TIKHONOV REGULARISATION FOR FUNCTIONAL MINIMUM DISTANCE ESTIMATORS

P. Gagliardini* and O. Scaillet†

This version: June 2006 ‡

(First version: May 2006)

*University of Lugano and Swiss Finance Institute.
†HEC Genève and Swiss Finance Institute.
‡Both authors received support by the Swiss National Science Foundation through the National Center of Competence in Research: Financial Valuation and Risk Management (NCCR FINRISK). We would like to thank the seminar participants at University of Geneva for helpful comments.
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 asymptotic 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. Finally, we provide an empirical application to nonparametric estimation of an Engel curve.

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

JEL classification: C13, C14, C15, D12.

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 is 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 Parzen-Rosenblatt kernel estimator. Recently, mainly motivated by the interest for IV functional estimation of structural equations, several authors have made contributions to correct for endogeneity in the nonparametric context as well. Newey and Powell (NP, 2003) consider the problem of estimating nonparametrically 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 nonparametric analog of the Two-Stage Least Squares estimator. The NP methodology extends to the general case of conditional moment restrictions which depend nonlinearly on functional parameters. Ai and Chen (AC, 2003) opt for a similar approach to estimate semiparametric specifications, in which the conditional moment restrictions contain both a finite-dimensional and a functional parameter. Although their focus is on the efficient estimation of the finite-dimensional component, AC 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) concentrate on nonparametric estimation of an instrumental regression function. 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. For further background, Florens (2003) and Blundell and Powell (2003) present surveys on endogenous nonparametric regressions.

There is a growing literature extending the above methods and considering empirical applications to different fields. Among others, Blundell, Chen and Kristensen (2004) estimate nonparametrically Engel curves with endogenous total expenditures; Chen and Ludvigson (2004) consider asset pricing models with functional specifications of habit formation; Chernozhukov, Imbens and Newey (2006) estimate non-separable models via quantile conditions; Loubes and Vanhems (2004) discuss the estimation of the solution of a differential equation with endogenous variables for microeconomic applications. Other related references include Newey, Powell, and Vella (1999), Chernozhukov and Hansen (2005), Carrasco and Florens (2005), Florens, Johannes and Van Bellegem (2005), and Horowitz (2006).

The main theoretical difficulty in nonparametric estimation with endogeneity is overcoming ill-posedness [see Kress (1999), Chapter 15, for a general treatment of ill-posed inverse problems, and Carrasco, Florens and Renault (2005) for a survey on ill-posed inverse problems in econometrics and on main properties of operators in function spaces]. 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 has 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 functional parameter of interest and its derivatives, which amounts to set compacity on the 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 $G(\varphi)$ is given by the Sobolev norm of function $\varphi$, which involves the $L^2$ norms of both $\varphi$ and its derivative $\nabla \varphi$. The basic idea is that the penalty term $\lambda_T G(\varphi)$ damps highly oscillating components of the estimator. These oscillations are otherwise unduly amplified 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 by reference to the pioneering papers of Tikhonov (1963a,b) where regularisation is achieved via a penalty term incorporating the function and its derivative [see Kress (1999) and Groetsch (1984) for an extensive discussion]. We stress that the regularisation approach in DFR and HH can be viewed as a Tikhonov regularisation, but with a penalty term in-
volving the $L^2$ norm of the function (without any derivative) instead of the Sobolev norm of the parameter. To avoid confusion, we refer to the DFR and HH estimators as regularised estimators with $L^2$ norm.

Our paper contributes to the literature along several directions. First, we introduce an estimator admitting the following appealing features: (i) it applies in a 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 $L^2$ regularised estimators in the linear case (DFR, HH); (iv) it has a faster rate of convergence than estimators based on bounding the Sobolev norm (NP, AC); (v) 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 their theoretical results. Concerning point (iii), we give in Section 4 the condition under which this property holds. In our Monte-Carlo experiments in Section 6, we find a clear-cut superior performance of the TiR estimator compared to the regularised estimator with $L^2$ norm.  

Point (iv) is induced by the requirement of a fixed bound in the approach of NP and AC. Point (v) is not shared by NP and AC estimators because of the inequality constraint. We will further explain the links between the TiR estimator and the literature in Section 2.4.

Second, we study in depth the asymptotic properties of our estimator. In particular: (a)

---

1 The advantage of the Sobolev norm compared to the $L^2$ norm for regularisation of ill-posed inverse problems is also pointed out in a numerical example in Kress (1999), Example 16.21.
we prove the consistency of the TiR estimator; (b) we derive the asymptotic expansion of the Mean Integrated Squared Error (MISE) as a function of the sample size and the (deterministic) regularisation parameter; (c) we characterize the MSE, and prove the pointwise asymptotic normality of the TiR estimator. To the best of our knowledge, results (b) and (c), as well as (a) for a sequence of stochastic regularisation parameters, are new for non-parametric instrumental regression estimators. 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 by their own for the study of the properties of $L^2$ regularised 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 resampling techniques for example. The finite sample properties seem very appealing from our numerical experiments on two examples mimicking possible shapes of Engel curves and with 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 nonparametric 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 its consistency. Section 4 is first devoted to the characterisation of the asymptotic MISE and of the optimal rates of convergence of the TiR estimator. We compare these results with those obtained under regularization via an $L^2$ norm instead of a Sobolev norm. We further discuss the suboptimality of bounding the Sobolev norm. We derive the asymptotic MSE and establish pointwise asymptotic normality of the TiR estimator. Implementation of the TiR estimator for linear moment restrictions is outlined in Section 5. In Section 6 we illustrate numerically our theoretical findings, and present a Monte-Carlo study of the finite sample properties of the TiR estimator. We also describe two data driven selection procedures of the regularization parameter, and show that they perform well in practice. We provide an empirical example in Section 7 where we estimate an Engel curve nonparametrically. Section 8 concludes. The proofs of all results in the paper are gathered in the Appendices. Some lengthy proofs of technical Lemmas are omitted for the sake of space, and are collected in a technical report, which is available from 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 nonparametric 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. copies of variables \((Y, X, Z)\), and let the support of \((Y, Z)\) be a subset of an Euclidean space while the support of \(X\) is \(X = [0, 1]\). \(^2\) Suppose that the parameter of interest is a function \(\varphi_0\) defined on \(X\), which satisfies the conditional moment restriction

\[
E_0 \left[ g(Y, \varphi_0(X)) \mid Z \right] = 0, \tag{1}
\]

where \(g\) is a known function. Parameter \(\varphi_0\) belongs to a 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 \(\langle \varphi, \psi \rangle = \int_X \varphi(x)\psi(x)dx\). Sobolev space \(H^2[0, 1]\) is an Hilbert space w.r.t. the scalar product \(\langle \varphi, \psi \rangle = \langle \varphi, \psi \rangle_H = \langle \varphi, \psi \rangle + \langle \nabla \varphi, \nabla \psi \rangle\), and the corresponding Sobolev norm is denoted by \(\|\varphi\|_H = \langle \varphi, \varphi \rangle_H^{1/2}\). The \(L^2\) norm \(\|\varphi\| = \langle \varphi, \varphi \rangle_{L^2}^{1/2}\) is used as consistency norm, and set \(\Theta\) is bounded and closed w.r.t. this norm. Further, we assume the following identification condition. \(^3\)

**Assumption 1 (Identification):** \(\varphi_0\) is the unique function \(\varphi \in \Theta\) that satisfies the conditional moment restriction (1).

\(^2\) We need compactness of the support of \(X\) for technical reasons. Mapping in [0,1] can be achieved by simple linear or nonlinear monotone transformations. Moreover, assuming univariate \(X\) simplifies the exposition. Extension of our theoretical results to higher dimensions is straightforward.

\(^3\) See NP, Theorems 2.2-2.4, for sufficient conditions ensuring identification (Assumption 1) in the linear setting, and Chernozhukov and Hansen (2005) for sufficient conditions in a nonlinear setting.
The nonparametric Minimum Distance approach to estimate $\varphi_0$ as in AC and NP relies on $\varphi_0$ minimizing the criterion
\[
Q_\infty(\varphi) = E_0 \left[ m(\varphi, Z)' \Omega_0(Z)m(\varphi, Z) \right], \quad \varphi \in \Theta,
\] (2)
where $m(\varphi, z) = E_0 [g(Y, \varphi(X)) \mid Z = z]$, 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 $\langle \psi_1, \psi_2 \rangle_{L^2_{\Omega_0}(F_Z)} = E_0 \left[ \psi_1(Z)' \Omega_0(Z) \psi_2(Z) \right]$. 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 polynomial approach, and minimize the empirical criterion over a finite-dimensional Sieve approximation of $\Theta$ based on polynomial or spline functions.

The main difficulty in nonparametric Minimum Distance estimation is that, contrary to the standard parametric case, Assumption 1 on identification 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) and Carrasco, Florens and Renault (2005)]. To briefly explain what ill-posedness is, note that solving the integral 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 \( F_0(y,x|z) \) 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 lack of 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 unidentifiability.

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

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

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

\[
Q_\infty(\varphi) = E_0[(A\Delta \varphi)(Z) \Omega_0(Z)(A\Delta \varphi)(Z)] = \langle \Delta \varphi, A^*A\Delta \varphi \rangle_H,
\]

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

Under weak regularity conditions, integral operator \( A \) is compact in \( L^2[0,1] \). Thus, \( A^*A \) is compact and self-adjoint in \( H^2[0,1] \). We denote by \( \{\phi_j : j \in \mathbb{N}\} \) an orthonormal basis in \( H^2[0,1] \) of eigenfunctions of operator \( A^*A \), and by \( \nu_1 \geq \nu_2 \geq \cdots \), with \( \nu_j > 0 \), 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 $\nu_j \to 0$, and it can be shown that $\nu_j / \|\phi_j\|^2 \to 0$. The limit criterion $Q_\infty(\varphi)$ can be minimized by a sequence in $\Theta$ such as

$$\varphi_n = \varphi_0 + \varepsilon \frac{\phi_n}{\|\phi_n\|}, \quad n \in \mathbb{N},$$

for $\varepsilon > 0$, which does not converge to $\varphi_0$ in $L^2$-norm $\|\|$. Indeed, we have $Q_\infty(\varphi_n) = \varepsilon^2 \langle \phi_n, A^*A\phi_n \rangle / \|\phi_n\|^2 = \varepsilon^2 \nu_n / \|\phi_n\|^2 \to 0$ 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 "identifiable uniqueness" assumption [e.g., White and Wooldridge (1991)]

$$\inf_{\varphi \in \Theta: \|\varphi - \varphi_0\| \geq \varepsilon} Q_\infty(\varphi) > 0 = Q_\infty(\varphi_0), \quad \text{for } \varepsilon > 0,$$

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 Condition (6) despite validity of Assumption 1 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 2 (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 $(A\Delta \varphi)(z) = \int \frac{\partial q}{\partial \varphi_0}(y, \varphi_0(x)) f(w|z) \Delta \varphi(x) dw$ is a compact operator in $L^2[0, 1]$. 

11
(ii) the second-order term \( R(\varphi, z) \) is such that \( \sup_{\varphi \in \Theta} \| R(\varphi, .) \|_{L^2_{\Omega_0}(F_2)} / \| A\Delta \varphi \|_{L^2_{\Omega_0}(F_2)} < 1 \).

Under Assumption 2, the identification condition (6) is not satisfied, and the Minimum Distance estimator which minimizes the empirical counterpart of criterion \( Q_\infty(\varphi) \) over set \( \Theta \) (or a Sieve approximation of \( \Theta \)) is not consistent w.r.t. the \( L^2 \)-norm \( \| . \| \). Finally, note that Assumption 2 (ii) and injectivity of operator \( A \) imply Assumption 1.

2.3 The Tikhonov Regularised (TiR) estimator

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 nonparametrically by \( \hat{m}(\varphi, z) = \int g(y, \varphi(x)) \hat{f}(w \mid z) \, dw \), where \( \hat{f}(w \mid z) \) 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 $G(\varphi) = \|\varphi\|^2_H$. 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

$$\hat{\varphi} = \arg\inf_{\varphi \in \Theta} Q_T(\varphi) + \lambda_T \|\varphi\|^2_H,$$  \hspace{1cm} (7)

where $\lambda_T$ is a stochastic sequence such that $\lambda_T \geq 0$ and $\lambda_T \to 0$, $P$-a.s..

The name Tikhonov Regularised (TiR) estimators that we use to characterize the Minimum Distance estimators introduced in Definition 1 goes back to Tikhonov (1963a,b), in his pioneering papers 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. They would be otherwise unduly amplified, since the criterion $Q_T(\varphi)$ becomes asymptotically flat along some directions because of ill-posedness. For instance, in the linear IV case where $Q_\infty(\varphi) = \langle \Delta \varphi, A^* A \Delta \varphi \rangle_H$, these directions are spanned by the eigenfunctions $\phi_n$ of operator $A^* A$ to eigenvalues $\nu_n$ close to zero, that is for large $n$ [see Equation (5) and the discussion in Section 2.2]. Typically, $\psi_n := \phi_n / \|\phi_n\|$ 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 \( G(\varphi) = \|\varphi\|_{H}^2 \), 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 the choice of the sequence \((\lambda_T)\) to achieve an optimal rate of convergence of TiR estimator \( \hat{\varphi} \), and we will present two global data driven selection procedures for \( \lambda_T \) in Section 6.

### 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 (2004)] suggest considering a compact parameter set \( \Theta \). In this case, by the same argument as in the standard parametric setting, Assumption 1 implies identification Condition (6).

Compact sets in \( L^2[0,1] \) w.r.t. the \( L^2 \) norm \( \| \cdot \| \) can be obtained by imposing a bound on the Sobolev norm \( \| \varphi \|_H \leq \bar{B} \) of the functional parameter. Then, the estimator is derived by solving minimization problem (7), where \( \lambda_T \) is interpreted as a Kuhn-Tucker multiplier.

Our approach differs from AC and NP along two directions. On the one hand, NP and AC use finite-dimensional Sieve estimators [see Chen (2006) for an extensive introduction on Sieve estimation in econometrics]. By contrast, we define the TiR estimator and study
its asymptotic properties as an estimator on a function space \(^4\), and we introduce a finite dimensional basis of functions only to approximate numerically the estimator (see Section 5).

On the other hand, \(\lambda_T\) is a free regularisation parameter for TiR estimators, whereas \(\lambda_T\) is tight down by the slackness condition in NP and AC approach, namely either \(\lambda_T = 0\) or \(\|\hat{\varphi}\|_H = \bar{B}, \text{ P-a.s.}\). As a consequence, our approach presents three important advantages.

i) Although, for given sample size \(T\), selecting different \(\lambda_T\) amounts to select different \(\bar{B}\) when the constraint is binding, the asymptotic properties of the TiR estimator and of the estimators with fixed \(\bar{B}\) are different. In particular, putting a bound \(\bar{B}\) on the Sobolev norm independent of sample size \(T\) implies in general the selection of a sub-optimal sequence of regularisation parameters \(\lambda_T\) (see Section 4.3). Thus, the estimators with fixed \(\bar{B}\) share rates of convergence which are slower than that of the TiR estimator with optimally selected sequence of regularisation parameter. \(^5\)

ii) 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 \(\bar{B}\) 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. For instance, we prove consistency in Theorem 1 and Proposition 2 of Section 3.

iii) Finally, we emphasize that the TiR estimator enjoys important computational ad-

---

\(^4\) See also NP at p. 1573 for such a suggestion.

\(^5\) Note that letting \(\bar{B} = \bar{B}_T\) grow (slowly) with sample size \(T\) without introducing a penalty term is not equivalent to our approach and does not guarantee the consistency of the estimator. Indeed, when \(\bar{B}_T \to \infty\), the resulting limit parameter set \(\Theta\) is not compact.
vantages. This is because, for given $\lambda_T$, the TiR estimator is defined by an unconstrained optimization problem, whereas inequality constraint $\|\phi\|_H \leq \bar{B}$ has to be accounted for in the minimization defining estimators with given $\bar{B}$. In particular, in the case of linear conditional moment restrictions, the TiR estimator admits a closed form (see Section 5), whereas the computation of the NP and AC estimator requires a numerical constrained quadratic optimization routine.

2.4.2 Regularisation with $L^2$ norm

For the special case of nonparametric linear 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 $\|\phi\|_H$ replaced by $L^2$ norm $\|\phi\|$ in the penalty term, and $\Omega_0(z) = 1$. Indeed, 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 (see the remark by DFR at p. 19). Thus, our approach differs from DFR and HH for the norm adopted for penalization. As already pointed out in the Introduction, the adoption of the Sobolev norm allows us to achieve a faster rate of convergence of the regularised estimator under the conditions detailed in Section 4, and a superior finite-sample performance in the Monte-Carlo experiments of Section 6. Intuitively incorporating the derivative $\nabla \phi$ in the penalty helps to control tightly the oscillating components induced by ill-posedness.
3 Consistency of the TiR estimator

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 $G(\varphi)$

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

(8)

where $Q_T(\varphi)$ and $(\lambda_T)$ are as in Definition 1. This estimator exists and is measurable under weak conditions. In particular, we do not need compactness of $\Theta$ (see Appendix 2.1). Note that the estimator $\hat{\varphi}$ exists by construction in the linear case, since it can be explicitly computed (see Section 5).

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

Theorem 1: Let

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

(ii) $\varphi_0 \in \Theta$;
(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 $\lim_{\lambda \to 0} \lambda^{-a} C_\varepsilon (\lambda) > 0$ and $T^a \tilde{\delta}_T = O_p (1)$, for any $\varepsilon > 0$.

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

$$(\lambda_T T)^{-1} \to 0, \quad P$-a.s., \tag{9}$$

the estimator $\hat{\varphi}$ defined in (8) is consistent, namely $\| \hat{\varphi} - \varphi_0 \| \overset{p}{\to} 0$.

**Proof:** See Appendix 2.

If $G = 0$, Theorem 1 corresponds to the standard result of consistency for extremum estimators [e.g., White and Wooldridge (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 $\tilde{\delta}_T \overset{p}{\to} 0$. In Appendix 2, we provide technical regularity conditions to quantify this bound and to verify Conditions...
(i), (ii), and (iv) of Theorem 1 for the TiR estimator. 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. Thus, this result applies to estimators with data-driven selection of the regularisation parameter. Condition (9) on \(\lambda_T\) for consistency requires that \(\lambda_T\) converges a.s. to zero at a rate smaller than \(1/T\). Finally, we remark that Theorem 1 applies also in the case where estimator \(\hat{\varphi}\) is defined by \(\hat{\varphi} = \arg \inf_{\varphi \in \Theta_T} Q_T (\varphi) + \lambda_T G(\varphi)\) and \((\Theta_T)\) is an increasing sequence of subsets of \(\Theta\) (Sieve). Then, Theorem 1 remains true if we define \(\delta_T := \sup_{\varphi \in \Theta_T} |Q_T (\varphi) - Q_\infty (\varphi)|\), and we assume that \(\bigcup_{T=1}^\infty \Theta_T\) is dense in \(\Theta\) and that \(a > 0\) in Condition (iv) is such that \(T^a \hat{\rho}_T = O(1)\) for any \(\varepsilon > 0\), where \(\hat{\rho}_T := \inf_{\varphi \in \Theta_T : \|\varphi - \varphi_0\| \leq \varepsilon} Q_\infty (\varphi) + |G(\varphi) - G(\varphi_0)|\) (see technical report).

The rest of this section will focus on the key assumption of Theorem 1, that is identification 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

\[
Q_\infty (\varphi_n) \to Q_\infty (\varphi_0) \text{ as } n \to \infty \implies G (\varphi_n) \to \infty \text{ as } n \to 0.
\]

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

**Proof:** See Appendix 2.

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 $G(\varphi) = \| \varphi \|_H^2$.

### 3.2 Penalization with Sobolev norm

When the penalty function $G(\varphi) = \| \varphi \|_H^2$ is used, Condition (10) in Proposition 2 can be 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 2 and the quadratic approximation of limit criterion $Q_\infty$ is given by $\langle \Delta \varphi, A^*A\Delta \varphi \rangle_H$.

**Assumption 3:** Let $\{ \phi_j : j \in \mathbb{N} \}$ be an orthonormal basis in $H^2[0,1]$ of eigenfunctions of operator $A^*A$ to eigenvalues $\nu_j$, ordered such that $\nu_1 \geq \nu_2 \geq \cdots$. Then, $M_n := \inf_{\varphi \in S_n : \| \varphi \| = 1} \| \varphi \|_H \rightarrow \infty$ as $n \rightarrow \infty$, where $S_n = \text{span} \{ \phi_j : j \geq n \}$.

Assumption 3 basically requires that the subspace spanned by the eigenfunctions of $A^*A$ to eigenvalues close to zero consists of highly oscillating functions. In Lemma A.1 in Appendix 2, we show that Assumptions 1-3 imply Condition (10) in Proposition 2. Then, from Theorem 1 and Proposition 2, the consistency of the TiR estimator follows.

### 4 Asymptotic distribution of the TiR estimator

In this section, we start with studying the asymptotic MISE of the TiR estimator as a function of the regularisation parameter before examining optimal rates of convergence. Then we discuss suboptimality induced by constraining the Sobolev norm. We end with
analysing the asymptotic MSE and the pointwise asymptotic distribution. All the theoretical results are stated in terms of operators $A$ and $A^*$ underlying the linearization of the moment function $m(\varphi, z)$ in Assumption 2. The proofs are however derived for the nonparametric linear IV regression setting in order to avoid the technical burden induced by the second order term $R(\varphi, z)$.

### 4.1 The Mean Integrated Square Error

In analogy with bandwidth choice in kernel estimation and as a first contribution to the literature, we derive the asymptotic expansion of the MISE with a deterministic sequence of regularisation parameters converging to zero. To simplify the exposition, we assume that an optimal weighting matrix is used.

**Assumption 4:** The asymptotic weighting matrix $\Omega_0(z)$ is $V_0[g(Y, \varphi_0(X)) \mid Z = z]^{-1}$.

The asymptotic expansion of the MISE is characterized in the next proposition.

**Proposition 3:** Under Assumptions 1-4, Assumptions B in Appendix 1, and the bandwidth conditions

$$h_T^m = o(\lambda_T b(\lambda_T)), \quad (T\lambda_T)^{-1} = o\left(h_T^{dz}\right), \quad (11)$$

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

$$E \left[ \|\hat{\varphi} - \varphi_0\|^2 \right] = \frac{1}{T} \sum_{j=1}^{\infty} \frac{\nu_j}{(\lambda_T + \nu_j)^2} \|\phi_j\|^2 + b(\lambda_T)^2 =: M_T(\lambda_T) \quad (12)$$

up to terms which are asymptotically negligible w.r.t. the RHS, where function $b(\lambda_T)$ is given.
by

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

(13)

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

\textbf{Proof:} See Appendix 3.

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 \( \varphi^* \), recall that the quadratic approximation of the limit criterion is given by \( \langle \Delta \varphi, A^*A\Delta \varphi \rangle_H \). Thus, function \( \varphi^* \) minimizes \( \langle \Delta \varphi, A^*A\Delta \varphi \rangle_H + \lambda_T \| \varphi \|_H^2 \) w.r.t. \( \varphi \in \Theta \). 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 \( (J_0 V_0^{-1} J_0)^{-1} \) 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 nonparametric setting, the inverse of operator \( A^*A \) is unbounded, and its eigenvalues \( 1/\nu_j \to \infty \) diverge. The penalty term \( \lambda_T \| \varphi \|_H^2 \) in the criterion defining the TiR estimator implies that inverse eigenvalues \( 1/\nu_j \) are

\[ \text{Since } \nu_j/(\lambda_T + \nu_j)^2 \leq \nu_j, \text{ the infinite sum converges under Assumption B.6 (i) in Appendix 1.} \]
replaced by $\nu_j/(\lambda_T + \nu_j)^2$.

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

**Assumption 5:** 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 5, the series $k_T := \sum_{j=1}^{\infty} \|\phi_j\|^2 [\nu_j/(\lambda_T + \nu_j)^2]$ 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 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 (42) in Appendix 3, first term, and the proof of Lemma A.4]. 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 $h_T^m = o(\lambda_T b(\lambda_T))$ in (11) implies that the estimation bias is asymptotically negligible compared to the regularisation bias [see Lemma A.5 in Appendix 3].

23
Finally, it is also possible to derive a similar asymptotic expansion of the MISE for the estimator \( \tilde{\phi} \) regularised by the \( L^2 \) norm. This characterisation is new in the nonparametric instrumental regression setting: \(^7\)

\[
E \left[ \| \tilde{\phi} - \varphi_0 \|^2 \right] = \frac{1}{T} \sum_{j=1}^{\infty} \frac{\tilde{\nu}_j}{(\lambda_T + \tilde{\nu}_j)^2} + \tilde{b}(\lambda_T)^2,
\]

(14)

where \( \tilde{\nu}_j \) are the eigenvalues of operator \( \tilde{A}A \), \( \tilde{A} \) denotes the adjoint of \( A \) w.r.t. the scalar products \( \langle ., . \rangle \) and \( \langle ., . \rangle_{L^2_0(F_Z)} \), and \( \tilde{b}(\lambda_T) = \left\| \left( \lambda_T + \tilde{A}A \right)^{-1} \tilde{A}A \varphi_0 - \varphi_0 \right\|. \)

\(^8\)

Let us now come back to the MISE \( M_T(\lambda_T) \) 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 traditional 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.

\(^7\) A similar formula has been derived by Carrasco and Florens (2005) for the density deconvolution problem.

\(^8\) Note that the adjoint defined by reference to the \( L^2 \) scalar product is denoted by the superscripted * in DFR or Carrasco, Florens, and Renault (2005). We stress that the adjoint \( A^* \) is here defined by reference to a Sobolev scalar product.
4.2 Optimal rates of convergence

The eigenvalues \( \nu_j \) and the \( L^2 \)-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 \( \nu_j \) share geometric decay and the norms of eigenfunctions \( \| \phi_j \| \) share hyperbolic decay. Results for the other cases are summarised at the end of the section.

**Assumption 6:** 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 \( C_1, C_2, \)

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

Assumption 6 (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 induce eigenvalues with decay of (at least) exponential type [see Theorem 15.20 in Kress (1999)]. 9 Assumption 6 (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 7:** The bias function is such that \( b(\lambda) = C_3 \lambda^\delta, \ \delta > 0, \) for \( \lambda \) close to 0, where

---

9 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 exhibit exponential decay.
$C_3$ 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 6 and 7, for some positive constants $c_1, c_2, c$ and $\tau$, we have

(i) The MISE is $M_T(\lambda) = \frac{1}{T} c_1 \frac{1}{\lambda \left[ \log \left(1/\lambda \right) \right]^\beta} + c_2 \lambda^{2\delta}$, 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 $M_T^* = cT^{-\frac{1+\beta}{1+2\delta}} (\log T)^{-\frac{2\delta}{1+2\delta}}$, up to a term which is negligible w.r.t. the RHS.

**Proof:** See Appendix 4.

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 [consistently with Condition (9)], 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 6 and 7, the bandwidth conditions (11) are fulfilled for the optimal sequence of regularisation parameters (15) if $h_T = C \cdot T^{-\eta}$, with $\frac{1}{d^2} \frac{2\delta}{1 + 2\delta} > \eta > \frac{1}{m} \frac{1 + \delta}{1 + 2\delta}$. This condition can be satisfied if $\frac{m}{d^2} > \frac{1 + \delta}{2\delta}$.

To conclude this section, we briefly discuss the optimal rate of convergence of the MISE when the eigenvalues have 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 6 (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 $\hat{A}A$ for $L^2$ regularisation, respectively. We assume $\alpha, \tilde{\alpha} > 1$, and $\alpha > \beta - 1$ to satisfy Assumption 5. 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 7, where $b(\lambda)$ is defined in (13) for the TiR estimator, and $\tilde{b}(\lambda)$ in (14) for $L^2$ regularisation, respectively.
With hyperbolic spectrum, the rate of convergence (power of $T$) of the TiR estimator includes 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 $\tilde{A}A$ instead of $A^*A$. With geometric spectrum, the TiR estimator enjoys 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 4.1.10

10 To see this, note that their Assumption A.3 implies hyperbolic decay of the eigenvalues and is consistent with $\tilde{\delta} = (2\beta_{HH} - 1)/(2\tilde{\alpha})$, where $\beta_{HH}$ is the $\beta$ coefficient of HH [see also the remark at p. 19 in DFR].
4.3 Suboptimality of bounding the Sobolev norm

The approach of NP and AC forces compactness by a direct bounding of the Sobolev norm. Unfortunately this leads to a suboptimal rate of convergence of the regularised estimator as stated in the following theorem, which is proved in the technical report.

**Proposition 5:** Let $\bar{B} \geq \|\varphi_0\|_H^2$ be a fixed constant. Let $\tilde{\varphi}$ be the estimator defined by $\tilde{\varphi} = \arg \inf_{\varphi \in \Theta} Q_T(\varphi)$ s.t. $\|\varphi\|_H^2 \leq \bar{B}$, and denote by $\tilde{\lambda}_T$ the associated stochastic Kuhn-Tucker multiplier. Suppose that:

(i) Function $b(\lambda)$ in (13) is non-decreasing, for $\lambda$ small enough, and s.t. $\lim_{\lambda \to 0} b(c\lambda)/b(\lambda) > 0$ for any $c > 0$;

(ii) The MISE $M_T(\lambda)$ of the TiR estimator in (12) is such that for any deterministic sequence $(l_T)$:

$$l_T = o(\lambda_T^*) \text{ or } \lambda_T^* = o(l_T) \implies M_T(l_T) / M_T^* \to \infty,$$

where $\lambda_T^*$ is the optimal deterministic regularisation sequence for the TiR estimator and $M_T^* = M_T(\lambda_T^*)$;

(iii) $P\left(c_1 l_T \leq \tilde{\lambda}_T \leq c_2 l_T \right) \to 1$, for two constants $c_1 \leq c_2$ and a deterministic sequence $l_T$ such that either $l_T = o(\lambda_T^*)$ or $\lambda_T^* = o(l_T)$.

Then:

$$E \left[ \|\tilde{\varphi} - \varphi_0\|^2 \right] / M_T^* \to \infty.$$

This proposition states that, whenever the stochastic regularisation parameter $\tilde{\lambda}_T$ implied by the bound $\bar{B}$ does not exhibit the same rate of convergence as the optimal deterministic
TiR sequence $\lambda_T$, the regularised estimator with fixed bound on the Sobolev norm has a slower rate of convergence than the optimal TiR estimator. Intuitively, this will generally be the case, since imposing a fixed bound $\bar{B}$ on the Sobolev norm offers no guarantee to select an optimal rate for $\lambda_T$. Conditions (i) and (ii) of Proposition 5 are satisfied under Assumptions 6 and 7 [geometric spectrum; see also Proposition 4 (i)]. In the technical report, we prove that also Condition (iii) of Proposition 5 is satisfied in such a setting, and we characterise the degree of sub-optimality of regularised estimators with fixed bound on the Sobolev norm.

4.4 Mean Squared Error and pointwise asymptotic normality

The asymptotic MSE can be computed along the same lines as the asymptotic MISE in Proposition 3. Therefore we only state the result without proof. It is immediately seen that the integral of the MSE below over the support $\mathcal{X} = [0, 1]$ gives the MISE in (12).

**Proposition 6:** Under Assumptions 1-4, Assumption B in Appendix 1, and the bandwidth conditions $h_T^{m} = o\left(\lambda_T b(\lambda_T)\right)$, $(T\lambda_T)^{-1} = o\left(h_T^{dz}\right)$, the MSE of the TiR estimator $\hat{\varphi}$ with deterministic sequence $(\lambda_T)$ is given by

$$E \left[\hat{\varphi}(x) - \varphi_0(x)\right]^2 = \frac{1}{T} \sum_{j=1}^{\infty} \frac{\nu_j}{(\lambda_T + \nu_j)^2} \phi_j^2(x) + B_T(x)^2 =: \frac{1}{T} \sigma_T^2(x) + B_T(x)^2,$$

up to terms which are asymptotically negligible w.r.t. the RHS, where the bias term is

$$B_T(x) = (\lambda_T + A^*A)^{-1} A^* A \varphi_0(x) - \varphi_0(x).$$

A couple of remarks are in order. (i) In a given point $x \in \mathcal{X}$, the rate of convergence of the MSE depends on the decay behaviour of eigenvalues $\nu_j$ and eigenfunctions $\phi_j(x)$. It
is possible to follow an analysis similar to Sections 4.1 and 4.2 to derive the locally optimal sequence of regularisation parameters and the associated optimal MSE. (ii) The asymptotic variance \( \sigma_T^2(x)/T \) of \( \hat{\varphi}(x) \) depends on \( x \in \mathcal{X} \) through the eigenfunctions \( \phi_j \), whereas the asymptotic bias of \( \hat{\varphi}(x) \) as a function of \( x \in \mathcal{X} \) is given by \( B_T(x) \). This implies that not only the scale but also the rate of convergence of the MSE may differ across the points of the support \( \mathcal{X} \). This feature is not shared by standard nonparametric estimators, and is peculiar of the ill-posed setting.

Finally, under a regularisation with an \( L^2 \) norm, we get

\[
E[\tilde{\varphi}(x) - \varphi_0(x)]^2 = \frac{1}{T} \sum_{j=1}^{\infty} \frac{\tilde{\nu}_j}{(\lambda_T + \tilde{\nu}_j)^2} \tilde{\phi}_j^2(x) + B_T(x)^2, \tag{18}
\]

where \( B_T(x) = (\lambda_T + \tilde{A}A)^{-1} \tilde{A}A\varphi_0(x) - \varphi_0(x) \) and \( \tilde{\phi}_j \) denotes an orthonormal basis in \( L^2[0,1] \) of eigenvectors of \( \tilde{A}A \) to eigenvalues \( \tilde{\nu}_j \).

The pointwise asymptotic normality of the TiR estimator is stated in the next proposition.

**Proposition 7:** Suppose the Assumptions of Proposition 6 hold, and that for a strictly positive sequence \( (a_j) \) such that \( \sum_{j=1}^{\infty} 1/a_j < \infty \), we have

\[
\frac{\sum_{j=1}^{\infty} \frac{\nu_j}{(\lambda_T + \nu_j)^2} \phi_j^2(x) \|g_j\|_3^2 a_j}{\sum_{j=1}^{\infty} \frac{\nu_j}{(\lambda_T + \nu_j)^2} \phi_j^2(x)} = o\left(T^{-1/3}\right), \tag{19}
\]

where \( \|g_j\|_3 := E\left[g_j (Y, X, Z)^3\right]^{1/3} \) and \( g_j(y, x, z) = (A\phi_j)(z) g(y, \varphi_0(x))/\sqrt{\nu_j} \). Then the
**TiR estimator is asymptotically normal:**

\[
\sqrt{T/\sigma_T^2(x)} (\hat{\varphi}(x) - \varphi_0(x) - \mathcal{B}_T(x)) \overset{d}{\rightarrow} N(0,1).
\]

**Proof:** See Appendix 5.

Condition (19) is used to apply a Lyapunov CLT. In general, it is satisfied when \( \lambda_T \) converges to zero not too fastly. Under Assumption A.6 (i) of geometric spectrum for the eigenvalues \( \nu_j \), and an assumption of hyperbolic decay for the eigenvectors \( \phi_j^2(x) \) and \( \|g_j\|_3 \), Lemma A.6 in Appendix 4 implies that (19) is satisfied whenever \( \lambda_T \geq cT^{-\gamma} \) for some \( c, \gamma > 0 \).

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 nonparametric IV estimation of a single equation model, with \( g(y, \varphi(x)) = \varphi(x) - y \), and conditional moment as in (3). Then, the estimated moment function is given by

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

To simplify the exposition, we assume that \( \Omega_0(z) = V_0[Y - \varphi_0(X) \mid Z = z]^{-1} = 1 \) in Assumption 4. The objective function of the TiR estimator in Definition 1 can be rewritten as [see Appendix 3.1]

\[
Q_T(\varphi) + \lambda_T \|\varphi\|_H^2 = \langle \varphi, \hat{A}^* \hat{A}\varphi \rangle_H - 2\langle \varphi, \hat{A}^* \hat{r} \rangle_H + \lambda_T \langle \varphi, \varphi \rangle_H, \quad \varphi \in H^2[0,1], \tag{20}
\]
up to a term independent of $\varphi$, where $\hat{A}^*$ denotes the linear operator defined on $L^2_{\hat{f}_{0}}(F_Z)$ by

$$\langle \varphi, \hat{A}^* \psi \rangle_H = \frac{1}{T} \sum_{t=1}^{T} \left( \hat{A}_t \varphi \right)(Z_t) \left( \psi(Z_t) \right), \quad \varphi \in H^2[0,1], \quad \psi \in L^2_{\hat{f}_{0}}(F_Z). \quad (21)$$

Under the regularity conditions in Appendix 1, Criterion (20) admits a global minimum $\hat{\varphi}$ on $H^2[0,1]$, which is characterized by the first order condition

$$\left( \lambda_T + \hat{A}^* \hat{A} \right) \hat{\varphi} = \hat{A}^* \hat{r}. \quad (22)$$

This is a Fredholm integral equation of Type II $^{11}$. The transformation of the ill-posed problem (1) in the well-posed estimating equation (22) is induced by the penalty term involving the Sobolev norm. The TiR estimator is the unique solution of Equation (22) and is given by

$$\hat{\varphi} = \left( \lambda_T + \hat{A}^* \hat{A} \right)^{-1} \hat{A}^* \hat{r}. \quad (23)$$

The TiR estimator can be approximated numerically by introducing a finite-dimensional basis of functions $\{P_j : j = 1, \ldots, K\}$ in $H^2[0,1]$ and solving Equation (23) on the subspace spanned by $\{P_j : j = 1, \ldots, K\}$, which yields

$$\varphi \simeq \sum_{j=1}^{K} \theta_j P_j =: \theta^\prime P, \quad \theta \in \mathbb{R}^K. \quad (24)$$

The $K \times K$ matrix corresponding to operator $\hat{A}^* \hat{A}$ on the subspace spanned by $\{P_j\}$ is given by [using (21)]

$$\langle P_i, \hat{A}^* \hat{A} P_j \rangle_H = \frac{1}{T} \sum_{t=1}^{T} \left( \hat{A} P_i \right)(Z_t) \left( \hat{A} P_j \right)(Z_t) = \frac{1}{T} \left( \hat{P}^\prime \hat{P} \right)_{i,j}, \quad i, j = 1, \ldots, K,$$

$^{11}$ See e.g. Linton and Mammen (2005), (2006), Gagliardini and Gouriéroux (2006), and the survey by Carrasco, Florens and Renault (2005) for other examples of estimation problems leading to Type II equations.
where \( \hat{P} \) is the \( T \times K \) matrix with rows \( \hat{P}(Z_t) = \int P(x) f(w|Z_t) \, dw, \ t = 1, \ldots, T \). Matrix \( \hat{P} \) is the matrix of the "fitted values" in the regression of \( P(X) \) on \( Z \) at the sample points. Then, Equation (22) reduces to a matrix equation

\[
\mu \lambda^T D + 1^T \hat{P} \hat{P} \theta = 1^T \hat{P} \hat{R},
\]

where \( \hat{R} = (\hat{r}(Z_1), \ldots, \hat{r}(Z_T))^T \), and \( D \) is the \( K \times K \) matrix of Sobolev scalar products

\[
D_{i,j} = \langle P_i, P_j \rangle_H, \quad i, j = 1, \ldots, K.
\]

The solution is given by

\[
\hat{\theta} = \left( \lambda_T D + \frac{1}{T} \hat{P}' \hat{P} \right)^{-1} \frac{1}{T} \hat{P}' \hat{R},
\]

which yields the approximation of the TiR estimator \( \hat{\varphi} \simeq \hat{\theta}' P \). \(^{12}\)

Estimator \( \hat{\theta} \) is a 2SLS estimator with a ridge correction term. It is easy to verify that this estimator is also obtained if we replace Approximation (24) in Criterion (20) and we minimize w.r.t. \( \theta \). This latter approach has been followed by NP, AC, and Blundell, Chen and Kristensen (2004), who use Sieve estimators. However, it is important to emphasize that, the introduction of a series of basis functions as in (24) is simply a method to compute approximately the true TiR estimator \( \hat{\varphi} \) in (23), 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, recall the different role of \( \lambda_T \) in NP, AC, and Blundell, Chen and Kristensen (2004), which has been pointed out in Section 2.4.1.

Finally, we mention that a similar approach can be followed when the \( L^2 \) norm is used for regularisation, and Formula (23) is akin to the estimator of DFR and HH. The approximation with a finite-dimensional basis of functions gives an estimator \( \hat{\theta} \) similar to above, with matrix

\[^{12}\text{Note that the matrix } D \text{ is by construction positive definite, since its entries are the scalar products of linearly independent basis functions. Hence, } \lambda_T D + \frac{1}{T} \hat{P}' \hat{P} \text{ is non-singular, } P\text{-a.s.}.
\]
$D$ replaced by matrix $B$ of $L^2$ scalar products $B_{i,j} = \langle P_i, P_j \rangle$, $i, j = 1, ..., K$.  

6 A Monte-Carlo study

6.1 Data generating process

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

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

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.

---

13 DFR follow a different approach to compute exactly the estimator (see DFR, p. 46). Their method requires solving a $T \times T$ linear system of equations. For the case where $X$ and $Z$ are univariate, HH implement an estimator which uses the same basis for estimating conditional expectation $m(\varphi, z)$ and for approximating function $\varphi(x)$. 

35
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, \( x \in [0, 1] \). The chosen functions are akin to possible shapes of Engel curves, namely monotone increasing or concave.

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 \( \varphi(x) \simeq \sum_{j=0}^{5} \theta_j P_j (x) \), where \( P_0 (x) = T_0^* (x)/\sqrt{\pi} \), \( P_j (x) = T_j^* (x)/\sqrt{\pi/2} \), \( j \neq 0 \), 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 \( \| \varphi \|_H^2 = \int_0^1 \varphi^2 + \int_0^1 (\nabla \varphi)^2 \) is approximated by

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

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:

\[
D = \begin{pmatrix}
\frac{1}{\pi} & 0 & -\sqrt{2} & 0 & -\sqrt{2} & 0 \\
\vdots & \frac{26}{3\pi} & 0 & \frac{38}{5\pi} & 0 & \frac{166}{21\pi} \\
& \frac{218}{3\pi} & 0 & \frac{1182}{35\pi} & 0 \\
& \frac{3898}{35\pi} & 0 & \frac{5090}{63\pi} \\
& \vdots & \frac{67894}{315\pi} & 0 \\
\ldots & \ldots & \ldots & \ldots & \ldots & \frac{82802}{231\pi}
\end{pmatrix}.
\]

The \(L_2\) norm \(||\varphi||^2\) can be approximated in a similar way with \(\theta' B \theta\) where

\[
B = \begin{pmatrix}
\frac{1}{\pi} & 0 & -\sqrt{2} & 0 & -\sqrt{2} & 0 \\
\vdots & \frac{2}{3\pi} & 0 & \frac{2}{5\pi} & 0 & \frac{2}{21\pi} \\
& \frac{14}{15\pi} & 0 & \frac{38}{105\pi} & 0 \\
& \frac{34}{35\pi} & 0 & \frac{52}{63\pi} \\
& \vdots & \frac{62}{63\pi} & 0 \\
\ldots & \ldots & \ldots & \ldots & \ldots & \frac{98}{99\pi}
\end{pmatrix}.
\]

Such simple and exact forms ease implementation\(^\text{14}\) , and improve on speed. They also contribute to the numerical stability of the estimation procedure because of their convexity in \(\theta\) (quadratic penalty).

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

\(^{14}\) The Gauss programs developed for this section and the empirical application are available on request from the authors.
\[ \theta' \hat{P}(z) - \hat{r}(z) \]

where

\[ \hat{P}(z) \approx \frac{\sum_{t=1}^{T} P(X_t) K \left( \frac{Z_t - z}{h} \right)}{\sum_{t=1}^{T} K \left( \frac{Z_t - z}{h} \right)}, \]

\[ \hat{r}(z) \approx \frac{\sum_{t=1}^{T} Y_t K \left( \frac{Z_t - z}{h} \right)}{\sum_{t=1}^{T} K \left( \frac{Z_t - z}{h} \right)}, \]

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 observed \( Z_t \). \(^{15}\) Observe that an advantage of directly smoothing observed \( Z_t \) is absence of a boundary bias if the support of the instrument is unbounded.

The weighting function \( \Omega_0(z) \) is taken equal to unity, satisfying Assumption 4.

### 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

\(^{15}\) This choice is motivated by ease of implementation. Moderate deviations from this simple rule do not seem to affect estimation results significantly.
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 ($\rho = 0$).

Several 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 the $L_2$ 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 $2.5 = 1000/400$ 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 6 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 $\log \lambda^*_T = \log c - \gamma \log T$, 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 $\hat{c} = .012$, $\hat{\gamma} = .428$ 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, com-
paring 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 induces a more convex ISB curve. This implies $\delta > \tilde{\delta}$ and thus the 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. $^{16}$ 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:

$^{16}$ A similar approach has been successfully applied in Carrasco and Florens (2005) for density deconvolution.
Algorithm

(i) Perform the spectral decomposition of the matrix $D^{-1}\hat{P}^{T}/T$ to get eigenvalues $\hat{\nu}_j$ and

eigenvectors $\hat{w}_j$, normalized to $\hat{w}_j' D \hat{w}_j = 1$, $j = 1, ..., K$.

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

(iii) Estimate the MISE:

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

$$
+ \bar{\theta} \left[ \frac{1}{T} \hat{P}^{T} \left( \lambda D + \frac{1}{T} \hat{P}^{T} \hat{P} \right)^{-1} - I \right] B \left[ \frac{1}{T} \hat{P}^{T} \left( \lambda D + \frac{1}{T} \hat{P}^{T} \hat{P} \right)^{-1} - I \right] \bar{\theta},
$$

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 6 and 7, 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 $\bar{\lambda} = \{0.0005, 0.001\}$ 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 $0.0028$ ($0.0014$, $0.0020$, $0.0033$) when $\bar{\lambda} = 0.0005$, and $0.0027$ ($0.0007$, $0.0014$, $0.0029$) when
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 (quartiles) of .6081 (.4908, .6134, .6979), when $\bar{\lambda} = .0005$, and .7224 (.5171, .6517, .7277), when $\bar{\lambda} = .0001$. In Case 2, we get an
average (quartiles) of 5597 (.4918, .5333, .5962), when $\bar{\lambda} = .0005$, and .5764 (.4946, .5416, .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 bandwidth selection in semiparametric estimation, we believe that they could be extended to our case as well. Note that we have shown in Proposition 7 that a limit distribution exists which is a prerequisite for applying subsampling. Recognizing that $\lambda_T = cT^{-\gamma}$ for the optimal bandwidth, we propose to choose $c$ and $\gamma$ which minimize the following estimator of the MISE:

$$M(c, \gamma) = \frac{1}{IJ} \sum_{i,j} \int_{0}^{1} (\hat{\varphi}_{i,j}(x; c, \gamma) - \bar{\varphi}(x))^2 dx,$$

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 our small scale study we have chosen 500 subsamples ($J = 500$) for each subsample size $m_i \in \{50, 60, 70, ..., 100\}$ ($I = 6$), $\bar{\lambda} = \{.0005, .0001\}$, 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 $\{.005, .01, .015\} \times \{.35, .4, .45\}$. 17 Note that the two grids yield a similar

\[17\] A full scale Monte Carlo study based on large $J$ and $I$ and a fine grid for $(c, \gamma)$ is computationally too demanding because of the resampling nature of the selection procedure.
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 $\bar{\lambda} = 0.0005$, and 87% when $\bar{\lambda} = 0.0001$. In Case 2 we have found 77% when $\bar{\lambda} = 0.0005$, and 82% when $\bar{\lambda} = 0.0001$. These frequencies are scattered among the grid values.

7 An empirical example

This section presents an empirical example with the data of Horowitz (2006). We estimate an Engel curve based on the moment condition $E_0 [Y - \varphi_0(X) | Z] = 0$, with $X = \Phi(X^*)$. Variable $Y$ denotes the food expenditure share, $X^*$ denotes the standardized logarithm of total expenditures, and $Z$ denotes the standardized logarithm of annual income from wages and salaries. We have 785 household-level observations from the 1996 US Consumer Expenditure Survey. The estimation procedure is the same as the one described in the Monte-Carlo study. We need however to estimate the optimal weighting matrix since $\Omega_0(z) = V_0 [Y - \varphi_0(X) | Z = z]^{-1}$ is doubtfully constant in the application. A pilot regularisation parameter $\bar{\lambda} = .0001$ is used to get a first step estimator of $\varphi_0$. The kernel estimator $\hat{\sigma}^2(Z_t)$ of the conditional variance $\sigma^2(Z_t) = \Omega_0(Z_t)^{-1}$ at observed sample points is of the same type as for the conditional moment restriction. Then the estimation procedure and data driven selection procedure remain analogous after substituting $\hat{P}(Z_t)/\hat{\sigma}(Z_t)$,

---

18 We would like to thank Joel Horowitz for kindly providing the dataset.
\( \hat{r}(Z_t)/\hat{\sigma}(Z_t) \) for \( \hat{P}(Z_t) \), \( \hat{r}(Z_t) \) in the lines of the previous section. The subsampling relies on
1000 subsamples \( (J = 1000) \) for each subsample size \( m_i \in \{50, 53, \ldots, 197, 200\} \) \( (I = 51) \), and the extended grid \( \{0.005, .01, .05, .1, .25, .5, 1, 2, 3, 4, 5, 6\} \times \{.3, .35, \ldots, .85, .9\} \) for \( (c, \gamma) \).

We obtain a selected value of \( \hat{\lambda} = .01113 \) with the spectral based approach, together
with regression estimates \( \hat{\alpha} = 2.05176, \hat{\beta} = 3.31044, \hat{\gamma} = .90889, \hat{\delta} = .05012 \). We obtain
a selected value of \( \hat{\lambda} = .01240 \) with the subsampling procedure, which corresponds to the
selected pair \( (5.9) \) for \( (c, \gamma) \). Figure 13 plots the estimated functions \( \hat{\varphi}(x) \) for \( x \in \mathcal{X} \), and
\( \hat{\varphi}(\Phi(x^*)) \) for \( x^* \in \mathbb{R} \), using \( \hat{\lambda} = .01113 \). The plotted shape corroborates the findings of
Horowitz (2006), who rejects a linear curve but not a quadratic curve at the 5% significance
level to explain \( \ln Y \). Banks, Blundell and Lewbel (1997) have considered demand systems
that accommodate such empirical Engel curves.

8 Concluding remarks

We have studied a new estimator of a functional parameter identified by conditional moment
restrictions. The estimator exploits a Tikhonov regularisation scheme to solve ill-posedness,
and is referred to as the TiR estimator. Our framework proves to be numerically tractable,
well-behaved in finite samples, and amenable to in-depth asymptotic analysis. Numerical
tractability and good finite sample properties are key advantages for finding a route to-
wards numerous empirical applications. Our theoretical analysis paves the way to further
extensions along the lines of asymptotics for data driven estimation, estimation of functional
derivatives, estimation of semiparametric models, etc.
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 $T = 1000$. 

Figure 11: The eigenvalues (left Panel) and the $L^2$-norms of the corresponding eigenfunctions (right Panel) of operator $A^*A$ using the approximation with six polynomials.

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 $\rho = 0.5$. 
Figure 13: Estimated Engel curves for 785 household-level observations from the 1996 US Consumer Expenditure Survey. In the right Panel, food expenditure share $Y$ is plotted as a function of the standardized logarithm $X^*$ of total expenditures. In the left Panel, $Y$ is plotted as a function of transformed variable $X = \Phi(X^*)$ with support $[0, 1]$, where $\Phi$ is the cdf of the standard normal distribution. Instrument $Z$ is standardized logarithm of annual income from wages and salaries.
Appendix 1

List of regularity conditions

B.1: \( \{ R_t = (Y_t, X_t, Z_t) : t = 1, \ldots, T \} \) is an i.i.d. sample from a distribution admitting a density \( f \) with support \( S \subset \mathbb{R}^d \).

B.2: The density \( f \) is in class \( C^m(S) \), with \( m \geq 2 \).

B.3: The kernel \( K \) is a Parzen-Rosenblatt kernel of order \( m \) on \( \mathbb{R}^d \), that is (i) \( \int K(u)du = 1 \), and \( K \) is bounded; (ii) \( \int u^\alpha K(u)du = 0 \) for any multi-index \( \alpha \in \mathbb{N}^d \) with \( |\alpha| < m \), and \( \int \|u\|^m |K(u)| du < \infty \).

B.4: The kernel \( K \) is such that \( Z |K(z)|^q(u)du < \infty \) where \( q(u) = \int |K(u + z)||z|^2 dz \).

B.5: The density \( f \) is such that there exist a constant \( h > 0 \), and a function \( \omega \in L^2(F) \) satisfying

\[
\sup_{t \leq h} \int |K(z)| \left| \frac{f(y + tz) - f(y)}{f(y)} \right| dz \leq h^2 \omega^4(y),
\]

\[
\sup_{t \leq h} \int |\tilde{K}(z)| \left| \frac{f(y + tz) - f(y)}{f(y)} \right| dz \leq h^2 \omega^4(y),
\]

\[
\sup_{t \leq h} \int |K(z)| \left| \frac{f(y + tz) - f(y)}{f(y)} \right|^2 dz \leq h^2 \omega^4(y),
\]

for any \( y \in S \), where \( K(z) = \int |K(u + z)K(u)| du \) and \( \tilde{K}(z) = \int |K(u + z)K(u)| q(u)du \).

B.6: The orthonormal basis of eigenvectors \( \{ \phi_j : j \in \mathbb{N} \} \) of operator \( A^*A \) with eigenvalues \( \nu_j \) satisfies (i) \( \sum_{j=1}^\infty \nu_j \|\phi_j\|^2 < \infty \); (ii) \( \sum_{j=1}^\infty \frac{\langle \phi_j, \phi_i \rangle^2}{\|\phi_j\|^2 \|\phi_i\|^2} < \infty \).
B.7: The eigenfunctions $\phi_j$ and the eigenvalues $\nu_j$ satisfy

$$\sup_{j \in \mathbb{N}} \frac{1}{\nu_j} E_0 \left[ \omega (R)^4 \left( A\phi_j \right) (Z)^2 g_0 (W)^2 \right] < \infty,$$

$$\sup_{j \in \mathbb{N}} \frac{1}{\nu_j} E_0 \left[ (1 + \omega^2 (R)) \| (\nabla A\phi_j) (Z) \|^2 g_0 (W)^2 \right] < \infty,$$

$$\sup_{j \in \mathbb{N}} \frac{1}{\nu_j} E_0 \left[ (1 + \omega^2 (R)) \left( A\phi_j \right) (Z)^2 \| (\nabla g_0) (W) \|^2 \right] < \infty,$$

where $g_0 (W) = g(Y, \varphi_0 (X))$ and $\omega$ is as in Assumption B.5.

B.8: There exist a constant $h > 0$ and a function $\omega_m \in L^2 (F)$ such that

$$\sup_{\alpha \in \mathbb{N}^m: |\alpha| = m} \sup_{t \leq h} \int |K(u)| \left| \frac{D^\alpha f(y + tu)}{f(y)} \right| \|u\|^m du \leq \omega_m (y),$$

for any $y \in S$. 
Appendix 2

Consistency of the TiR estimator

In this Appendix we prove the consistency of penalized extremum estimators

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

(25)

This covers the special case of the TiR estimator in Definition 1, where \( G(\varphi) = \|\varphi\|_{H}^2 \).

A.2.1 Existence of the estimator

Since \( Q_T \) is positive, a function \( \hat{\varphi} \in \Theta \) is solution of optimization problem (25) if and only if it is a solution of:

\[
\hat{\varphi} = \arg\inf_{\varphi \in \Theta} Q_T(\varphi) + \lambda_T G(\varphi), \quad \text{s.t.} \quad \lambda_T G(\varphi) \leq L_T,
\]

(26)

where \( L_T := Q_T(\varphi_0) + \lambda_T G(\varphi_0) \). From Theorem 2.2 of White and Wooldridge (1991), the solution \( \hat{\varphi} \) of (26) exists if

(i) function \( Q_T : \Omega \times \Theta \to \mathbb{R} \) is Borel-measurable, where \( Q_T(\omega, \varphi) \) denotes the value of random variable \( Q_T(\varphi) \) for event \( \omega \in \Omega \), and \( (\Omega, \mathcal{F}, P) \) is a complete probability space;

(ii) mappings \( \varphi \to G(\varphi) \) and \( \varphi \to Q_T(\omega, \varphi) \) are lower semicontinuous on \( \Theta \), \( P\text{-a.s.} \), for any \( T \), w.r.t. the \( L^2 \) norm \( \| \cdot \| \);

(iii) set \( \{ \varphi \in \Theta : \lambda_T G(\varphi) \leq L_T \} \) is compact w.r.t. the \( L^2 \) norm \( \| \cdot \| \), \( P\text{-a.s.} \), for any \( T \).

A.2.2 Consistency of penalized extremum estimators
Proof of Theorem 1: For any $T$ and any given $\varepsilon > 0$, we have

$$P \left[ \| \hat{\varphi} - \varphi_0 \| > \varepsilon \right] \leq P \left[ \inf_{\varphi \in \Theta : \| \varphi - \varphi_0 \| \geq \varepsilon} Q_T (\varphi) + \lambda T G (\varphi) \leq Q_T (\varphi_0) + \lambda_T G (\varphi_0) \right].$$

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

$$\inf_{\varphi \in \Theta : \| \varphi - \varphi_0 \| \geq \varepsilon} Q_T (\varphi) + \lambda_T G (\varphi) \leq \lambda_T G (\varphi_0) + \sup_{\varphi \in \Theta} |\Delta Q_T (\varphi)| \leq \lambda_T G (\varphi_0) + \sup_{\varphi \in \Theta} |\Delta Q_T (\varphi)| = 2\delta T.$$

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

$$P \left[ \| \hat{\varphi} - \varphi_0 \| > \varepsilon \right] \leq P \left[ C_\varepsilon (\lambda_T) \leq 2\delta T \right] = P \left[ 1 \leq \frac{1}{\lambda_T^a C_\varepsilon (\lambda_T) (T\lambda_T)^a} (2T^n \delta T) \right] =: P \left[ 1 \leq Z_T \right].$$

Since $\lambda_T \to 0$ such that $(T\lambda_T)^{-1} \to 0$, $P$-a.s., for $a$ chosen as in (iv) we have $Z_T \to 0$, and we deduce $P \left[ \| \hat{\varphi} - \varphi_0 \| > \varepsilon \right] \leq P \left[ Z_T \geq 1 \right] \to 0$. 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

$$C_\varepsilon (\lambda_n) \leq 0, \quad \forall n \in \mathbb{N}. \tag{27}$$

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 (27) that there exists a sequence $(\varphi_n)$ such that $\varphi_n \in \Theta, \| \varphi_n - \varphi_0 \| \geq \varepsilon$, and

$$Q_\infty (\varphi_n) + \lambda_n G (\varphi_n) - \lambda_n G (\varphi_0) \leq \lambda_n. \tag{28}$$
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

$$G(\varphi_n) \leq G(\varphi_0) + 1. \quad (29)$$

Moreover, since $G(\varphi_n) \geq G_0$, where $G_0$ is the lower bound of function $G$, we get $Q_\infty(\varphi_n) + \lambda_n G_0 - \lambda_n G(\varphi_0) \leq \lambda_n$ from (28), that is $Q_\infty(\varphi_n) \leq \lambda_n (1 + G(\varphi_0) - G_0)$, which implies

$$\lim_{n} Q_\infty(\varphi_n) = 0 = Q_\infty(\varphi_0). \quad (30)$$

Obviously, the simultaneous holding of (29) and (30) violates Assumption (10).

A.2.3 Penalization with Sobolev norm

In this Section we check that the assumptions in A.2.1 and A.2.2 hold for the special case $G(\varphi) = \|\varphi\|^2_H$ under Assumptions 1-3.

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

ii) The set $\{ \varphi \in \Theta : \lambda_T \|\varphi\|^2_H \leq L_T \}$ is compact w.r.t. the norm $\|\cdot\|$, P-a.s., for any $T$.

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

**Lemma A.1:** Assumptions 1-3 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

$$Q_\infty(\varphi_n) \to 0 \text{ as } n \to \infty. \quad (31)$$
We have to prove
\[ \| \varphi_n \|_H \to \infty \text{ as } n \to 0. \] (32)

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
\[ \langle e_n, A^* A e_n \rangle_H = \frac{\langle \Delta \varphi_n, A^* A \Delta \varphi_n \rangle_H}{\| \varphi_n - \varphi_0 \|^2} \leq \frac{1}{\varepsilon^2} \| A \Delta \varphi_n \|_{L_{0,0}^2(F_Z)}^2. \] Moreover, we have:
\[
Q_\infty (\varphi_n) = \| A \Delta \varphi_n \|_{L_{0,0}^2(F_Z)}^2 + 2 \| A \Delta \varphi_n, R (\varphi_n, \cdot) \|_{L_{0,0}^2(F_Z)}^2 + \| R (\varphi_n, \cdot) \|_{L_{0,0}^2(F_Z)}^2 \\
\geq \| A \Delta \varphi_n \|_{L_{0,0}^2(F_Z)}^2 - 2 \| A \Delta \varphi_n \|_{L_{0,0}^2(F_Z)} \| R (\varphi_n, \cdot) \|_{L_{0,0}^2(F_Z)} \| R (\varphi_n, \cdot) \|_{L_{0,0}^2(F_Z)}^2 ,
\]
which implies from Assumption 2 (ii):
\[
\frac{Q_\infty (\varphi_n)}{\| A \Delta \varphi_n \|_{L_{0,0}^2(F_Z)}^2} \geq \left( 1 - \frac{\| R (\varphi_n, \cdot) \|_{L_{0,0}^2(F_Z)}^2}{\| A \Delta \varphi_n \|_{L_{0,0}^2(F_Z)}^2} \right)^2 \geq \left( 1 - \sup_{\varphi \in \mathcal{B}} \frac{\| R (\varphi, \cdot) \|_{L_{0,0}^2(F_Z)}^2}{\| A \Delta \varphi \|_{L_{0,0}^2(F_Z)}^2} \right)^2 =: c_2 > 0 .
\]
Thus, \( \langle e_n, A^* A e_n \rangle_H \leq \frac{1}{c_2 \varepsilon^2} Q_\infty (\varphi_n) \to 0 \) as \( n \to \infty \), from (31). Let \( \Pi (N) \) denote the orthogonal projection [w.r.t. the scalar product \( \langle \cdot, \cdot \rangle_H \)] on the subspace spanned by \( \{ \phi_1, \ldots, \phi_N \} \).

Then we have for any \( N \in \mathbb{N} \)
\[
\| \Pi (N) e_n \|_H^2 = \sum_{j=1}^{N} \langle \phi_j, e_n \rangle_H^2 \leq \frac{1}{\nu_N} \sum_{j=1}^{N} \nu_j \langle \phi_j, e_n \rangle_H^2 \leq \frac{1}{\nu_N} \sum_{j=1}^{\infty} \nu_j \langle \phi_j, e_n \rangle_H^2 \\
= \frac{1}{\nu_N} \langle e_n, A^* A e_n \rangle_H \to 0, \text{ as } n \to \infty,
\]
that is \( \| \Pi (N) e_n \|_H \to 0 \) as \( n \to \infty \), for any \( N \in \mathbb{N} \).

Let us now derive a lower bound for the Sobolev norm \( \| e_n \|_H \). We have
\[
\| e_n \|_H \geq \| \Pi (N) e_n \|_H - \| \Pi (N) e_n \|_H , \quad (33)
\]

67
where $\Pi_{(N)} = 1 - \Pi_{(N)}$ denotes the orthogonal projection on $\text{span}\{\phi_j : j \geq N + 1\}$. Let us derive a bound for the first term in the RHS of (33). We have

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

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

Thus, we get from (33)

$$
\|e_n\|_H \geq M_{N+1} \left( 1 - \|\Pi_{(N)} e_n\|_H \right) - \|\Pi_{(N)} e_n\|_H ,
$$

(34)

for any $N$ and $n \in \mathbb{N}$. Since the bound (34) holds for any $N \in \mathbb{N}$, it follows

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

(35)

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 (35) diverges. To this goal, define the sequence $n(N), N = 1, 2...$ recursively by

$$
n(1) = \min \left\{ n^* \in \mathbb{N} \mid \|\Pi_{(1)} e_n\|_H \leq 1 \text{ for all } n \geq n^* \right\},
$$

$$
n(N) = \min \left\{ n^* \in \mathbb{N} \mid n^* > n(N-1) \ , N \|\Pi_{(N)} e_n\|_H \leq 1 \text{ for all } n \geq n^* \right\} , \quad N = 2, ...
$$

Since $N \|\Pi_{(N)} e_n\|_H \to 0$ 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 $n(N), N = 1, 2, ...$ 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

\[
N_n \| \Pi(N_n) e_n \|_H \leq 1, \quad (36)
\]

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

\[
\| \Pi(N_n) e_n \|_H \leq 1/2, \quad \forall n \text{ large enough}. \quad (37)
\]

Using Bounds (36) and (37) in Inequality (35), we get \(\| e_n \|_H \geq M_{N_n+1} (1/2) - 1/2 \to \infty\), as \(n \to \infty\), from Assumption 3.

Finally, we get

\[
\| \varphi_n \|_H = \| \varphi_n - \varphi_0 \|_H e_n + \varphi_0 \|_H \geq \| \varphi_n - \varphi_0 \|_H e_n - \| \varphi_0 \|_H (38)
\]

\[
\geq \varepsilon \| e_n \|_H - \| \varphi_0 \|_H \to \infty.
\]

Therefore, (32) follows, and the proof is concluded. \(\blacksquare\)
Appendix 3

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

\[ m(\varphi, z) = E_0 [\varphi(X) - Y \mid Z = z] = (A\varphi)(z) - r(z), \]

where \( (A\varphi)(z) = \int \varphi(x)f(w|z)dw \)

and \( r(z) = \int yf(w|z)dw, \) with \( \Omega_0(z) = 1. \)

A.3.1 The first-order condition

The estimated moment function is

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

The objective function of the TiR estimator becomes

\[ Q_T(\varphi) + \lambda_T \|\varphi\|_H^2 = \frac{1}{T} \sum_{t=1}^T [(\hat{A}\varphi)(Z_t) - \hat{r}(Z_t)]^2 + \lambda_T \langle \varphi, \varphi \rangle_H. \] (39)

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.2:** Under regularity conditions, the following properties hold \( P\text{-}a.s. : \)

(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

\[ \langle \varphi, \hat{A}^*\psi \rangle_H = \frac{1}{T} \sum_{t=1}^T (\hat{A}\varphi)(Z_t)\psi(Z_t), \] for any \( \psi \in L^2(F_Z) \) and \( \varphi \in H^2[0, 1]; \)

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

**Proof:** See technical report.
Then, from Lemma A.2 i)-iii), Criterion (39) can be rewritten as

\[ Q_T(\varphi) + \lambda T \| \varphi \|^2_H = \langle \varphi, (\lambda_T + \hat{A}^* \hat{A}) \varphi \rangle_H - 2 \langle \varphi, \hat{A}^* \hat{r} \rangle_H, \]  

up to a term independent of \( \varphi \). From Lemma A.2 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 (40) admits a global minimum over \( H^2[0,1] \). It is given by the first-order condition \( (\hat{A}^* \hat{A} + \lambda_T) \hat{\varphi} = \hat{A}^* \hat{r}, \) that is

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

A.3.2 Asymptotic expansion of the first-order condition

Let us now expand the estimator in (41). We can write

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

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

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

\[ =: V_T + S(\lambda_T) + R_T, \]  

where the remaining term \( R_T \) is given by

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

\[ + \left[ (\lambda_T + \hat{A}^* \hat{A})^{-1} \hat{A}^* \hat{A} - (\lambda_T + A^* A)^{-1} A^* A \right] \varphi_0. \]  

71
Lemma A.3: Assume the bandwidth conditions $h^m_T = o(\lambda_T)$, $(\lambda_T T)^{-1} = o\left(h_T^{dz}\right)$, where $m$ is the order of the kernel $K$, and $d_Z$ the dimension of $Z$. Then, under regularity assumptions, $E \left[ \|R_T\|^2 \right] = o\left( E \left[ \|V_T + S(\lambda_T)\|^2 \right] \right)$.

Proof: See technical report.

From (42) we deduce

$$E \left[ \|\hat{\varphi} - \varphi_0\|^2 \right] = E \left[ \|V_T + S(\lambda_T)\|^2 \right] + E \left[ \|R_T\|^2 \right] + 2E \left[ \langle V_T + S(\lambda_T), R_T \rangle \right] = E \left[ \|V_T + S(\lambda_T)\|^2 \right] + o\left( E \left[ \|V_T + S(\lambda_T)\|^2 \right] \right),$$

by applying twice the Cauchy-Schwarz inequality and Lemma A.3. Since

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

(44)

we get

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

(45)

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.4 and A.5. Combining these two Lemmas and the asymptotic expansion in (45) results in Proposition 3.

A.3.3 Asymptotic expansion of the variance term
Lemma A.4: Under regularity conditions, up to a term which is asymptotically negligible w.r.t. the RHS, we have
\[ E \left[ \left\| (\lambda_T + A^*A)^{-1} A^* \left( \hat{\psi} - E\hat{\psi} \right) \right\|^2 \right] = \frac{1}{T} \sum_{j=1}^{\infty} \frac{\nu_j}{(\lambda_T + \nu_j)^2} \left\| \phi_j \right\|^2. \]

**Proof:** See technical report.

A.3.4 Asymptotic expansion of the bias term

Lemma A.5: 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 technical report.
Appendix 4

Rate of convergence with geometric spectrum

In this Appendix we prove Proposition 4.

i) The next Lemma A.6 characterizes the variance term of the asymptotic expansion of the MISE in Proposition 3.

**Lemma A.6:** Let $\nu_j$ and $\|\phi_j\|^2$ satisfy Assumption 6, and define the function

$$I(\lambda) = \sum_{j=1}^{\infty} \frac{\nu_j}{(\lambda + \nu_j)^2} \|\phi_j\|^2, \quad \lambda > 0.$$  

Then,

$$\lim_{\lambda \to 0} \lambda \left[ \log (1/\lambda) \right]^{\beta} I(\lambda) = \left( \frac{1}{\alpha} \right)^{1-\beta} C_2.$$

**Proof:** See technical report.

From Lemma A.6 and using Assumption 7, we get

$$M_T(\lambda) = \frac{1}{T} C_1 \frac{1}{\lambda \left[ \log (1/\lambda) \right]^{2\beta}} + c_2 \lambda^{2\delta},$$

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

$$\frac{dM_T(\lambda)}{d\lambda} = -\frac{1}{T} C_1 \frac{1}{\lambda^2 \left[ \log (1/\lambda) \right]^{2\beta}} \left( \frac{\left[ \log (1/\lambda) \right]^{\beta} - \lambda \beta \left[ \log (1/\lambda) \right]^{\beta-1} \frac{1}{\lambda}}{\lambda} \right) + 2c_2 \delta \lambda^{2\delta-1}$$

$$= -\frac{1}{T} C_1 \frac{\log (1/\lambda) - \beta}{\lambda^2 \left[ \log (1/\lambda) \right]^{2\beta+1}} + 2c_2 \delta \lambda^{2\delta-1}. $$

Thus

$$\frac{dM_T(\lambda_T^*)}{d\lambda} = 0 \iff \frac{1}{T} C_1 \frac{\log (1/\lambda_T^*) - \beta}{2c_2 \delta \left[ \log (1/\lambda_T^*) \right]^{2\beta+1}} = (\lambda_T^*)^{2\delta+1}. \quad (46)$$

To solve the latter equation for $\lambda_T^*$, define $\tau_T := \log (1/\lambda_T^*)$. Then $\tau_T$ satisfies

$$\tau_T = c_3 + \frac{1}{1 + 2\delta} \log T + \frac{1 + \beta}{1 + 2\delta} \log \tau_T - \frac{1}{1 + 2\delta} \log (\tau_T - \beta),$$

74
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 a 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

\[
M_T(\lambda^*_T) = \frac{1}{T} c_1 \frac{1}{\lambda^*_T \log(1/\lambda^*_T)} + c_2 (\lambda^*_T)^{2\delta} = \frac{1}{T} c_1 \frac{1}{\lambda^*_T \tau_T} + c_2 (\lambda^*_T)^{2\delta}.
\]

From (46), \( \lambda^*_T = \left( \frac{c_1}{2c_2\delta} \right)^{\frac{1}{\delta+1}} T^{-\frac{1}{\delta+1}} \left( \frac{\tau_T - \beta}{\tau_T} \right)^{\frac{1}{\delta+1}} = c_5 T^{-\frac{1}{\delta+1}} \tau_T^{-\frac{\beta}{\delta+1}} \), for a constant \( c_5 \), up to a term which is negligible w.r.t. the RHS. Thus we get

\[
M_T(\lambda^*_T) = \frac{1}{T} c_1 T^{\frac{1}{\delta+1}} \frac{1}{\omega T^{\frac{1}{\delta+1} + \beta}} + c_2 c_5^{2\delta} T^{-\frac{2\delta}{\delta+1}} \tau_T^{-\frac{2\delta\beta}{\delta+1}}
\]

\[
= c_6 T^{-\frac{2\delta}{\delta+1} \tau_T^{\frac{\beta}{\delta+1}}} = c_7 T^{-\frac{2\delta}{\delta+1} (\log T)^{\frac{\beta}{\delta+1}}},
\]

for some constants \( c_6 \) and \( c_7 \), up to a term which is negligible w.r.t. the RHS.
Appendix 5

Asymptotic normality of the TiR estimator

In this Appendix we prove the asymptotic normality of the TiR estimator. From Equation (42) in Appendix 3, we have

\[
\sqrt{T/\sigma_T^2(x)} (\hat{\psi} (x) - \varphi_0 (x)) = \sqrt{T/\sigma_T^2(x)} \left( (\lambda_T + A^*A)^{-1} A^* \left( \hat{\psi} - E\hat{\psi} \right) (x) + \sqrt{T/\sigma_T^2(x)} B_T(x) \right) \\
+ \sqrt{T/\sigma_T^2(x)} (\lambda_T + A^*A)^{-1} A^* E\hat{\psi}(x) + \sqrt{T/\sigma_T^2(x)} R_T(x) \\
=: (I) + (II) + (III) + (IV),
\]

where \( R_T(x) \) is defined in (43). We now show that the term (I) is asymptotically \( N(0, 1) \) distributed and the terms (III) and (IV) are \( o_p(1) \), which implies Proposition 7.

A.5.1 Asymptotic normality of (I)

Since \( \{ \varphi_j : j \in \mathbb{N} \} \) is an orthonormal basis w.r.t. \( \langle ., . \rangle_H \), we can write:

\[
(\lambda_T + A^*A)^{-1} A^* \left( \hat{\psi} - E\hat{\psi} \right) (x) = \sum_{j=1}^{\infty} \left\langle \varphi_j, (\lambda_T + A^*A)^{-1} A^* \left( \hat{\psi} - E\hat{\psi} \right) \right\rangle_H \varphi_j (x) \\
= \sum_{j=1}^{\infty} \frac{1}{\lambda_T + \nu_j} \left\langle \varphi_j, A^* \left( \hat{\psi} - E\hat{\psi} \right) \right\rangle_H \varphi_j (x).
\]

Then, we get

\[
\sqrt{T/\sigma_T^2(x)} (\lambda_T + A^*A)^{-1} A^* \left( \hat{\psi} - E\hat{\psi} \right) (x) = \sum_{j=1}^{\infty} w_{j,T}(x) Z_{j,T},
\]

where
\[ Z_{j,T} := \frac{1}{\sqrt{\nu_j}} \langle \phi_j, \sqrt{T} A^* \left( \hat{\psi} - E \hat{\psi} \right) \rangle_H, \quad j = 1, 2, \ldots, \]

and

\[ w_{j,T}(x) := \frac{\sqrt{\nu_j} \phi_j(x)}{\lambda_T + \nu_j \left( \sum_{j=1}^{\infty} \frac{\nu_j}{(\lambda_T + \nu_j)^2} \phi_j(x)^2 \right)^{1/2}}, \quad j = 1, 2, \ldots. \]

Note that \( \sum_{j=1}^{\infty} w_{j,T}(x)^2 = 1 \). We need the following two Lemmas A.7 and A.8, which are proved in the technical report.

**Lemma A.7:** Under regularity conditions:

\[ \sum_{j=1}^{\infty} w_{j,T}(x) Z_{j,T} = \sqrt{T} \int G_T(r) \left[ \hat{f}(r) - E \hat{f}(r) \right] dr + o_p(1), \quad (47) \]

where \( r = (w, z) \), \( G_T(r) := \sum_{j=1}^{\infty} w_{j,T}(x) g_j(r) \) and \( g_j(r) = (A\phi_j)(z) g_0(w) / \sqrt{\nu_j} \).

**Lemma A.8:** Under regularity conditions:

\[ \sqrt{T} \int G_T(r) \left[ \hat{f}(r) - E \hat{f}(r) \right] dr = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} Y_{tT} + o_p(1), \]

where \( Y_{tT} := G_T(R_t) = \sum_{j=1}^{\infty} w_{j,T}(x) g_j(R_t) \).

From Lemma A.7 and A.8, it is sufficient to prove that \( T^{-1/2} \sum_{t=1}^{T} Y_{tT} \) is asymptotically \( N(0, 1) \) distributed. Note that

\[ E \left[ g_j(R) \right] = \frac{1}{\sqrt{\nu_j}} \sum_{j=1}^{\infty} E \left[ (A\phi_j)(Z) E \left[ g_0(W) | Z \right] \right] = 0, \]

77
and
\[
\text{Cov}[g_j(R), g_l(R)] = E \left[ \frac{1}{\sqrt{\nu_j \nu_l}} (A\phi_j)(Z) (A\phi_l)(Z) \right] E \left[ g_0(W)^2 \right] = \frac{1}{\sqrt{\nu_j \nu_l}} \langle \phi_j, A^* A\phi_l \rangle_H = \delta_{j,l}.
\]

Thus \( E(Y_{iT}) = 0 \) and
\[
V[Y_{iT}] = \sum_{j,l=1}^{\infty} w_{j,T}(x) w_{l,T}(x) \text{Cov}[g_j(R), g_l(R)] = \sum_{j=1}^{\infty} w_{j,T}(x)^2 = 1.
\]

From application of a Lyapunov CLT, it is sufficient to show that
\[
\frac{1}{T^{1/2}} E \left[ |Y_{iT}|^3 \right] \to 0, \quad T \to \infty.
\]

To this goal, using \(|Y_{iT}| \leq \sum_{j=1}^{\infty} |w_{j,T}(x)||g_j(R_i)|\) and the triangular inequality, we get
\[
\frac{1}{T^{1/2}} E \left[ |Y_{iT}|^3 \right] \leq \frac{1}{T^{1/2}} E \left[ \left( \sum_{j=1}^{\infty} |w_{j,T}(x)||g_j| \right)^3 \right] = \frac{1}{T^{1/2}} \left( \sum_{j=1}^{\infty} \nu_j \right)^{3/2} \left( \sum_{j=1}^{\infty} \frac{\nu_j}{\lambda_T + \nu_j} \phi_j(x) \|g_j\|_3 \right)^{3/2} \left( \sum_{j=1}^{\infty} \frac{1}{a_j} \right)^{1/2}.
\]

Moreover, from the Cauchy-Schwarz inequality we have
\[
\sum_{j=1}^{\infty} \frac{\sqrt{\nu_j}}{\lambda_T + \nu_j} \phi_j(x) \|g_j\|_3 \leq \left( \sum_{j=1}^{\infty} \frac{\nu_j}{(\lambda_T + \nu_j)^2} \phi_j(x)^2 \|g_j\|_3^2 a_j \right)^{1/2} \left( \sum_{j=1}^{\infty} \frac{1}{a_j} \right)^{1/2},
\]
and \( \sum_{j=1}^{\infty} a_j^{-1} < \infty, a_j > 0 \). Thus, we get:
\[
\frac{1}{T^{1/2}} E \left[ |Y_{iT}|^3 \right] \leq \left( \sum_{j=1}^{\infty} \frac{1}{a_j} \right)^{3/2} \left( \frac{1}{T^{1/3}} \sum_{j=1}^{\infty} \frac{\nu_j}{(\lambda_T + \nu_j)^2} \phi_j(x)^2 \|g_j\|_3^2 a_j \right)^{3/2} \left( \sum_{j=1}^{\infty} \frac{\nu_j}{(\lambda_T + \nu_j)^2} \phi_j(x)^2 \right)^{-3/2}.
\]
and Condition (48) is implied by Condition (19).

**A.5.2 Terms (III) and (IV) are $o_p(1)$**

**Lemma A.9:** Under regularity conditions:

$$\sqrt{T/\sigma^2_T(x)}(\lambda_T + A^* A)^{-1} A^* E\hat{\psi}(x) = o_p(1).$$

**Proof:** See technical report.

**Lemma A.10:** Under regularity conditions:

$$\sqrt{T/\sigma^2_T(x)}R_T(x) = o_p(1).$$

**Proof:** See technical report.