moment specification, they show that the estimator of the functional component converges at a rate faster than $T^{-1/4}$ in an appropriate metric.
Darolles, Florens and Renault (DFR, 2003) and Hall and Horowitz (HH, 2005) also consider non-parametric estimation of an instrumental regression function, but focus on the linear case. Their estimation approach is based on the empirical analog of the conditional moment restriction, seen as a linear integral equation in the unknown functional parameter. HH derive the optimal rate of convergence of their estimator in quadratic mean. [Cite some other peripheric papers (Chernozukov, Vanhems)].

The main theoretical difficulty to overcome in non-parametric estimation with endogeneity is ill-posedness. Ill-posedness occurs since the mapping of the reduced form parameter (that is, the distribution of the data) into the structural parameter (the instrumental regression function) is not continuous. This may have serious consequences, in particular it can lead to inconsistency of the estimators. The problem of ill-posedness has been addressed in the literature in different ways. NP and AC propose to introduce bounds on the derivatives of the functional parameter of interest, which amounts to assume a compact parameter space. In the linear case, DFR and HH adopt a regularisation technique, which results in a kind of ridge regression approach in a functional setting.

The aim of this paper is to introduce a new minimum distance estimator for a functional parameter identified by conditional moment restrictions. To address the issue of ill-posedness, we consider penalized extremum estimators which minimize a criterion of the type $Q_T(\varphi) + \lambda_T G(\varphi)$, where $Q_T(\varphi)$ is a minimum distance criterion in the functional parameter $\varphi$, $G(\varphi)$ is a penalty function and $\lambda_T$ is a positive sequence converging to zero. The penalty function $G(\varphi)$ corresponds to the Sobolev norm of function $\varphi$, which involves the
$L^2$ norms of both $\varphi$ and its derivative $\nabla \varphi$. The basic idea behind our estimator is that the term $\lambda_T G(\varphi)$ penalizes highly oscillating components of the estimator, which are otherwise unduly enhanced by the minimum distance criterion $Q_T (\varphi)$ because of ill-posedness. The amount of regularisation is tuned by parameter $\lambda_T$. We call our estimator a Tikhonov Regularised (TiR) estimator, since the penalty term is inspired by the pioneering paper of Tikhonov (1963) on the regularisation of ill-posed inverse problems. We stress that also the regularisation approach of DFR and HH is an example of Tikhonov regularisation, but with penalty term involving the $L^2$ norm instead of the Sobolev norm of the parameter. To avoid confusion, we refer to the DFR and HH estimator as a regularised estimator with $L^2$ norm.

Our paper contributes to the literature along several directions. First, we introduce a nonparametric estimator for conditional moment restrictions, which admits the following features: (i) it applies in the general (linear and nonlinear) setting; (ii) the tuning parameter is allowed to depend on sample size and to be stochastic; (iii) it may have a faster rate of convergence than the DFR and HH estimator in the linear case; (iv) it admits a closed form in the linear case. We emphasize that point (ii) is crucial to develop estimators with data-driven selection of the tuning parameter. This point is not addressed in the setting of NP and AC, where the tuning parameter is the bound on the Sobolev norm of the estimator and is assumed fixed in all theoretical results. For the same reason, feature (iv) is not shared by NP and AC estimator [see Section 2.4 for more details on the links between the TiR estimator and the literature]. Concerning point (iii), we give in Section 4 the condition under which this property holds. In our Monte-Carlo experiments in Section 6, we have
found a superior performance of the TiR estimator compared to the regularised estimator with $L^2$ norm. ¹

Second, we study rather in depth the asymptotic properties of our estimator. In particular: (i) we prove the consistency of the TiR estimator; (ii) we derive the asymptotic expansion of the Mean Integrated Squared Error (MISE) as a function of the sample size and the (deterministic) regularisation parameter; (iii) we prove the pointwise asymptotic normality of the TiR estimator. To the best of our knowledge, results (ii) and (iii), as well as (i) for a sequence of stochastic regularisation parameters, are new for non-parametric estimators of this type. In particular, the asymptotic expansion of the MISE allows us to study the effect of the regularisation parameter on the variance term and on the bias term of the TiR estimator, to define the optimal sequence of regularisation parameters, and to derive the associated optimal rate of convergence of the TiR estimator. The methodology is easily extended to the case of regularisation with $L^2$ norm, so that these results are interesting also for the study of the properties of the DFR and HH estimators. Finally, the asymptotic expansion of the MISE suggests a procedure for the data-driven selection of the regularisation parameter, that we implement in the Monte-Carlo study.

Third, we investigate the attractiveness of the TiR estimator from an applied point of view. In the nonlinear case, the TiR estimator only requires running an unconstrained optimisation routine instead of a constrained one, and in the linear case it even takes a closed form. Such a numerical tractability is a key advantage in practice, when using heavy

¹ The advantage of the Sobolev norm compared to the $L^2$ norm for regularisation is also pointed out in a numerical example in Kress (1999), Example 16.21.
resampling techniques for example. The finite sample properties seem very appealing from our numerical experiments on two examples and two data driven selection procedures of the regularisation parameter.

The rest of the paper is organized as follows. In Section 2, we first introduce the general setting of non-parametric estimation under conditional moment restrictions and the problem of ill-posedness. We then define the TiR estimator and discuss the links with the literature. In Section 3 we prove the consistency of the TiR estimator. Section 4 is devoted to the analysis of the MISE and of the optimal rates of convergence of the TiR estimator. The case of linear moment restrictions is detailed in Section 5. In section 6 we present a Monte-Carlo study of the finite sample properties of the TiR estimator. Finally, Section 7 concludes. The proofs of all results in the paper are gathered in the Appendices. The proofs of technical Lemmas are collected in a document, which is available by the authors on request.

2 Minimum Distance estimators under Tikhonov regularisation

In this section we introduce the class of Tikhonov Regularised (TiR) estimators. In Section 2.1 we present the general setting of non-parametric Minimum Distance estimation. In Section 2.2 we highlight its main issue, namely ill-posedness. In Section 2.3 TiR estimators are defined as a regularisation method for the ill-posedness problem. Finally, links with estimators and results currently available in the literature are discussed in detail in Section 2.4.
2.1 Nonparametric Minimum Distance estimation

Let \( \{(Y_t, X_t, Z_t) : t = 1, ..., T\} \) be i.i.d. variables, and let the support of \( X_t \) be \( \mathcal{X} = [0, 1] \). Suppose that the parameter of interest is a function \( \varphi_0 \) defined on \( \mathcal{X} \), which satisfies the conditional moment restriction

\[
E_0 [g(Y, \varphi_0(X)) | Z] = 0,
\]

where \( g \) is a known function. Parameter \( \varphi_0 \) belongs to a subset \( \Theta \) of \( L^2[0,1] \), equipped with the \( L^2 \) scalar product \( \langle \varphi, \psi \rangle = \int_{\mathcal{X}} \varphi(x)\psi(x)dx \) and the \( L^2 \) norm \( \|\varphi\| = \langle \varphi, \varphi \rangle^{1/2} \). It is assumed that \( \Theta \) is bounded and closed, and that \( \varphi_0 \) is the unique function \( \varphi \in \Theta \) that satisfies the conditional moment restriction (1).

The non-parametric Minimum Distance approach to estimate \( \varphi_0 \) as in AC and NP relies on \( \varphi_0 \) minimizing the criterion

\[
Q_\infty(\varphi) = E_0 \left[ m(\varphi, Z)' \Omega_0(Z)m(\varphi, Z) \right], \quad \varphi \in \Theta,
\]

where \( m(\varphi, z) = E_0 [g(Y, \varphi(X)) | Z = z] \), and \( \Omega_0(z) \) is a p.d. matrix for any given \( z \). This criterion is well-defined if \( m(\varphi, z) \) belongs to \( L^2_{\Omega_0}(F_Z) \), for any \( \varphi \in \Theta \), where \( L^2_{\Omega_0}(F_Z) \) denotes the \( L^2 \) space of square integrable vector-valued functions of \( Z \) defined by scalar product \( \langle \psi_1, \psi_2 \rangle_{L^2_{\Omega_0}(F_Z)} = E_0 [\psi_1(Z)' \Omega_0(Z)\psi_2(Z)] \). Then, the idea is to estimate \( \varphi_0 \) by the minimizer of the empirical counterpart of Criterion (2). For instance, AC and NP estimate the conditional moment \( m(\varphi, z) \) by an orthogonal polynomials approach, and minimize the empirical criterion over a finite-dimensional Sieve approximation of \( \Theta \) based on polynomial or spline functions.
The main difficulty in non-parametric Minimum Distance estimation is that, contrary to the standard parametric case, the assumption that function $\varphi_0$ is identified in a bounded and closed parameter set $\Theta$ is not sufficient in general to get the consistency of the estimator. This is due to the so-called ill-posedness of such an estimation problem.

2.2 Unidentifiability and ill-posedness in Minimum Distance estimation

The goal of this section is to highlight the issue of ill-posedness in Minimum Distance estimation [NP; see also Kress (1999), Chapter 15, for a general treatment of ill-posed inverse problems, and Carrasco, Florens and Renault (2005) for a survey on inverse problems in econometrics]. To briefly explain what ill-posedness is, note that solving the equation $E_0 [g(Y, \varphi(X)) \mid Z] = 0$ for unknown function $\varphi \in \Theta$ can be seen as an inverse problem, which maps the conditional distribution $F_0(y,x \mid z)$ of $(Y,X)$ given $Z = z$ into the solution $\varphi_0$ [see Equation (1)]. Ill-posedness arises when this mapping is not continuous. As a consequence, the estimator $\hat{\varphi}$ of $\varphi_0$, which is the solution of the inverse problem corresponding to a consistent estimator $\hat{F}$ of $F_0$, is not guaranteed to be consistent. Indeed, by non-continuity, small deviations of $\hat{F}$ from $F_0$ may result in large deviations of $\hat{\varphi}$ from $\varphi_0$. We refer to NP for a more in-depth discussion along these lines. In this paper, we prefer to emphasize the link between ill-posedness and a classical concept in econometrics, namely parameter identification.

To illustrate the main point, let us consider the case of non-parametric linear IV estimation, where $g(y, \varphi(x)) = \varphi(x) - y$. The moment function $m(\varphi,z) = E_0 [\varphi(X) - Y \mid Z = z]$
can be written as

\[ m(\varphi, z) = (A \varphi)(z) - r(z) = (A \Delta \varphi)(z), \quad (3) \]

where \( \Delta \varphi := \varphi - \varphi_0 \), operator \( A \) is defined by \( (A \varphi)(z) = \int \varphi(x)f(w|z)dw \) and \( r(z) = \int yf(w|z)dw \). Conditional moment restriction (1) identifies \( \varphi_0 \) if and only if operator \( A \) is injective. The limit criterion in (2) becomes

\[ Q_{\infty}(\varphi) = E_0 \left[ (A\Delta \varphi)(Z)^\prime \Omega_0(Z) (A\Delta \varphi)(Z) \right] = \langle \Delta \varphi, A^*A\Delta \varphi \rangle, \quad (4) \]

where \( A^* \) denotes the adjoint operator of \( A \) w.r.t. the scalar products \( \langle ., . \rangle \) and \( \langle ., . \rangle_{L^2_{\Omega_0}(F_Z)} \).

Under weak regularity conditions, operator \( A \) is compact. Thus, \( A^*A \) is compact and self-adjoint. We denote by \( \{ \psi_j : j \in \mathbb{N} \} \) an orthonormal basis in \( L^2[0,1] \) of eigenfunctions of operator \( A^*A \), and by \( \mu_1 \geq \mu_2 \geq \cdots \), with \( \mu_j > 0 \), the corresponding eigenvalues [see Kress (1999), Section 15.3, for the spectral decomposition of compact, self-adjoint operators]. By compactness of \( A^*A \), the eigenvalues are such that \( \mu_j \to 0 \). Assume that \( \varphi_0 \) is an interior point of parameter set \( \Theta \). Then, the limit criterion \( Q_{\infty}(\varphi) \) can be minimized by a sequence in \( \Theta \) such as

\[ \varphi_n = \varphi_0 + \varepsilon \psi_n, \quad n \in \mathbb{N}, \quad (5) \]

for \( \varepsilon > 0 \), which does not converge to \( \varphi_0 \). Indeed, \( Q_{\infty}(\varphi_n) = \varepsilon^2 \langle \psi_n, A^*A\psi_n \rangle = \varepsilon^2 \mu_n \to 0 \) as \( n \to \infty \), but \( \| \varphi_n - \varphi_0 \| = \varepsilon, \forall n \). Since we can chose \( \varepsilon > 0 \) as small as we want, the usual identification assumption [e.g., White and Wooldridge (1991)]

\[ \inf_{\varphi \in \Theta : \| \varphi - \varphi_0 \| \geq \varepsilon} Q_{\infty}(\varphi) > 0 = Q_{\infty}(\varphi_0), \quad \text{for} \ \varepsilon > 0, \quad (6) \]
is not satisfied. In other words, function $\varphi_0$ is not identified in $\Theta$ as an isolated minimum of $Q_\infty$. This is the identification problem of Minimum Distance estimation with functional parameter. Failure of identification condition (6) is due to 0 being a limit point of the eigenvalues of operator $A^*A$. It applies in the general setting of conditional moment restriction (1), whenever the linearization of moment function $m(\varphi, z)$ around $\varphi = \varphi_0$ involves a compact operator. This is the maintained assumption in our paper, and is stated below.

**Assumption 1 (Ill-posedness):** The moment function $m(\varphi, z)$ is such that $m(\varphi, z) = (A\Delta \varphi)(z) + R(\varphi, z)$, for any $\varphi \in \Theta$, where

(i) the operator $A$ defined by $(A\Delta \varphi)(z) = \int \frac{\partial g}{\partial v} (y, \varphi_0(x)) f(w|z) \Delta \varphi(x) \, dw$ is a compact operator in $L^2[0,1]$;

(ii) the second-order term $R(\varphi, z)$ is such that $\sup_{\varphi \in \Theta} \|R(\varphi, .)\|_{L^2_b(F_{\varphi})}/\|A\Delta \varphi\|_{L^2_b(F_{\varphi})} < 1$.

Under Assumption 1, the identification condition (6) is not satisfied, and the Minimum Distance estimator which minimizes the empirical counterpart of criterion $Q_\infty(\varphi)$ over (a Sieve approximation of) set $\Theta$ is not consistent.

### 2.3 Tikhonov Regularised (TiR) estimators

In this paper, we address the issue of ill-posedness by introducing Minimum Distance estimators based on Tikhonov regularisation. We consider extremum estimators which minimize a criterion of the type $Q_T(\varphi) + \lambda_T G(\varphi)$, where $Q_T(\varphi)$ is an empirical counterpart of criterion $Q_\infty(\varphi)$ in (2), $G(\varphi)$ is a penalty function introduced to solve the unidentifiability problem
arising from ill-posedness, and \( \lambda_T \) is a sequence converging to zero as sample size \( T \) increases.

Functions \( Q_T(\varphi) \) and \( G(\varphi) \) are defined next.

The conditional moment \( m(\varphi, z) = E_0[g(Y, \varphi(X)) \mid Z = z] \) can be estimated non-parametrically by \( \hat{m}(\varphi, z) = \int g(y, \varphi(x)) \hat{f}(w|z) \, dw \), where \( \hat{f}(w|z) \) denotes a kernel estimator of the density of \( (Y, X) \) given \( Z = z \) with kernel \( K \), bandwidth \( h_T \), and \( w = (y, x) \). Then, the criterion \( Q_T(\varphi) \) is defined by

\[
Q_T(\varphi) = \frac{1}{T} \sum_{t=1}^{T} \hat{m}(\varphi, Z_t)' \Omega_T(Z_t) \hat{m}(\varphi, Z_t),
\]

where \( \Omega_T(z), T \in \mathbb{N}, \) is a sequence of p.d. matrices converging to \( \Omega_0(z), \) P-a.s., for any \( z \).

Different choices of penalty function \( G(\varphi) \) are possible, leading to consistent estimators under the assumptions of Theorem 1 in Section 3 below. In this paper, we focus on the Sobolev norm \( G(\varphi) = \|\varphi\|_{H^2}^2 \). More precisely, we assume that \( \varphi_0 \) belongs to some subset \( \Theta \) of the Sobolev space \( H^2[0,1] \), which is defined as the completion of linear space \( \{\varphi \in C^1[0,1] \mid \nabla \varphi \in L^2[0,1]\} \) with respect to the \( L^2 \) scalar product \( \langle . , . \rangle \). Sobolev space \( H^2[0,1] \) is an Hilbert space w.r.t. the scalar product \( \langle \varphi, \psi \rangle_H := \langle \varphi, \psi \rangle + \langle \nabla \varphi, \nabla \psi \rangle \), and the corresponding Sobolev norm is denoted by \( \|\varphi\|_H = \langle \varphi, \varphi \rangle_H^{1/2} \).

The Minimum Distance estimator under Tikhonov regularisation with Sobolev norm is defined next.

**Definition 1:** The Tikhonov Regularised (TiR) Minimum Distance estimator is defined by

\[
\hat{\varphi} = \arg \inf_{\varphi \in \Theta_T} Q_T(\varphi) + \lambda_T \|\varphi\|_{H}^2,
\]

where \( \lambda_T \) is a stochastic sequence such that \( \lambda_T \geq 0 \) and \( \lambda_T \to 0 \) P-a.s., and \( (\Theta_T) \) is an
increasing sequence of subsets of the Sobolev space $H^2[0,1]$, which are compact w.r.t. the $L^2$-norm $\| \cdot \|$.

The name Tikhonov Regularised (TiR) estimators that we use to characterize the Minimum Distance estimators introduced in Definition 1 goes back to Tikhonov (1963), in his pioneering paper on the regularisation of ill-posed inverse problems [see Kress (1999), Chapter 16]. The main intuition is that the term $\lambda_T \| \varphi \|^2_H$ in the criterion penalizes highly oscillating components of the estimated function, which would be otherwise unduly enhanced, since the criterion $Q_T (\varphi)$ becomes asymptotically flat along some directions because of ill-posedness. For instance, in the linear IV case where $Q_{\infty} (\varphi) = \langle \Delta \varphi, A^* A \Delta \varphi \rangle$, these directions correspond to the eigenfunctions $\psi_n$ of operator $A^* A$ to eigenvalues $\mu_n$ close to zero, that is for large $n$ [see Equation (5) and the discussion in Section 2.2]. Typically, $\psi_n$ is an highly oscillating function and $\| \psi_n \|_H \to \infty$ as $n \to \infty$, so that these directions are penalized by term $G (\varphi) = \| \varphi \|^2_H$ in the empirical criterion $Q_T (\varphi) + \lambda_T \| \varphi \|^2_H$. In Theorem 1 in Section 3 below, we provide precise conditions under which the penalty function $G (\varphi) = \| \varphi \|^2_H$ restores the validity of the identification condition (6) and ensures the consistency of the TiR estimator.

The sequence $(\lambda_T)$ in Definition 1 controls for the amount of regularisation introduced by term $G (\varphi) = \| \varphi \|^2_H$, and how this depends on sample size $T$. Therefore, $\lambda_T$ can be seen as a tuning parameter (or as a sequence of tuning parameters). The rate of convergence of $\lambda_T$ to zero affects the rate of convergence of the TiR estimator $\hat{\varphi}$, We will discuss in Section 4 below the choice of the sequence $(\lambda_T)$ to achieve an optimal rate of convergence of TiR estimator $\hat{\varphi}_T$, and we will present two global data driven selection procedures for $\lambda_T$ in
2.4 Links with the literature

The goal of this Section is to discuss the links between the TiR estimator and the different approaches proposed in the literature on nonparametric estimation under conditional moment restrictions.

2.4.1 Regularisation by compactness

To address the issue of ill-posedness, NP and AC [see also Blundell, Chen and Kristensen (2005)] suggest considering a compact parameter set $\Theta$. In this case, by the same argument as in the standard parametric setting, the assumption that $\varphi_0$ is the unique function in $\Theta$ which satisfies (1) implies identification condition (6). Compact sets in $L^2[0,1]$ can be defined by imposing a bound on the Sobolev norm $\|\varphi\|_H \leq B$ of the functional parameter. Then, the estimator is obtained by minimization problem (7), where $\lambda_T$ is interpreted as a Kuhn-Tucker multiplier.

Our approach by TiR estimators differ from AC and NP along two directions. On the one hand, for TiR estimators $\lambda_T$ is a free regularisation parameter, whereas $\lambda_T$ is tight down by the slackness condition in NP and AC approach: either $\lambda_T = 0$ or $\|\varphi\|_H = \overline{B}$, P-a.s. As a consequence, the approach by TiR estimators presents three important advantages.

i) Optimal rates of convergence. Although, for given sample size $T$, selecting different $\lambda_T$ amounts to select different $\overline{B}$ when the constraint is binding, the asymptotic properties of the TiR estimator and of the estimators with fixed $\overline{B}$ are different. In partic-
ular, the adoption of a bound $B$ on the Sobolev norm independent of sample size $T$ implies in general the selection of a sub-optimal sequence of regularisation parameters $\lambda_T$. Thus, the NP and AC estimators share rates of convergence which are slower than that of the TiR estimator with optimally selected sequence of regularisation parameter. The optimal rates of convergence for the TiR estimator are characterized in Section 4. Finally, note that letting $B = B_T$ grow (slowly) with sample size $T$ is not equivalent to our approach and does not guarantee the consistency of the estimator. Indeed, when $B_T \to \infty$, the resulting limit parameter set $\Theta$ is not compact.

ii) Data-driven selection of tuning parameters. For the TiR estimator, the tuning parameter $\lambda_T$ is allowed to depend on sample size $T$ and sample data, whereas in the theoretical setting of NP and AC the tuning parameter $B$ is treated as fixed. Thus, our approach allows for a discussion of asymptotic properties of regularised estimators with data-driven selection of the tuning parameter [see Proposition 3 in Section 3 for consistency].

iii) Computational tractability. Finally, we emphasize that the TiR estimator features computational advantages compared to NP and AC estimators. This is because, for given $\lambda_T$, the TiR estimator is defined by an unconstrained optimization problem, whereas inequality constraint $\|\varphi\|_H \leq B$ has to be accounted for in the minimization defining estimators with given $B$. In particular, in the case of linear conditional moment restrictions, TiR estimators admit a closed form [see Section 5], whereas the computation of the NP and AC estimator requires a numerical constrained quadratic optimization routine.

On the other hand, a second difference is that NP, AC and BCK use finite-dimensional
2.4.2 Regularisation with $L^2$ norm

For the special case of non-parametric IV estimation of a single equation model [see Equation (3)], DFR and HH [see also Carrasco, Florens and Renault (2005)] introduce a regularised estimator defined by minimization problem (7) with Sobolev norm $\|\varphi\|_H$ replaced by $L^2$ norm $\|\varphi\|$ in the penalty term, and $\Omega_0(Z) = 1$. Indeed, it is possible to show that the first order condition for such an estimator corresponds to the linear equation (4.1) in DFR, or to the estimator defined at p. 4 in HH. DFR and HH study the consistency and the optimal rates of convergence of their estimator for a deterministic sequence of regularisation parameters $\lambda_T$. In Section 4, we will compare the optimal rate of convergence of regularised estimators with Sobolev norm and with $L^2$ norm, and give conditions under which the first one is larger. Finally, note that the techniques used in this paper to study the asymptotic properties of TiR estimators are easily extended to estimators with $L^2$ regularisation and allow to derive new results for the DFR and HH estimators, such as the asymptotic expansion of the Mean Integrated Square Error (MISE) in Section 4.

3 Consistency of TiR estimators

In this section we show the consistency of the TiR estimator. To highlight the main idea, we first provide in Section 3.1 a consistency theorem for penalized extremum estimators minimizing the criterion $Q_T(\varphi) + \lambda_T G(\varphi)$ with a general penalty function $G(\varphi)$. Then, in Section 3.2 the assumptions of the theorem are particularized to the Sobolev penalty
function $G(\varphi) = \|\varphi\|_H^2$ used for the TiR estimator.

3.1 A general consistency result for penalized extremum estimators

Let us consider an extremum estimator of the TiR-type as in Definition 1 with a general penalty function $G(\varphi)$

$$\hat{\varphi} = \arg \inf_{\varphi \in \Theta_T} Q_T(\varphi) + \lambda_T G(\varphi),$$

(8)

where $Q_T(\varphi)$, $(\lambda_T)$ and $\Theta_T$ are as in Definition 1. This estimator is well-defined and measurable under weak conditions [see Appendix 1].

The consistency of estimator $\hat{\varphi}$ defined in (8) is stated in the next Theorem.

**Theorem 1:** Let

(i) $\delta_T := \sup_{\varphi \in \Theta_T} |Q_T(\varphi) - Q_\infty(\varphi)| \overset{p}{\to} 0$;

(ii) $\varphi_0 \in \Theta$, and $\bigcup_{T=1}^\infty \Theta_T$ is dense in $\Theta \subset H^2[0,1]$;

(iii) For any $\varepsilon > 0$, $C_\varepsilon(\lambda) := \inf_{\varphi \in \Theta: \|\varphi - \varphi_0\| \geq \varepsilon} Q_\infty(\varphi) + \lambda G(\varphi) - Q_\infty(\varphi_0) - \lambda G(\varphi_0) > 0$, for any $\lambda > 0$ small enough;

(iv) $\exists a > 0$ such that $\lim_{\lambda \to 0} \lambda^{-a} C_\varepsilon(\lambda) > 0$, $T^a \delta_T \overset{p}{\to} 0$, and $T^a \rho_T \to 0$, for any $\varepsilon > 0$,

where $\rho_T := \inf_{\varphi \in \Theta_T: \|\varphi - \varphi_0\| \leq \varepsilon} Q_\infty(\varphi) + |G(\varphi) - G(\varphi_0)|$.

Then, under (i)-(iv), for any sequence $(\lambda_T)$ such that $\lambda_T > 0$, $\lambda_T \to 0$, P-a.s., and

$$\lambda_T/T \to 0, \quad P\text{-a.s.},$$

(9)
The estimator \( \hat{\varphi} \) defined in (8) is consistent, namely \( \| \hat{\varphi}_T - \varphi_0 \| \xrightarrow{p} 0 \).

**Proof:** See Appendix 1.

If \( G = 0 \), Theorem 1 corresponds to the standard result of consistency for extremum estimators [e.g., Wooldridge and White (1991), Corollary 2.6]. Indeed, in this case, Condition (iii) is the usual identification condition (6), whereas Condition (iv) is satisfied. Theorem 1 extends this consistency result to situations where Condition (6) does not hold, as it is the case for our ill-posed setting (see Section 2.2). The identification of \( \varphi_0 \) as isolated minimum is restored by including a small additional component \( \lambda G(\varphi) \) in the limit criterion. Thus, Condition (iii) in Theorem 1 is the condition on penalty function \( G(\varphi) \) to overcome ill-posedness and achieve consistency of the estimator \( \hat{\varphi} \). To interpret Condition (iv), note that in the ill-posed setting we have \( C_\varepsilon(\lambda) \to 0 \) as \( \lambda \to 0 \), and the rate of this convergence can be seen as a measure for the severity of ill-posedness. Thus, Condition (iv) introduces a bound on ill-posedness severity, related to the rates of uniform convergence \( \delta_T \xrightarrow{p} 0 \) and approximation error \( \rho_T \to 0 \) of the Sieve \( \Theta_T \). In Appendix 1, we provide technical regularity conditions to quantify this bound and to verify Conditions (i), (ii), and (iv) of Theorem 1 for the TiR estimator. Finally, it is important to emphasize that Theorem 1 is more general than the results currently known in the literature, since sequence \( (\lambda_T) \) is allowed to be stochastic, possibly data dependent, in a fully general way. Condition (9) on \( \lambda_T \) for consistency requires that \( \lambda_T \) converges to zero at a rate smaller than \( 1/T \).

The rest of this Section will focus on the key assumption of Theorem 1, that is identifi-
cational assumption (iii). The next Proposition provides a sufficient condition for the validity of this assumption.

**Proposition 2:** Assume that the function $G$ is bounded from below. Furthermore, suppose that, for any $\varepsilon > 0$ and any sequence $(\varphi_n)$ in $\Theta$ such that $\|\varphi_n - \varphi_0\| \geq \varepsilon$ for all $n \in \mathbb{N}$, we have

$$Q_\infty(\varphi_n) \to Q_\infty(\varphi_0) \text{ as } n \to \infty \implies G(\varphi_n) \to \infty \text{ as } n \to 0. \quad (10)$$

Then, Condition (iii) of Theorem 1 is satisfied.

**Proof:** See Appendix 1.

Condition (10) provides a simple intuition to explain why the penalty function $G(\varphi)$ restores identification. Indeed, it basically requires that the sequences $(\varphi_n)$ in $\Theta$, which minimize $Q_\infty(\varphi)$ without converging to $\varphi_0$, are penalized by function $G(\varphi)$. In the next section, we particularize this condition for the penalty function which is relevant for the TiR estimator in Definition 1, that is the Sobolev norm $G(\varphi) = \|\varphi\|^2_{H}$. 

### 3.2 Penalization with Sobolev norm

When the penalty function $G(\varphi) = \|\varphi\|^2_{H}$ is used, Condition (10) in Proposition 2 can be nicely stated in terms of the spectrum of the operator $A^*A$, where $A$ is the operator in the linearization of the moment function defined in Assumption 1.

**Assumption 2:** Let $\{\psi_j : j \in \mathbb{N}\}$ be an orthonormal basis in $L^2[0,1]$ of eigenfunctions of operator $A^*A$ to eigenvalues $\mu_j$, ordered such that $\mu_1 \geq \mu_2 \geq \cdots$, and let function
\( \psi_j \in H^2[0,1], \text{ for any } j \in \mathbb{N}. \) Then, \( M_n := \inf_{\varphi \in S_n : \|\varphi\| = 1} \|\varphi\|_H \to \infty \text{ as } n \to \infty, \) where \( S_n = \text{span} \{ \psi_j : j \geq n \}. \)

Assumption 2 basically requires that the subspace spanned by the eigenfunctions of \( A^*A \) to eigenvalues close to zero consists of highly oscillating functions with large Sobolev norm. Then, deviations of the estimator \( \hat{\varphi} \) from \( \varphi_0 \) along such directions are penalized by \( G(\varphi) = \|\varphi\|^2_H. \) This compensates the inability of the empirical criterion \( Q_T(\varphi) \) to achieve this task because of its becoming asymptotically flat in such directions.

In Lemma A.1 in Appendix 1, we show that Assumptions 1 and 2 imply Condition (10) in Proposition 2. Then, from Theorem 1 and Proposition 2, the consistency of the TiR estimator follows.

4 Mean Integrated Square Error analysis of the TiR estimator

4.1 The Mean Integrated Square Error

In this section, we derive the Mean Integrated Square Error (MISE) of the TiR estimator with deterministic sequence of regularisation parameters. To simplify the exposition, we assume that an optimal weighting matrix is used.

**Assumption 3:** The asymptotic weighting matrix \( \Omega_0(z) \) is \( V_0 [g(Y_t, \varphi_0(X_t)) | Z = z]^{-1}. \)

The asymptotic expansion of the MISE is characterized in the next Proposition.
Proposition 3: Under Assumptions 1-3, in Appendix B, and the bandwidth conditions

\[ h_m^T = o(\lambda_T b(\lambda_T)), \quad (T\lambda_T)^{-1} = o\left(h_{TZ}^T\right), \]  

the MISE of the TiR estimator \( \hat{\varphi} \) with deterministic sequence \( (\lambda_T) \) is given by

\[ E \left[ \| \hat{\varphi} - \varphi_0 \|^2 \right] = \frac{1}{T} \sum_{j=1}^{\infty} \frac{\nu_j}{(\lambda_T + \nu_j)^2} \| \phi_j \|^2 \| b(\lambda_T) \|^2 =: M_T(\lambda) \]  

up to terms which are asymptotically negligible w.r.t. the RHS, where \( \{ \phi_j : j \in \mathbb{N} \} \) are the orthonormal eigenfunctions of operator \( A^*A \) to eigenvalues \( \nu_j \), \( A^* \) denotes the adjoint operator of \( A \) w.r.t. the scalar products \( \langle ., . \rangle_H \) and \( \langle ., . \rangle_{L^2_{\text{loc}}(F_Z)} \), function \( b(\lambda_T) \) is given by

\[ b(\lambda_T) = \| (\lambda_T + A^*A)^{-1} A^*A\varphi_0 - \varphi_0 \|, \]  

\( m \) is the order of the kernel \( K \), and \( d_Z \) the dimension of \( Z \).

Proof: See Appendix 2.

The asymptotic expansion of the MISE consists of two components, which are a variance term and a bias term, respectively.

(i) The bias function \( b(\lambda_T) \) is the \( L^2 \) norm of \( (\lambda_T + A^*A)^{-1} A^*A\varphi_0 - \varphi_0 =: \varphi^* - \varphi_0 \). To interpret function \( \varphi^* \), note that the quadratic approximation of the limit criterion [see (4) and Assumption 1] can be written as

\[ \langle \Delta \varphi, A^*A\Delta \varphi \rangle = E_0 \left[ (A\Delta \varphi)(Z) \right] \Omega_0(Z) (A\Delta \varphi)(Z) = \langle \Delta \varphi, A^*A\Delta \varphi \rangle_H, \quad \varphi \in \Theta. \]

Then, function \( \varphi^* \) minimizes the penalized asymptotic criterion \( \langle \Delta \varphi, A^*A\Delta \varphi \rangle_H + \lambda_T \| \varphi \|_H^2 \).

Thus, \( b(\lambda_T) \) is the asymptotic bias arising from introducing penalty \( \lambda_T \| \varphi \|_H^2 \) in the criterion.
It corresponds to the so-called regularisation bias in the theory of Tikhonov regularisation [see e.g. Kress (1999), Groetsch (1984)]. Under general conditions on operator $A^*A$ and true function $\varphi_0$, the bias function $b(\lambda)$ is increasing w.r.t. $\lambda$ and such that $b(\lambda) \to 0$ as $\lambda \to 0$.

(ii) The variance term $T^{-1} \sum_{j=1}^{\infty} \|\phi_j\|^2 \left[ \nu_j / (\lambda_T + \nu_j)^2 \right]$ involves a weighted sum of the "regularised" inverse eigenvalues $\nu_j / (\lambda_T + \nu_j)^2$ of operator $A^*A$, with weights $\|\phi_j\|^2$. To have an interpretation, note that the inverse of operator $A^*A$ corresponds to the standard asymptotic variance matrix $(J_0V_0^{-1}J_0)^{-1}$ of the efficient GMM in the parametric setting, where $J_0 = E_0 \left[ \partial g / \partial \theta \right]$ and $V_0 = V_0 [g]$. In the ill-posed non-parametric setting, the inverse of operator $A^*A$ is unbounded, and its eigenvalues $1/\nu_j \to \infty$ diverge. The penalty term $\lambda_T \|\varphi\|^2_H$ in the criterion defining the TiR estimator implies that inverse eigenvalues $1/\nu_j$ are replaced by $\nu_j / (\lambda_T + \nu_j)^2$.

The variance term $T^{-1} \sum_{j=1}^{\infty} \|\phi_j\|^2 \left[ \nu_j / (\lambda_T + \nu_j)^2 \right]$ is a decreasing function of $\lambda_T$. To study its behaviour when $\lambda_T \to 0$, we introduce the next assumption.

**Assumption 4:** The eigenfunctions $\phi_j$ and the eigenvalues $\nu_j$ of $A^*A$ satisfy

$$\sum_{j=1}^{\infty} \nu_j^{-1} \|\phi_j\|^2 = \infty.$$

Under Assumption 4, the series $k_T := \sum_{j=1}^{\infty} \|\phi_j\|^2 \left[ \nu_j / (\lambda_T + \nu_j)^2 \right]$ diverges as $\lambda_T \to 0$. When $k_T \to \infty$ such that $k_T/T \to 0$, the variance term converges to zero. However, the rate of convergence is smaller than the parametric rate $1/T$. This smaller rate of convergence is typical in nonparametric estimation. Note, however, that the smaller rate of convergence

---

2 Since $\nu_j / (\lambda_T + \nu_j)^2 \leq \nu_j$, the infinite sum converges under Assumption B.6 (i) in Appendix B.
is not coming from localization as for kernel estimation, but from the ill-posedness of the problem, which implies \( \nu_j \to 0 \).

The asymptotic expansion of the MISE of the TiR estimator given in Proposition 3 does not involve the bandwidth \( h_T \), as long as Conditions (11) are satisfied. The variance term is asymptotically independent of \( h_T \) since the asymptotic expansion of \( \hat{\varphi} - \varphi_0 \) involves the kernel density estimator integrated w.r.t. \((Y, X, Z)\) [see Equation (36) in Appendix 2, first term, and the proof of Lemma A.3]. The integral averages the localization effect of the bandwidth \( h_T \). On the contrary, kernel estimation \( \hat{m}(\varphi, z) \) of the conditional moment function does have an effect on the bias of the TiR estimator. However, the assumption \( h_T^2 = o (\lambda_T b (\lambda_T)) \) in (11) implies that the estimation bias is asymptotically negligible compared to the regularisation bias [see Lemma A.4 in Appendix 2].

Finally, it is also possible to derive a similar asymptotic expansion of the MISE for the estimator \( \hat{\varphi}_T \) regularised by the \( L^2 \) norm. This characterisation is new

\[
E \left[ \| \hat{\varphi}_T - \varphi_0 \|^2 \right] = \frac{1}{T} \sum_{j=1}^{\infty} \frac{\mu_j}{(\lambda_T + \mu_j)^2} + \tilde{b} (\lambda_T)^2,
\]

where \( \mu_j \) are the eigenvalues of operator \( A^* A \), and \( \tilde{b} (\lambda_T) = \| (\lambda_T + A^* A)^{-1} A^* A \varphi_0 - \varphi_0 \| \).

Let us now come back to the MISE \( M_T (\lambda) \) of the TiR estimator in Proposition 3 and discuss the optimal choice of the regularisation parameter \( \lambda_T \). Since the bias term is increasing in the regularisation parameter, whereas the variance term is decreasing, we face a kind of bias-variance trade-off. The optimal sequence of deterministic regularisation parameters is given by \( \lambda_T^* = \arg \min_{\lambda > 0} M_T (\lambda) \), and the corresponding optimal MISE of the TiR estimator is given by \( M_T^* := M_T (\lambda_T^*) \).
The optimal sequence of regularisation parameters $\lambda^*_T$, in particular its rate of convergence to zero, depends on the decay behaviour of the eigenvalues $\nu_j$ and of the norms of eigenfunctions $\|\phi_j\|$, as well as on the bias function $b(\lambda)$ close to $\lambda = 0$. In the next section, we characterize the optimal sequence of regularisation parameters $\lambda^*_T$, the corresponding optimal MISE $M^*_T$, and their rate of convergence in a broad class of models.

4.2 Optimal rates of convergence

The eigenvalues $\nu_j$ and the norms of eigenfunctions $\|\phi_j\|$ can feature different types of decay as $j \to \infty$, for instance geometric or hyperbolic decay. Intuitively, the first type is associated with a faster convergence of the spectrum to zero, and thus to a more serious problem of ill-posedness. In this section, we focus our analysis on the case where the eigenvalues $\nu_j$ feature geometric decay and the norms of eigenfunctions $\|\phi_j\|$ feature hyperbolic decay. Results for the other cases are summarised at the end of the section.

Assumption 5: The eigenvalues $\nu_j$ and the norms of the eigenfunctions $\|\phi_j\|$ of operator $A^*A$ are such that, for $j = 1, 2, \ldots$, and some positive constants $C_1, C_2$,

(i) $\nu_j = C_1 \exp(-\alpha j), \alpha > 0$,  
(ii) $\|\phi_j\|^2 = C_2 j^{-\beta}, \beta > 0$.  

Assumption 5 (i) is satisfied for a large number of models, including for instance the two examples that we consider below in our Monte-Carlo analysis. In general, it is known that, under appropriate regularity conditions, compact integral operators with smooth kernel feature eigenvalues with decay of (at least) exponential type [see Theorem 15.20 in Kress]
Assumption 5 (ii) is adopted e.g. in Wahba (1977), and is also satisfied in the examples of our Monte-Carlo analysis.

We further assume that the bias function features a power-law behaviour close to $\lambda = 0$.

**Assumption 6:** The bias function is such that $b(\lambda) = C_3 \lambda^\delta$, $\delta > 0$, for $\lambda$ close to 0, where $C_3$ is a positive constant.

Then, the MISE and the optimal sequence of regularisation parameters are characterised in the next Proposition.

**Proposition 4:** Under the Assumptions of Proposition 3, Assumptions 5 and 6, for some positive constants $c_1, c_2, c$ and $\bar{c}$, we have

(i) The MISE is $M_T(\lambda) = \frac{1}{T} c_1 \frac{1}{\lambda [\log (1/\lambda)]^\delta} + c_2 \lambda^{2\delta}$, up to terms which are negligible when $\lambda \to 0$ and $T \to \infty$.

(ii) The optimal sequence of regularisation parameters is

$$
\log \lambda_T^* = \log c - \frac{1}{1 + 2\delta} \log T, \quad T \in \mathbb{N},
$$

up to a term which is negligible w.r.t. the RHS.

(iii) The optimal MISE is $M_T^* = \bar{c} T^{-\frac{2\delta}{1+2\delta}} (\log T)^{-\frac{2\delta+\beta}{1+2\delta}}$, up to a term which is negligible w.r.t. the RHS.

---

3 In the case of linear IV estimation and regularisation with $L^2$ norm, the eigenvalues correspond to the nonlinear canonical correlations of $(X, Z)$. When $X$ and $Z$ are monotonic transformations of variables which are jointly normally distributed with correlation parameter $\rho$, the canonical correlations of $(X, Z)$ are $\rho^j$, $j \in \mathbb{N}$ [see e.g. DFR]. Thus the eigenvalues feature exponential decay.
Proof: See Appendix 3.

The log of the optimal regularisation parameter is linear in the log sample size. The slope coefficient \( \gamma := 1/(1 + 2\delta) \) is smaller than 1, and depends on the convexity parameter \( \delta \) of the bias function close to \( \lambda = 0 \). We have \( \gamma < 1/2 \) when the squared bias function \( b(\lambda)^2 \) is convex, that is \( 2\delta > 1 \), respectively \( \gamma \geq 1/2 \) when \( 2\delta < 1 \). The optimal MISE converges to zero as a power of \( T \) and of \( \log T \). The negative exponent of the dominant term \( T \) is \( 2\delta/(1 + 2\delta) \). This rate of convergence is smaller than 1, that is the parametric rate, because of ill-posedness, and is increasing w.r.t. convexity parameter \( \delta \) of the bias function. Note that the geometric decay rate \( \alpha \) does not affect neither the rate of convergence of the optimal regularisation sequence, nor that of the MISE, whereas coefficient \( \beta \) of eigenfunction norms affects the exponent of the log \( T \) term in the MISE only. Finally, under Assumptions 5 and 6, the bandwidth conditions (11) are fulfilled for the optimal sequence of regularisation parameters (15) if \( h_T = C \cdot T^{-\eta} \), with \( \frac{1}{dZ} \frac{2\delta}{1 + 2\delta} < \eta < \frac{1}{m} \frac{1 + \delta}{1 + 2\delta} \). This condition can be satisfied if \( m dZ > \frac{1 + \delta}{2\delta} \).

To conclude this section, we briefly discuss the optimal rate of convergence of the MISE when the eigenvalues feature hyperbolic decay, that is \( \nu_j = C j^{-\alpha}, \alpha > 0 \), or when regularisation with \( L^2 \) norm is adopted. The results are summarized in Table 1 below, and are found using Formula (14) and an argument similar to the proof of Proposition 4. In Table 1, parameter \( \beta \) is defined as in Assumption 5 (ii) for the TiR estimator. Parameters \( \alpha \) and \( \tilde{\alpha} \) denote the hyperbolic decay rates of the eigenvalues of operator \( A^*A \) for the TiR estimator, and of operator \( A^*A \) for \( L^2 \) regularisation, respectively. We assume \( \alpha, \tilde{\alpha} > 1 \), and \( \alpha > \beta - 1 \).
to satisfy Assumption 4. Finally, parameters $\delta$ and $\tilde{\delta}$ are the power-law coefficients of the bias function $b(\lambda)$ and $\tilde{b}(\lambda)$ for $\lambda \to 0$ as in Assumption 6, where $b(\lambda)$ is defined in (13) for the TiR estimator, and $\tilde{b}(\lambda)$ in (14) for $L^2$ regularisation, respectively.

<table>
<thead>
<tr>
<th>Spectrum</th>
<th>TiR estimator</th>
<th>$L^2$ regularisation</th>
</tr>
</thead>
<tbody>
<tr>
<td>geometric spectrum</td>
<td>$T^{-\frac{2\delta}{1+2\delta}} (\log T)^{-\frac{2\delta}{1+2\delta}}$</td>
<td>$T^{-\frac{2\delta}{1+2\delta}}$</td>
</tr>
<tr>
<td>hyperbolic spectrum</td>
<td>$T^{-\frac{2\delta}{1+2\delta}(1-\beta)/\alpha}$</td>
<td>$T^{-\frac{2\delta}{1+2\delta+1/\alpha}}$</td>
</tr>
</tbody>
</table>

Table 1: Optimal rate of convergence of the MISE. The decay factors are $\alpha$ and $\tilde{\alpha}$ for the eigenvalues, $\delta$ and $\tilde{\delta}$ for the bias, and $\beta$ for the squared norm of the eigenfunctions.

With hyperbolic spectrum, the rate of convergence (power of $T$) of the TiR estimator features an additional term $(1 - \beta)/\alpha$ in the denominator, which involves both the $\alpha$ and $\beta$ coefficients. When $\beta > 1$, the rate of convergence is faster than that with geometric spectrum. This is an effect of the less severe ill-posedness problem. The rate of convergence with geometric spectrum is recovered letting $\alpha \to \infty$ (up to the $\log T$ term).

The rate of convergence with $L^2$ regularisation coincides with that of the TiR estimator with $\beta = 0$ and coefficients $\alpha, \delta$ corresponding to operator $A^*A$ instead of $A^+A$. With geometric spectrum, the TiR estimator features a faster rate of convergence than the regularised estimator with $L^2$ norm if $\delta > \tilde{\delta}$, that is if the bias function of the TiR estimator is more
convex. Finally, note that with hyperbolic spectrum and $L^2$ regularisation, the formula given in Table 1 corresponds to that derived by HH, Theorem 4.1.4

5 The TiR estimator for linear moment restrictions

In this section we derive the TiR estimator when the moment restrictions are linear w.r.t. the functional parameter $\varphi_0$. We consider the case of non-parametric IV estimation of a single equation model, with $g(y, \varphi_0(x)) = \varphi_0(x) - y$, and conditional moment as in (3). Then, the estimated moment function is given by

$$\hat{m} (\varphi, z) = \int \varphi (x) \hat{f} (w|z) \, dw - \int y \hat{f} (w|z) \, dw =: (\hat{A} \varphi) (z) - \hat{r} (z).$$

To simplify the exposition, we assume that $\Omega_0(z) = V_0 [Y_t - \varphi_0 (X_t) \mid Z = z]^{-1} = 1$ in Assumption 3. The objective function of the TiR estimator in Definition 1 can be rewritten as [see Appendix 2.1]

$$Q_T (\varphi) + \lambda_T \| \varphi \|^2_H = \langle \varphi, \hat{A}^* \hat{A} \varphi \rangle_H - 2 \langle \varphi, \hat{A}^* \hat{r} \rangle_H + \lambda_T \langle \varphi, \varphi \rangle_H, \quad \varphi \in H^2 [0, 1],$$

up to a term independent of $\varphi$, where $\hat{A}^*$ denotes the linear operator defined on $L^2_{H^0} (F_Z)$ by

$$\langle \varphi, \hat{A}^* \psi \rangle_H = \left\{ \begin{array}{ll}
\frac{1}{T} \sum_{t=1}^T (\hat{A} \varphi) (Z_t) \psi (Z_t), & \varphi \in H^2 [0, 1], \quad \psi \in L^2_{H^0} (F_Z).
\end{array} \right\}$$

Under the regularity conditions in Appendix B, Criterion (16) admits a global minimum $\hat{\varphi}$ on $H [0, 1]$, which is characterized by the first order condition

$$\left( \lambda_T + \hat{A}^* \hat{A} \right) \hat{\varphi} = \hat{A}^* \hat{r}.$$  

---

4 To see this, note that their Assumption A.3 corresponds to $\hat{\delta} = (2\beta_{HH} - 1) / (\hat{\alpha} + \beta_{HH})$, where $\beta_{HH}$ is the $\beta$ coefficient of HH.
This is a Fredholm integral equation of Type II \(^5\). The transformation of the ill-posed problem (1) in the well-posed estimating equation (18) is induced by the penalty term involving the Sobolev norm. The TiR estimator is the unique solution of Equation (18) and is given by

\[
\hat{\varphi} = \left( \lambda_T + \hat{A}^* \hat{A} \right)^{-1} \hat{A}^* \hat{r}.
\] (19)

The TiR estimator can be approximated numerically by introducing a finite-dimensional basis of functions \( \{ P_j : j = 1, ..., K \} \) in \( H^2[0, 1] \) and solving Equation (19) on the subspace spanned by \( \{ P_j : j = 1, ..., K \} \), which yields

\[
\varphi \simeq \sum_{j=1}^{K} \theta_j P_j =: \hat{\theta} P, \quad \theta \in \mathbb{R}^K.
\] (20)

The \( K \times K \) matrix corresponding to operator \( \hat{A}^* \hat{A} \) on the subspace spanned by \( \{ P_j \} \) is given by [using (17)]

\[
\langle P_i, \hat{A}^* \hat{A} P_j \rangle_H = \frac{1}{T} \sum_{t=1}^{T} (\hat{A} P_i)(Z_t)(\hat{A} P_j)(Z_t) = \frac{1}{T} (\hat{P}^T \hat{P})_{i,j}, \quad i, j = 1, ..., K,
\]

where \( \hat{P} \) is the \( T \times K \) matrix with rows \( \hat{P}(Z_t)' = \int P(x)' \hat{f}(w|Z_t) dw, \ t = 1, ..., T \). Matrix \( \hat{P} \) is the matrix of the "fitted values" in the regression of \( P(X) \) on \( Z \) at the sample points. Then, Equation (18) reduces to a matrix equation \( \left( \lambda_T D + \frac{1}{T} \hat{P}^T \hat{P} \right) \theta = \frac{1}{T} \hat{P}^T \hat{R} \), where \( \hat{R} = (\hat{r}(Z_1)',..., \hat{r}(Z_T))' \) and \( D \) is the \( K \times K \) matrix of Sobolev scalar products \( D_{i,j} = \langle P_i, P_j \rangle_H, \quad i, j = 1, ..., K \). The solution is given by \( \hat{\theta} = \left( \lambda_T D + \frac{1}{T} \hat{P}^T \hat{P} \right)^{-1} \frac{1}{T} \hat{P}^T \hat{R} \), which yields the approximation of the TiR estimator \( \hat{\varphi} \simeq \hat{\theta} P \).

---

\(^5\) See e.g. Linton and Mammen (2005), (2006), Gagliardini and Gouriéroux (2006), and the survey by Carrasco, Florens, Renault (2005) for other examples of estimation problems leading to Type II equations.
Estimator $\hat{\theta}$ is a 2SLS estimator with a ridge correction term. It is easy to verify that, this is the estimator that we obtain, if we replace Approximation (20) in Criterion (16) and we minimize w.r.t. $\theta$. This latter approach has been considered by NP, AC, and Blundell et al (2005), which use Sieve estimators. However, it is important to emphasize that, the introduction of a series of basis functions as in (20) is simply a method to compute approximately the true TiR estimator $\hat{\varphi}$ in (19), which is a well-defined estimator on the function space. In particular, when $K = K_T \to \infty$ sufficiently fast with $T$, the asymptotic properties of estimator $\hat{\theta} P$ are the same as for estimator $\hat{\varphi}$. Moreover, the asymptotic properties (consistency) of the estimators proposed by NP, AC, and Blundell et al (2005), have been derived only in the case where parameter $\lambda_T$ is tight down by the inequality constraint $\|\hat{\varphi}\|_H \leq \bar{B}$ for fixed $\bar{B}$, whereas, for the TiR estimator, $\lambda_T$ is treated as a free regularization parameter depending on sample size.

6 A Monte Carlo study

6.1 Data generating process

Following NP we draw the errors $U$ and $V$ and the instruments $Z$ as:

$$
\begin{pmatrix}
U \\
V \\
Z
\end{pmatrix} \sim N
\begin{pmatrix}
\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \\
\begin{pmatrix} 1 & \rho & 0 \\ \rho & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\end{pmatrix},
\rho \in \{0, 0.5\},
$$

and build $X^* = Z + V$. Then we map $X^*$ into a variable $X = \Phi(X^*)$, which lives in $[0, 1]$. The function $\Phi$ denotes the cdf of a standard Gaussian variable, and is assumed to be known. To generate $Y$, we restrict ourselves to the linear case since a simulation analysis of
a nonlinear case would be very time consuming. We examine two designs

Case 1:  \( Y = B_{a,b} (X) + U, \)

where \( B_{a,b} \) denotes the cdf of a Beta\((a,b)\) variable;

Case 2:  \( Y = \sin (\pi X) + U. \)

The parameters of the beta distribution are chosen equal to \( a = 2 \) and \( b = 5. \)

When the correlation \( \rho \) between \( U \) and \( V \) is 50\% there is endogeneity in both cases. When \( \rho = 0 \) there is no need to correct for the endogeneity bias.

The moment condition is

\[
E_0 [Y - \varphi_0 (X) \mid Z] = 0,
\]

where the functional parameter is \( \varphi_0 (x) = B_{a,b} (x) \) in Case 1, and \( \varphi_0 (x) = \sin (\pi x) \) in Case 2, \( x \in [0,1] \).

### 6.2 Estimation procedure

Since we face an unknown function \( \varphi_0 \) on \([0,1]\), we use a series approximation based on standardized shifted Chebyshev polynomials of the first kind (see Section 22 on orthogonal polynomials of Abramowitz and Stegun (1970) for their mathematical properties). We take orders 0 to 5 which yields six coefficients \( (K = 6) \) to be estimated in the approximation

\[
\varphi (x) \approx \sum_{j=0}^{5} \theta_j P_j (x), \quad \text{where} \quad P_0 (x) = T_0^* (x) / \sqrt{\pi}, \quad P_j (x) = T_j^* (x) / \sqrt{\pi / 2}, \quad j \neq 0, \quad \text{and the shifted}
\]

Chebyshev polynomials of the first kind are

\[ T_0^*(x) = 1, \quad T_1^*(x) = 1 + 2x, \quad T_2^*(x) = 1 - 8x + 8x^2, \]
\[ T_3^*(x) = -1 + 18x - 48x^2 + 32x^3, \quad T_4^*(x) = 1 - 32x + 160x^2 - 256x^3 + 128x^4, \]
\[ T_5^*(x) = -1 + 50x - 400x^2 + 1120x^3 - 1280x^4 + 512x^5. \]

The (squared) Sobolev norm \( \| \varphi \|^2_{H^k} = \int_0^1 \varphi^2 + \int_0^1 (\nabla \varphi)^2 \) is approximated by

\[ \| \varphi \|^2_{H^k} \simeq \sum_{i=0}^5 \sum_{j=0}^5 \theta_i \theta_j \int_0^1 (P_i(x)P_j(x) + \nabla P_i(x)\nabla P_j(x)) \, dx. \]

The coefficients in this quadratic form \( \theta' D \theta \) take a closed form, and can be computed easily via integration with a symbolic calculus package:

\[ D = \begin{pmatrix} \frac{1}{\pi} & 0 & -\sqrt{2} & 0 & -\sqrt{2} & 0 \\ \\ : & \frac{26}{3\pi} & 0 & \frac{38}{3\pi} & 0 & \frac{166}{21\pi} \\ \\ : & \frac{218}{5\pi} & 0 & \frac{1182}{35\pi} & 0 \\ \\ : & \frac{3808}{39\pi} & 0 & \frac{5090}{63\pi} \\ \\ : & \frac{67894}{315\pi} & 0 \\ \\ \ldots & \ldots & \frac{82802}{231\pi} \end{pmatrix}. \]
The $L_2$ norm $\|\varphi\|^2$ can be approximated in a similar way with $\theta' B \theta$ where

$$
B = \begin{pmatrix}
\frac{1}{\pi} & 0 & -\sqrt{\frac{2}{3}} & 0 & -\sqrt{\frac{2}{15}} & 0 \\
\vdots & \frac{2}{3\pi} & 0 & -\frac{2}{5\pi} & 0 & -\frac{2}{21\pi} \\
& & \frac{14}{15\pi} & 0 & -\frac{38}{105\pi} & 0 \\
& & & \frac{34}{35\pi} & 0 & -\frac{22}{63\pi} \\
& & & \vdots & \frac{62}{63\pi} & 0 \\
& & & & \ldots & \ldots & \frac{98}{99\pi}
\end{pmatrix}.
$$

Such simple and exact forms ease implementation, improve on speed, and contribute to the numerical stability of the estimation procedure.

The kernel estimator $\hat{m}(\varphi, z)$ of the conditional moment is approximated through

$$
\hat{P}(z) - \hat{r}(z) = \theta' \hat{P}(z) - \hat{r}(z) = \theta' \left[ \sum_{t=1}^{T} P(X_t) K \left( \frac{Z_t - z}{h} \right) \right] - \left[ \sum_{t=1}^{T} Y_t K \left( \frac{Z_t - z}{h} \right) \right],
$$

where $h$ denotes the bandwidth, and $K$ is the Gaussian kernel. This kernel estimator is asymptotically equivalent to the one described in the lines above. We prefer it because of its numerical tractability. It has the advantage of avoiding bivariate numerical integration and the choice of two additional bandwidths. The bandwidth is selected via the standard rule of thumb $h = 1.06 \hat{\sigma}_Z T^{-1/5}$ (Silverman (1986)), where $\hat{\sigma}_Z$ is the empirical standard deviation of $Z_t$.

The weighting function $\Omega_0(z)$ is taken equal to unity, satisfying Assumption 3.

---

6 The Gauss programs developed for this section are available on request from the authors.
6.3 Simulation results

The sample size is initially fixed at $T = 400$. Estimator performance is measured in terms of the Mean Integrated Squared Error (MISE) and the Integrated Squared Bias (ISB) based on averages over 1000 repetitions. We use a univariate Gauss-Legendre quadrature with 40 knots to compute the integrals.

Figures 1 to 4 concern Case 1 while Figures 5 to 8 concern Case 2. In each figure the left panel plots the MISE on a grid of lambda, the central panel the ISB on a grid of lambda, and the right panel the mean estimated functions and the true function on the unit interval. Mean estimated functions correspond to averages obtained either from regularised estimates with a lambda achieving the lowest MISE or from OLS estimates. The regularization schemes use the Sobolev norm, corresponding to the TiR estimator (odd numbering of the figures), and the $L_2$ norm (even numbering of the figures). We consider designs exhibiting an endogeneity ($\rho = 0.5$) in Figures 1, 2, 5, 6, while Figures 3, 4, 7, 8 are dedicated to the designs without endogeneity ($\rho = 0$).

A couple of remarks can be made. First, the bias of the OLS estimator can be large under endogeneity. Second, the MISE of the TiR estimator is more convex in lambda than the one obtained from an $L_2$ norm, and performance is clearly better for the TiR estimator. The Sobolev norm should be strongly favoured over the $L_2$ norm in order to recover the shape of the true functions. Third, the fit obtained by the OLS estimator is almost perfect when endogeneity is absent. Using six polynomials delivers a very good approximation of the true functions.
We have also examined sample sizes $T = 100$ and $T = 1000$, as well as approximations based on polynomials with orders up to 10 and 15. The above conclusions remain qualitatively unaffected. This suggests that as soon as the order of the polynomials is sufficiently large to deliver a good numerical approximation of the underlying function, it is not necessary to link it with sample size, as explained in Section 5. For example Figures 9 and 10 are the analogues of Figures 1 and 5 with $T = 1000$. We can see that the bias term is almost identical, while the variance term decreases by a factor about 2.5 as predicted by Proposition 3.

In Figure 11 we display the six eigenvalues of operator $A^*A$ and the $L^2$-norms of the corresponding eigenfunctions when the same approximation basis of six polynomials is used. These true quantities have been computed by Monte Carlo integration. The eigenvalues $\nu_j$ feature a geometric decay w.r.t. the order $j$, whereas the decay of the norms $\|\phi_j\|^2$ is of an hyperbolic type. This is conform to Assumption 5 and the analysis conducted in Proposition 4. A linear fit of the plotted points gives a decay factor $\hat{\alpha} = 2.254$ for the eigenvalues and a decay factor $\hat{\beta} = 2.911$ for the norm of the eigenfunctions.

Figure 12 is dedicated to check whether the line $\log \lambda_r^* = \log c - \gamma \log T$, induced by Proposition 4 (ii), holds in small samples. For $\rho = 0.5$ the right panel for Case 1 as well as the left panel for Case 2 exhibit a linear relationship between the logarithm of the regularisation parameter minimizing the average MISE on the 1000 Monte Carlo simulations and the logarithm of sample size ranging from $T = 50$ to $T = 1000$. The OLS estimation of this linear relationship from the plotted pairs delivers $\hat{c} = .226, \hat{\gamma} = .752$ in Case 1, and $\hat{c} = .012, \hat{\gamma} = .752$ in Case 2.
\[ \hat{\gamma} = .428 \] in Case 2. Both estimated slope coefficients are smaller than 1, and qualitatively consistent with the implications of Proposition 4. Indeed, from Figures 9 and 10 the ISB curve appears to be more convex in Case 2 than in Case 1. This points to a larger \( \delta \) parameter, and thus to a smaller slope coefficient \( \gamma = 1/(1 + 2\delta) \), in Case 2. Inverting the relationship \( \gamma = 1/(1 + 2\delta) \) we get estimates for the decay factor \( \delta \), which are \( \hat{\delta} = .165 \) and \( \hat{\delta} = .668 \) in Case 1 and Case 2, respectively.

By a similar argument, Proposition 4 also explains the better performance of the TiR estimator compared to the \( L^2 \)-regularised estimator that we reported above. Indeed, comparing the ISB curves of the two estimators in Case 1 (Figures 1 and 2) and in Case 2 (Figures 5 and 6), it appears that the TiR estimator features a more convex ISB curve. This implies \( \delta > \tilde{\delta} \) and thus a faster rate of convergence of the TiR estimator.

Finally we wish to conclude by a brief discussion on data driven selection procedures of the regularisation parameter \( \lambda_T \). We investigate a first method based on the asymptotic spectral representation of the MISE provided in Proposition 3, and a second method based on a resampling approximation.

The first data driven selection procedure aims at estimating directly Expression (12) in order to derive the optimal regularisation parameter. In unreported results we have checked that the asymptotic MISE, the asymptotic ISB and the asymptotic variance are close to the ones exhibited in Figures 9 and 10. These true quantities have also been computed by Monte Carlo integration. We have found an asymptotic optimal lambda equal to .0018 in Case 1 and to .0009 in Case 2, which are of the same magnitudes as .0013 and .0007 in Figures 9
and 10. We have also checked that the linear relationship exhibited in Figure 12 holds true
when deduced from optimizing the asymptotic MISE. The OLS estimation delivers $\hat{c} = .418,$
$\hat{\gamma} = .795$ in Case 1, and $\hat{c} = .037, \hat{\gamma} = .546$ in Case 2, and thus $\hat{\delta} = .129$ and $\hat{\delta} = .418,$
respectively.

The data driven estimation algorithm goes as follows:

**Algorithm**

(i) Perform the spectral decomposition of the matrix $D^{-1}\hat{P}'\hat{P}/T$ to get eigenvalues $\hat{\nu}_j$ and
eigenvectors $\hat{w}_j,$ normalized to $\hat{w}'_jD\hat{w}_j = 1,$ $j = 1, ..., K.$

(ii) Get a first-step TiR estimator $\hat{\theta}$ using a pilot regularisation parameter $\bar{\lambda}.$

(iii) Estimate the MISE:

$$\hat{M}(\lambda) = \frac{1}{T} \sum_{j=1}^{K} \frac{\hat{\nu}_j}{(\lambda + \hat{\nu}_j)^2} \hat{w}'_j B \hat{w}_j$$

$$+ \hat{\vartheta}' \left[ \frac{1}{T} \hat{P}' \hat{P} \left( \lambda D + \frac{1}{T} \hat{P}' \hat{P} \right)^{-1} - I \right] \left[ \frac{1}{T} \hat{P}' \hat{P} \left( \lambda D + \frac{1}{T} \hat{P}' \hat{P} \right)^{-1} - I \right] \hat{\theta},$$

and minimize it w.r.t. $\lambda$ to get the optimal regularisation parameter $\hat{\lambda}.$

(iv) Compute the second-step TiR estimator $\hat{\theta}$ using regularisation parameter $\hat{\lambda}.$

A second-step estimated MISE viewed as a function of sample size $T$ and regularisation
parameter $\lambda$ can then be estimated with $\hat{\theta}$ instead of $\bar{\theta}.$ Besides, if we assume the decay
behaviour of Assumptions 5 and 6, the decay factors $\alpha$ and $\beta$ can be estimated via minus the
slopes of the linear fit on the pairs $(\log \hat{\nu}_j, j)$ and on the pairs $(\log \hat{w}'_j B \hat{w}_j, \log j), j = 1, ..., K.$
After getting lambdas minimizing the second-step estimated MISE on a grid of sample sizes we can also estimate $\gamma$ by regressing the logarithm of lambda on the logarithm of sample size.

We have used $\bar{\lambda} = \{.0005, .0001\}$ as the pilot regularisation parameter for $T = 1000$ and $\rho = .5$. In Case 1, the average (quartiles) of the selected lambda over 1000 simulations is equal to .0028 (.0014, .0020, .0033) when $\bar{\lambda} = .0005$, and .0027 (.0007, .0014, .0029) when $\bar{\lambda} = .0001$. In Case 2, the results are .0009 (.0007, .0008, .0009) when $\bar{\lambda} = .0005$, and .0008 (.0004, .0006, .0009) when $\bar{\lambda} = .0001$. The selection procedure tends to slightly overpenalize on average, especially in Case 1, but this does not seem to impact much the MISE of the two-step TiR estimator. Indeed if we use the optimal data driven regularisation parameter at each simulation, the MISE based on averages over the 1000 simulations is equal to .0120 for Case 1 and equal to .0144 for Case 2 when $\bar{\lambda} = .0005$ (resp., .0156 and .0175 when $\bar{\lambda} = .0001$), which are of the same magnitudes as the best MISE, which are .0099 and .0121 in Figures 9 and 10. In Case 1, the tendency of the selection procedure to overpenalized without unduly affecting efficiency is due to the flatness of the MISE curve.

We also get average values for the decay factors $\alpha$ and $\beta$ close to the asymptotic ones. These have been computed through estimating the coefficients of a linear fit for each simulation, and averaging over the 1000 simulations. For $\alpha$ the average (quartiles) is equal to 2.2502 (2.1456, 2.2641, 2.3628), and for $\beta$ it is equal to 2.9222 (2.8790, 2.9176, 2.9619).

To compute the average value for the decay factor $\gamma$ we have used an equally spaced grid of sample sizes $T \in \{500, 550, ..., 950, 1000\}$ in the variance component of the MISE,
together with the data driven estimate of $\theta$ in the bias component of the MISE. Optimizing on the grid of sample sizes yields an optimal lambda for each sample size per simulation. The logarithm of the optimal lambda is then regressed on the logarithm of the sample size, and the estimated slope is averaged over the 1000 simulations to obtain the average estimated gamma. In Case 1, we get an average (quartile) of $0.6081 (0.4908, 0.6134, 0.6979)$, when $\bar{\lambda} = .0005$, and $0.7224 (0.5171, 0.6517, 0.7277)$, when $\bar{\lambda} = .0001$. In Case 2, we get an average (quartile) of $0.5597 (0.4918, 0.5333, 0.5962)$, when $\bar{\lambda} = .0005$, and $0.5764 (0.4946, 0.5416, 0.6203)$, when $\bar{\lambda} = .0001$.

The second data driven selection procedure builds on the suggestion of Goh (2004) based on a subsampling procedure (also called the $m$-out-of-$n$ (moon) bootstrap). Even if his theoretical results are derived for semiparametric estimators we believe that they could be extended to our case as well. Recognizing that $\lambda_T = c T^{-\gamma}$ we propose to choose $c$ and $\gamma$ which minimize the following estimator of the MISE:

$$\hat{MISE}(c, \gamma) = \frac{1}{IJ} \sum_{i,j} \int_0^1 (\hat{\varphi}_{i,j}(x; c, \gamma) - \bar{\varphi}(x))^2 dx,$$

where $\hat{\varphi}_{i,j}(x; c, \gamma)$ denotes the estimator based on the $j$th subsample of size $m_i$ ($m_i << T$) with regularisation parameter $\lambda_{m_i} = cm_i^{-\gamma}$, and $\bar{\varphi}(x)$ denotes the estimator based on the original sample of size $T$ with a pilot regularisation parameter $\bar{\lambda}$ chosen sufficiently small to eliminate the bias.

In practice we have chosen 500 subsamples ($J = 500$) for each subsample size $m_i \in \{50, 60, 70, ..., 100\}$ ($I = 6$), $\bar{\lambda} = \{0.0005, 0.0001\}$, and $T = 1000$. To determine $c$ and $\gamma$ we have build a joined grid with values around the OLS estimates coming from Case 1, namely
\{.15, .2, .25\} \times \{.7, .75, .8\}, and with values around the OLS estimates coming from Case 2, namely \{.005, .01, .015\} \times \{.35, .4, .45\}. Note that the two grids yield a similar range for $\lambda_T$.

In the experiments for $\rho = 0.5$ we want to verify whether the data driven procedure is able to pick most of the time $c$ and $\gamma$ in the first set of values in Case 1, and in the second set of values in Case 2. On 1000 simulations we have found a frequency equal to 96\% of adequate choices in Case 1 when $\bar{\lambda} = .0005$, and 87\% when $\bar{\lambda} = .0001$. In Case 2 we have found 77\% when $\bar{\lambda} = .0005$, and 82\% when $\bar{\lambda} = .0001$. These frequencies are scattered among the grid values.
References


Figure 1: MISE (left panel), ISB (central panel) and estimated function (right panel) for the TiR estimator using Sobolev norm (solid line) and for OLS estimator (dashed line). The true function is the dotted line in the right panel, and corresponds to Case 1. Correlation parameter is $\rho = 0.5$, and sample size is $T = 400$. 
Figure 2: MISE (left panel), ISB (central panel) and estimated function (right panel) for the regularised estimator using $L^2$ norm (solid line) and for OLS estimator (dashed line). The true function is the dotted line in the right panel, and corresponds to Case 1. Correlation parameter is $\rho = 0.5$, and sample size is $T = 400$. 
Figure 3: MISE (left panel), ISB (central panel) and estimated function (right panel) for the TiR estimator using Sobolev norm (solid line) and for OLS estimator (dashed line). The true function is the dotted line in the right panel, and corresponds to Case 1. Correlation parameter is $\rho = 0$, and sample size is $T = 400$. 
Figure 4: MISE (left panel), ISB (central panel) and estimated function (right panel) for the regularised estimator using $L^2$ norm (solid line) and for OLS estimator (dashed line). The true function is the dotted line in the right panel, and corresponds to Case 1. Correlation parameter is $\rho = 0$, and sample size is $T = 400$. 
Figure 5: MISE (left panel), ISB (central panel) and estimated function (right panel) for the TiR estimator using Sobolev norm (solid line) and for OLS estimator (dashed line). The true function is the dotted line in the right panel, and corresponds to Case 2. Correlation parameter is $\rho = 0.5$, and sample size is $T = 400$. 
Figure 6: MISE (left panel), ISB (central panel) and estimated function (right panel) for the regularised estimator using $L^2$ norm (solid line) and for OLS estimator (dashed line). The true function is the dotted line in the right panel, and corresponds to Case 2. Correlation parameter is $\rho = 0.5$, and sample size is $T = 400$. 
Figure 7: MISE (left panel), ISB (central panel) and estimated function (right panel) for the TiR estimator using Sobolev norm (solid line) and for OLS estimator (dashed line). The true function is the dotted line in the right panel, and corresponds to Case 2. Correlation parameter is $\rho = 0$, and sample size is $T = 400$. 
Figure 8: MISE (left panel), ISB (central panel) and estimated function (right panel) for the regularised estimator using $L^2$ norm (solid line) and for OLS estimator (dashed line). The true function is the dotted line in the right panel, and corresponds to Case 2. Correlation parameter is $\rho = 0$, and sample size is $T = 400$. 
Figure 9: MISE (left panel), ISB (central panel) and estimated function (right panel) for the TiR estimator using Sobolev norm (solid line) and for OLS estimator (dashed line). The true function is the dotted line in the right panel, and corresponds to Case 1. Correlation parameter is $\rho = 0.5$, and sample size is $T = 1000$. 
Figure 10: MISE (left panel), ISB (central panel) and estimated function (right panel) for the TiR estimator using Sobolev norm (solid line) and for OLS estimator (dashed line). The true function is the dotted line in the right panel, and corresponds to Case 2. Correlation parameter is $\rho = 0.5$, and sample size is $T = 1000$. 

![Graph showing MISE, ISB, and estimated function for the TiR estimator using Sobolev norm and OLS estimator, with true function as a dotted line, and correlation parameter $\rho = 0.5$, and sample size $T = 1000$.]
Figure 11: The six largest eigenvalues (left Panel) and the $L^2$-norms of the corresponding eigenfunctions (right Panel) of operator $A^*A$. 
Figure 12: Log of optimal regularisation parameter as a function of log of sample size for Case 1 (left panel) and Case 2 (right panel). Correlation parameter is $\rho = 0.5$.

APPENDIX 1

Consistency of the TiR estimator

In this Appendix we prove the consistency of penalized extremum estimators

$$\hat{\varphi} = \arg \inf_{\varphi \in \Theta} Q_T(\varphi) + \lambda_T G(\varphi).$$

(A.1)

This covers the special case of the TiR estimator in Definition 1, where $G(\varphi) = \|\varphi\|_H^2$.

A.1.1 Existence and measurability of the estimator

From Theorem 2.2 of White and Wooldridge (1991), the estimator $\hat{\varphi}$ in (A.1) is well-defined and measurable if

(i) function $Q_T : \Omega \times \Theta_T \to \mathbb{R}$ is Borel-measurable, where $Q_T(\omega, \varphi)$ denotes the values of random variable $Q_T(\varphi)$ for event $\omega \in \Omega$, and $(\Omega, \mathcal{F}, P)$ is a complete probability space;
(ii) mappings $\varphi \rightarrow G(\varphi)$ and $\varphi \rightarrow Q_T(\omega, \varphi)$ are weakly lower semi-continuous on $\Theta_T$, $P$-a.s., for any $T$, w.r.t. the $L^2$ norm $\|\cdot\|$

(iii) set $\Theta_T$ is compact w.r.t. the $L^2$ norm $\|\cdot\|$ for any $T$.

A.1.2 Consistency of penalized extremum estimators

Proof of Theorem 1: For any $T$ and some given $\varepsilon > 0$, let us define $\varphi_T^* \in \Theta_T$ such that

$$Q_\infty (\varphi_T^*) + \lambda_T G(\varphi_T^*) = \inf_{\varphi \in \Theta_T : \|\varphi - \varphi_0\| \leq \varepsilon} Q_\infty (\varphi) + \lambda_T G(\varphi).$$

We have $P[\|\hat{\varphi} - \varphi_0\| > \varepsilon] \leq P[\inf_{\varphi \in \Theta_T : \|\varphi - \varphi_0\| \leq \varepsilon} Q_T(\varphi) + \lambda_T G(\varphi) \leq Q_T(\varphi_T^*) + \lambda_T G(\varphi_T^*)].$

Let us bound the probability on the RHS. Denoting $\Delta Q_T := Q_T - Q_\infty$, we get

$$\inf_{\varphi \in \Theta_T : \|\varphi - \varphi_0\| \geq \varepsilon} Q_T(\varphi) + \lambda_T G(\varphi) \leq Q_T(\varphi_T^*) + \lambda_T G(\varphi_T^*)$$

$$\Rightarrow \inf_{\varphi \in \Theta_T : \|\varphi - \varphi_0\| \geq \varepsilon} Q_\infty (\varphi) + \lambda_T G(\varphi) + \inf_{\varphi \in \Theta_T} \Delta Q_T(\varphi) \leq Q_\infty (\varphi_T^*) + \lambda_T G(\varphi_T^*) + \sup_{\varphi \in \Theta_T} |\Delta Q_T(\varphi)|$$

$$\Rightarrow \inf_{\varphi \in \Theta_T : \|\varphi - \varphi_0\| \geq \varepsilon} Q_\infty (\varphi) + \lambda_T G(\varphi) - \lambda_T G(\varphi_0) \leq \inf_{\varphi \in \Theta_T : \|\varphi - \varphi_0\| \leq \varepsilon} Q_\infty (\varphi) + \lambda_T [G(\varphi) - G(\varphi_0)]$$

$$+ 2 \sup_{\varphi \in \Theta_T} |\Delta Q_T(\varphi)|$$

$$\leq \inf_{\varphi \in \Theta_T : \|\varphi - \varphi_0\| \leq \varepsilon} Q_\infty (\varphi) + |G(\varphi) - G(\varphi_0)|$$

$$+ 2 \sup_{\varphi \in \Theta_T} |\Delta Q_T(\varphi)|$$

$$= \rho_T + 2\delta_T.$$

Thus, from (iii) we get for $a > 0$

$$P[\|\hat{\varphi} - \varphi_0\| > \varepsilon] \leq P[C_\varepsilon(\lambda_T) \leq \rho_T + 2\delta_T]$$

$$= P \left[ 1 \leq \frac{1}{\lambda_T^a} C_\varepsilon(\lambda_T) \frac{1}{(T\lambda_T)^\alpha} (T^a \rho_T + 2T^a \delta_T) \right] =: P[1 \leq Z_T].$$
Since $\lambda_T \to 0$ such that $(T\lambda_T)^{-1} \to 0$, $P$-a.s., for $a$ chosen as in (iv) we have $Z_T \overset{P}{\to} 0$, and we deduce $P[\|\hat{\varphi} - \varphi_0\| > \varepsilon] \leq P[Z_T \geq 1] \to 0$. Since $\varepsilon > 0$ is arbitrary, the proof is concluded.

Proof of Proposition 2: By contradiction, assume that Condition (iii) of Theorem 1 is not satisfied. Then there exists $\varepsilon > 0$ and a sequence $(\lambda_n)$ such that $\lambda_n \searrow 0$ and

$$C_\varepsilon (\lambda_n) \leq 0, \quad \forall n \in \mathbb{N}. \quad (21)$$

By definition of function $C_\varepsilon (\lambda)$, for any $\lambda > 0$ and $\eta > 0$, there exists $\varphi \in \Theta$ such that $\|\varphi - \varphi_0\| \geq \varepsilon$, and $Q_\infty(\varphi) + \lambda G (\varphi) - \lambda G (\varphi_0) \leq C_\varepsilon (\lambda) + \eta$. Setting $\lambda = \eta = \lambda_n$ for $n \in \mathbb{N}$, we deduce from (21) that there exists a sequence $(\varphi_n)$ such that $\varphi_n \in \Theta, \|\varphi_n - \varphi_0\| \geq \varepsilon$, and

$$Q_\infty(\varphi_n) + \lambda_n G (\varphi_n) - \lambda_n G (\varphi_0) \leq \lambda_n. \quad (22)$$

Now, since $Q_\infty(\varphi_n) \geq 0$, we get $\lambda_n G (\varphi_n) - \lambda_n G (\varphi_0) \leq \lambda_n$, that is

$$G (\varphi_n) \leq G (\varphi_0) + 1. \quad (23)$$

Moreover, since $G (\varphi_n) - G (\varphi_0) \geq G_0 - G (\varphi_0)$, we get $Q_\infty(\varphi_n) + \lambda_n G_0 - \lambda_n G (\varphi_0) \leq \lambda_n$ from (22), that is $Q_\infty(\varphi_n) \leq \lambda_n (1 + G (\varphi_0) - G_0)$, which implies

$$\lim_n Q_\infty(\varphi_n) = 0 = Q_\infty(\varphi_0). \quad (24)$$

Obviously, the simultaneous holding of (23) and (24) violates Assumption (10).

A.1.3 Penalization with Sobolev norm

54
In this Section we check that the assumptions in A.1.1 and A.1.2 hold for the special case

\[ G(\varphi) = \|\varphi\|_H^2 \] under Assumptions 1 and 2.

i) The mapping \( \varphi \rightarrow \|\varphi\|_H^2 \) is lower semi-continuous on \( H^2[0,1] \) w.r.t. the norm \( \|\cdot\| \) [see Reed and Simon (1980), p. 358].

ii) Let us verify that the assumptions of Proposition 2 are satisfied. Clearly function \( G(\varphi) = \|\varphi\|_H^2 \) is bounded from below by 0. Let us now check that Assumption (10) in Proposition 2 is satisfied.

**Lemma A.1:** Assumptions 1 and 2 imply Assumption (10) in Proposition 2.

**Proof:** Let \( \varepsilon > 0 \) and let \( (\varphi_n) \) be a sequence in \( \Theta \) such that \( \|\varphi_n - \varphi_0\| \geq \varepsilon \) for all \( n \in \mathbb{N} \), and

\[ Q_\infty(\varphi_n) \rightarrow 0 \text{ as } n \rightarrow \infty. \]  

(25)

We have to prove

\[ \|\varphi_n\|_H \rightarrow \infty \text{ as } n \rightarrow 0. \]  

(26)

To this aim, define sequence \( e_n = \frac{\varphi_n - \varphi_0}{\|\varphi_n - \varphi_0\|}, \ n \in \mathbb{N} \). Then, \( \|e_n\| = 1 \) for all \( n \in \mathbb{N} \), and from Assumption 1 and (25),

\[ \left\langle e_n, A^* Ae_n \right\rangle = \frac{\langle \Delta \varphi_n, A^* A \Delta \varphi_n \rangle}{\|\varphi_n - \varphi_0\|^2} \leq \frac{1}{\varepsilon^2 c_2} Q_\infty(\varphi_n) \rightarrow 0, \]

as \( n \rightarrow \infty \). Let \( \Pi(N) \) denote the orthogonal projection [w.r.t. the scalar product \( \langle \cdot, \cdot \rangle \)] on the subspace spanned by \( \{\psi_1, ..., \psi_N\} \). Then we have for any \( N \in \mathbb{N} \)

\[ \|\Pi(N)e_n\|^2 = \sum_{j=1}^{N} \langle \psi_j, e_n \rangle^2 \leq \frac{1}{\mu_N} \sum_{j=1}^{N} \mu_j \langle \psi_j, e_n \rangle^2 \leq \frac{1}{\mu_N} \sum_{j=1}^{\infty} \mu_j \langle \psi_j, e_n \rangle^2 \]

\[ = \frac{1}{\mu_N} \langle e_n, A^* Ae_n \rangle \rightarrow 0, \text{ as } n \rightarrow \infty, \]

that is \( \|\Pi(N)e_n\| \rightarrow 0 \) as \( n \rightarrow \infty \), for any \( N \in \mathbb{N} \).
Let us now derive a lower bound for the Sobolev norm $\|e_n\|_H$. We have

$$\|e_n\|_H \geq \|\Pi_{(N)}^+ e_n\|_H - \|\Pi_{(N)} e_n\|_H,$$

where $\Pi_{(N)}^+ = 1 - \Pi_{(N)}$ denotes the orthogonal projection on $\text{span}\{\psi_j : j \geq N + 1\}$. Let us derive bounds for the two terms in the RHS of (27). We have

$$\|\Pi_{(N)}^+ e_n\|_H = \left\| \sum_{j=N+1}^{\infty} \langle \psi_j, e_n \rangle \psi_j \right\|_H = \left\| \sum_{j=N+1}^{\infty} \frac{\langle \psi_j, e_n \rangle}{\sum_{j=N+1}^{\infty} \langle \psi_j, e_n \rangle^2} \right\|_H^{1/2} \left( \sum_{j=N+1}^{\infty} \langle \psi_j, e_n \rangle^2 \right)^{1/2} \geq \inf_{\varphi \in S_{N+1}:\|\varphi\|=1} \|\varphi\|_H \left( \sum_{j=N+1}^{\infty} \langle \psi_j, e_n \rangle^2 \right)^{1/2} = M_{N+1} \left( 1 - \|\Pi_{(N)} e_n\|_2^2 \right)^{1/2},$$

since $\|e_n\| = 1$, where $S_{N+1} = \text{span}\{\psi_j : j \geq N + 1\}$, and $M_{N+1} = \inf_{\varphi \in S_{N+1}:\|\varphi\|=1} \|\varphi\|_H$. Moreover,

$$\|\Pi_{(N)} e_n\|_H^2 = \left\| \sum_{j=1}^{N} \langle \psi_j, e_n \rangle \psi_j \right\|_H^2 = \sum_{j,l=1}^{N} \langle \psi_j, e_n \rangle \langle \psi_l, e_n \rangle \langle \psi_j, \psi_l \rangle_H \leq \max_{j=1,\ldots,N} \|\psi_j\|_H \sum_{j,l=1}^{N} |\langle \psi_j, e_n \rangle| |\langle \psi_l, e_n \rangle| = M_N \left( \sum_{j=1}^{N} |\langle \psi_j, e_n \rangle| \right)^2 \leq N \overline{M}_N \sum_{j,l=1}^{N} \langle \psi_j, e_n \rangle^2 = N \overline{M}_N \|\Pi_{(N)} e_n\|_2^2,$$

where $\overline{M}_N = \max_{j=1,\ldots,N} \|\psi_j\|_H^2$. Thus, we get from (27)

$$\|e_n\|_H \geq M_{N+1} \left( 1 - \|\Pi_{(N)} e_n\|_2^2 \right)^{1/2} - c_N \|\Pi_{(N)} e_n\|_H,$$

for any $N$ and $n \in \mathbb{N}$, where $c_N = \sqrt{N \overline{M}_N}$. Note that, since $\overline{M}_N \geq \|\psi_N\|_H^2 \geq M_N^2$, by Assumption 2 we have $c_N \to \infty$ as $N \to \infty$. Since the bound (28) holds for any $N \in \mathbb{N}$, it follows

$$\|e_n\|_H \geq M_{N+1} \left( 1 - \|\Pi_{(N)} e_n\|_2^2 \right)^{1/2} - c_N \|\Pi_{(N)} e_n\|_H, \text{ for any } n \in \mathbb{N},$$

(29)
for any sequence of integers \((N_n)\).

Let us now prove that there exists a sequence of integers \((N_n)\) such that the RHS of (29) diverges. To this goal, define the sequence \(n(N), N = 1, 2, \ldots\) recursively by

\[
\begin{align*}
n(1) &= \min \{ n^* \in \mathbb{N} \mid c_1 \| \Pi(1)e_n \| \leq 1 \text{ for all } n \geq n^* \}, \\
n(N) &= \min \{ n^* \in \mathbb{N} \mid n^* > n(N-1), c_N \| \Pi(N)e_n \| \leq 1 \text{ for all } n \geq n^* \}, \quad N = 2, \ldots
\end{align*}
\]

Since \(c_N \| \Pi(N)e_n \| \to 0\) as \(n \to \infty\), for any \(N \in \mathbb{N}\), it follows that \(n(N) < \infty\), for any \(N \in \mathbb{N}\), and the sequence \(n(N), N = 1, 2, \ldots\) is strictly increasing. Then, let the sequence of integers \((N_n)\), for \(n \geq n(1)\), be defined by

\[
N_n = \begin{cases} 
1 & \text{if } n(1) \leq n < n(2), \\
2 & \text{if } n(2) \leq n < n(3), \\
& \vdots \\
& \vdots
\end{cases}
\]

By construction, we have

\[
c_{N_n} \| \Pi(N_n)e_n \| \leq 1, \tag{30}
\]

for any \(n \geq n(1)\). Since \(c_N \to \infty\) as \(N \to \infty\), we deduce

\[
\| \Pi(N_n)e_n \| \leq 1/2, \quad \forall n \text{ large enough.} \tag{31}
\]

Using Bounds (30) and (31) in Inequality (29), we get \(\|e_n\|_H \geq M_{N_n+1} (3/4)^{1/2} - 1 \to \infty\), as \(n \to \infty\), from Assumption 2.

Finally, we get

\[
\| \varphi_n \|_H = \| \varphi_n - \varphi_0 \|_H + \| \varphi_n \|_H \geq \| \varphi_n - \varphi_0 \|_H + \| \varphi_0 \|_H \\
\geq \| e_n \|_H - \| \varphi_0 \|_H \to \infty. \tag{32}
\]
Therefore, (26) follows, and the proof is concluded. ■
Appendix 2

The MISE of the TiR estimator

In this Appendix we derive the asymptotic expansion of the MISE with deterministic sequence of regularisation parameters (Proof of Proposition 3). We focus on the linear IV case $m(\varphi, z) = E_0 [\varphi (X) - Y | Z = z] = (A\varphi) (z) - r (z)$, where $(A\varphi) (z) = \int \varphi (x) f (w | z) dw$ and $r(z) = \int y f (w | z) dw$.

A.2.1 The first-order condition

The objective function of the TiR estimator becomes in the linear case

$$Q_T (\varphi) + \lambda_T \| \varphi \|^2_H = \frac{1}{T} \sum_{t=1}^{T} \left[ (\hat{A}\varphi) (Z_t) - \hat{r} (Z_t) \right]^2 + \lambda_T \langle \varphi, \varphi \rangle_H. \quad (33)$$

Let us now prove that this objective function can be written as a quadratic form in $\varphi \in H^2[0, 1]$. To this aim, let us introduce the dual operator $\hat{A}^*$ of $\hat{A}$.

**Lemma A.1**: Under regularity conditions, the following properties hold $P$-a.s.:

(i) Function $\hat{r}$ is in $L^2(F_Z)$;

(ii) Operator $\hat{A}$ maps $H^2[0, 1]$ into $L^2(F_Z)$;

(iii) There exists a linear operator $\hat{A}^*$ from $L^2(F_Z)$ into $H^2[0, 1]$, such that

$$\left( h, \hat{A}^* \psi \right)_H = \frac{1}{T} \sum_{t=1}^{T} \left( \hat{A} h \right) (Z_t) \psi (Z_t), \text{ for any } \psi \in L^2(F_Z) \text{ and } h \in H^2[0, 1];$$

(iv) Operator $\hat{A}^* : H^2[0, 1] \to H^2[0, 1]$ is compact.

**Proof**: See Appendix B.
Then, from Lemma A.1 i)-iii), Criterion (33) can be rewritten as

$$Q_T (\varphi) + \lambda_T \| \varphi \|^2_H = \langle \varphi, \left( \lambda_T + \hat{A}^* \hat{A} \right) \varphi \rangle_H - 2 \langle \varphi, \hat{A}^* \hat{r} \rangle_H, $$

(34)

up to a term independent of $\varphi$. From Lemma A.1 iv), $\hat{A}^* \hat{A}$ is a compact operator from $H^2[0,1]$ in itself. Since $\hat{A}^* \hat{A}$ is positive, operator $\lambda_T + \hat{A}^* \hat{A}$ is invertible [Kress (1999), Theorem 3.4]. It follows that the quadratic criterion function (34) admits a global minimum over $H^2[0,1]$. It is given by the first-order condition $(\hat{A}^* \hat{A} + \lambda_T) \hat{\varphi}_T = \hat{A}^* \hat{r}$, that is

$$\hat{\varphi} = \left( \lambda_T + \hat{A}^* \hat{A} \right)^{-1} \hat{A}^* \hat{r}. $$

(35)

A.2.2 The asymptotic expansion of the first-order condition

Let us now expand the estimator in (35). We can write

$$\hat{r}(z) = \int (y - \varphi_0(x)) \Delta \hat{f}(w|z)dw + \int \varphi_0(x) \hat{f}(w|z)dw =: \hat{\psi}(z) + \left( \hat{A} \varphi_0 \right)(z),$$

where $\Delta \hat{f}(w|z) := \hat{f}(w|z) - f(w|z)$. Hence, $\hat{\varphi} = \left( \lambda_T + \hat{A}^* \hat{A} \right)^{-1} \hat{A}^* \hat{\psi} + \left( \lambda_T + \hat{A}^* \hat{A} \right)^{-1} \hat{A}^* \hat{A} \varphi_0,$

which yields

$$\hat{\varphi} - \varphi_0 = (\lambda_T + A^* A)^{-1} A^* \hat{\psi} + [(\lambda_T + A^* A)^{-1} A^* A \varphi_0 - \varphi_0] + R_T$$

$$=: V_T + D(\lambda_T) + R_T,$$

(36)

where the remaining term $R_T$ is given by

$$R_T = \left[ (\lambda_T + \hat{A}^* \hat{A})^{-1} \hat{A}^* - (\lambda_T + A^* A)^{-1} A^* \right] \hat{\psi}$$

$$+ \left[ (\lambda_T + \hat{A}^* \hat{A})^{-1} \hat{A}^* \hat{A} - (\lambda_T + A^* A)^{-1} A^* A \right] \varphi_0.$$
Lemma A.2: Assume the bandwidth conditions $h_m T = o(\lambda T)$, $\lambda / T = o(h_{dZ}^m)$, where $m$ is the order of the kernel $K$, and $dZ$ the dimension of $Z$. Then, under regularity assumptions, $E[\|R_T\|^2] = o\left(E[\|V_T + D(\lambda T)\|^2]\right)$.

Proof: See Appendix B.

From (36) we deduce

$$
E[\|\hat{\varphi} - \varphi_0\|^2] = E[\|V_T + D(\lambda T)\|^2] + E[\|R_T\|^2] + 2E[(V_T + D(\lambda T), R_T)]
$$

$$
= E[\|V_T + D(\lambda T)\|^2] + o\left(E[\|V_T + D(\lambda T)\|^2]\right),
$$

by applying twice the Cauchy-Schwarz inequality and Lemma A.2. Since

$$
E[\|V_T + D(\lambda T)\|^2] = \left\| (\lambda_T + A^* A)^{-1} A^* A \varphi_0 - \varphi_0 + (\lambda_T + A^* A)^{-1} A^* E \hat{\psi} \right\|^2
$$

$$
+ E\left[ \left\| (\lambda_T + A^* A)^{-1} A^* \left( \hat{\psi} - E \hat{\psi} \right) \right\|^2 \right],
$$

(37)

we get

$$
E[\|\hat{\varphi} - \varphi_0\|^2] = \left\| (\lambda_T + A^* A)^{-1} A^* A \varphi_0 - \varphi_0 + (\lambda_T + A^* A)^{-1} A^* E \hat{\psi} \right\|^2
$$

$$
+ E\left[ \left\| (\lambda_T + A^* A)^{-1} A^* \left( \hat{\psi} - E \hat{\psi} \right) \right\|^2 \right],
$$

(38)

up to a term which is asymptotically negligible w.r.t. the RHS. This asymptotic expansion consists of a bias term (regularisation bias plus estimation bias) and a variance term, which will be analysed separately below in Lemma A.3 and A.4. Combining these two Lemmas and the asymptotic expansion in (38) results in Proposition 3.

A.2.3 Asymptotic expansion of the variance term

61
Lemma A.3: Under regularity conditions, up to a term which is asymptotically negligible w.r.t. the RHS, we have
\[ E \left[ \left\| (\lambda T + A^* A)^{-1} A^* \left( \hat{\psi} - E \hat{\psi} \right) \right\|^2 \right] = \frac{1}{T} \sum_{j=1}^{\infty} \frac{\nu_j}{(\lambda T + \nu_j)^2} \| \phi_j \|^2. \]

Proof: See Appendix B.

A.2.4 Asymptotic expansion of the bias term

Lemma A.4: Define \( b(\lambda T) = \left\| (\lambda T + A^* A)^{-1} A^* A \varphi_0 - \varphi_0 \right\|. \) Then, under regularity conditions and the bandwidth condition \( h_T^m = o(\lambda_T b(\lambda_T)), \) where \( m \) is the order of the kernel \( K, \)
we have \( \left\| (\lambda T + A^* A)^{-1} A^* A \varphi_0 - \varphi_0 + (\lambda T + A^* A)^{-1} A^* E \hat{\psi} \right\| = b(\lambda T), \) up to a term which is asymptotically negligible w.r.t. the RHS.

Proof: See Appendix B.
Appendix 3

Rate of convergence with geometric spectrum

In this Appendix we prove Proposition 4.

i) The next Lemma A.5 characterizes the variance term of the asymptotic expansion of the MISE in Proposition 3.

**Lemma A.5:** Let \( \nu_j \) and \( \| \phi_j \| \) satisfy Assumption 5, and define the function
\[
I(\lambda) = \sum_{j=1}^{\infty} \frac{\nu_j}{(\lambda + \nu_j)^2} \| \phi_j \|^2,
\]
\( \lambda > 0 \). Then,
\[
\lim_{\lambda \to 0} \lambda \left[ \log \left( \frac{1}{\lambda} \right) \right]^{\beta} I(\lambda) = \left( \frac{1}{\alpha} \right)^{1-\beta} C_2.
\]

**Proof:** See Appendix B.

From Lemma A.5 and using Assumption 6, we get
\[
M_T(\lambda) = \frac{1}{T} c_1 \frac{1}{\lambda \left[ \log \left( \frac{1}{\lambda} \right) \right]^2} + c_2 \lambda^{2\delta},
\]
for \( \lambda \to 0 \) and \( T \to \infty \), where \( c_1 = \left( \frac{1}{\alpha} \right)^{1-\beta} C_2 \), \( c_2 = C_3^2 \).

ii) The optimal sequence \( \lambda^*_T \) is obtained by minimizing function \( M_T(\lambda) \) w.r.t. \( \lambda \). We have
\[
\frac{dM_T(\lambda)}{d\lambda} = -\frac{1}{T} c_1 \frac{1}{\lambda^2 \left[ \log \left( \frac{1}{\lambda} \right) \right]^{2\beta}} \left( \log \left( \frac{1}{\lambda} \right) - \lambda \beta \left[ \log \left( \frac{1}{\lambda} \right) \right]^{\beta-1} \frac{1}{\lambda} \right) + 2c_2 \delta \lambda^{2\delta-1}
\]
\[
= -\frac{1}{T} c_1 \frac{\log \left( \frac{1}{\lambda} \right) - \beta}{\lambda^{2\beta+1}} + 2c_2 \delta \lambda^{2\delta-1}.
\]
Thus
\[
\frac{dM_T(\lambda^*_T)}{d\lambda} = 0 \iff \frac{c_1}{T} \frac{1}{2c_2 \delta \left[ \log \left( \frac{1}{\lambda^*_T} \right) \right]^{2\beta+1}} = (\lambda^*_T)^{2\delta+1}.
\]
(39)

To solve the latter equation for \( \lambda^*_T \), define \( \tau_T := \log \left( \frac{1}{\lambda^*_T} \right) \). Then \( \tau_T \) satisfies
\[
\tau_T = c_3 + \frac{1}{1 + 2\delta} \log T + \frac{1 + \beta}{1 + 2\delta} \log \tau_T - \frac{1}{1 + 2\delta} \log (\tau_T - \beta),
\]
63
where $c_3 = (1 + 2\delta)^{-1} \log (2c_2\delta/c_1)$. It follows that

$$\tau_T = c_4 + \frac{1}{1 + 2\delta} \log T + \frac{1 + \beta}{1 + 2\delta} \log \log T + o (\log \log T),$$

for some constant $c_4$, that is

$$\log (\lambda_T^*) = -c_4 - \frac{1}{1 + 2\delta} \log T - \frac{1 + \beta}{1 + 2\delta} \log \log T + o (\log \log T).$$

iii) Finally, let us compute the MISE corresponding to $\lambda_T^*$. We have

$$M_T (\lambda_T^*) = \frac{1}{T} c_1 \frac{1}{\lambda_T^* [\log (1/\lambda_T^*)]^{\beta}} + c_2 (\lambda_T^*)^{2\delta} = \frac{1}{T} c_1 \frac{1}{\lambda_T^{*-\frac{\beta}{T}}} + c_2 (\lambda_T^*)^{2\delta}.$$

From (39), $\lambda_T^* = \left( \frac{c_1}{2c_2\delta} \right)^{1/2} T^{-1/2} \left( \frac{\tau_T - \beta}{\tau_T^*} \right)^{1/2} = c_5 T^{-1/2} \tau_T^{\beta - \frac{\beta}{T^*}}$, for some constant $c_5$, up to a term which is negligible w.r.t. the RHS. Thus we get

$$M_T (\lambda_T^*) = \frac{1}{T} c_5 T^{-\frac{1}{2\delta}} \frac{1}{\tau_T^{\frac{\beta}{2\delta}} + \beta} + c_2 c_5^{2\delta} T^{-\frac{2\delta}{2\delta + 1}} \tau_T^{-\frac{\beta}{2\delta + 1}} = c_6 T^{-\frac{2\delta - 1}{2\delta + 1}} \tau_T^{\frac{-\beta}{2\delta + 1}} = c_7 T^{-\frac{2\delta}{2\delta + 1}} (\log T)^{-\frac{2\delta - 1}{2\delta + 1}}$$

for some constants $c_6$ and $c_7$, up to a term which is negligible w.r.t. the RHS.