Tikhonov Regularisation for Functional Minimum Distance Estimators

Patrick Gagliardini
University of Lugano and Swiss Finance Institute

Olivier Scaillet
HEC Genève and Swiss Finance Institute
Outline of the talk

(1) Introduction

(2) The Tikhonov Regularised (TiR) estimator

(3) Asymptotic and finite-sample properties of the TiR estimator

(4) Empirical application to non-parametric estimation of an Engel curve
**Parametric IV estimation**

- Minimum Distance estimators are used to exploit (conditional) moment restrictions

- An example is parametric Instrumental Variable (IV) estimation

\[ Y = X'\beta_0 + U \quad , \quad E[XU] \neq 0 \]

Instrument \( Z \) satisfying \( E[ZU] = 0 \) implies the moment restriction

\[ E[Z(Y - X'\beta_0)] = 0 \]

The IV (2SLS) estimator is:

\[ \hat{\beta} = \arg \min_{\beta} \hat{m}(\beta)' \hat{\Omega} \hat{m}(\beta) = \left(X'Z \left(Z'Z\right)^{-1}Z'X\right)^{-1}X'Z \left(Z'Z\right)^{-1}Z'Y \]

where \( \hat{m}(\beta) = \frac{1}{T} \sum_{t=1}^{T} Z_t(Y_t - X_t'\beta) \) and \( \hat{\Omega} = \left(Z'Z\right)^{-1} \)

- Extension to non-linear parametric moment restrictions is GMM
Non-parametric IV estimation: Two examples

- The data generating process is
  \[
  \begin{pmatrix}
  U \\
  V \\
  Z
  \end{pmatrix}
  \sim IIN
  \begin{pmatrix}
    (0) \\
    (0) \\
    (0)
  \end{pmatrix},
  \begin{pmatrix}
    1 & \rho & 0 \\
    \rho & 1 & 0 \\
    0 & 0 & 1
  \end{pmatrix},
  \quad \rho \in \{0, 0.5\}
  \]

  Build \( X^* = Z + V \) and map \( X^* \) into a variable \( X = \Phi(X^*) \) in \([0, 1]\)

  Case 1: \( Y = B(X) + U \), \( B \) is the cdf of Beta(2, 5) distribution
  
  Case 2: \( Y = \sin(\pi X) + U \)

- When \( \rho \neq 0 \) regressor \( X \) is endogenous!

- The conditional moment restriction:
  \[
  E [Y - \varphi_0(X) \mid Z] = 0
  \]
  where \( \varphi_0(x) = B(x) \) in Case 1 and \( \varphi_0(x) = \sin(\pi x) \) in Case 2, \( x \in [0, 1] \)

- How to estimate functional parameter \( \varphi_0 \) by exploiting (1)?
Ill-posedness

The main mathematical difficulty in non-parametric IV estimation is

**ILL-POSEDNESS**

**Intuition:**

- The conditional moment restriction $E [Y - \varphi_0(X) \mid Z] = 0$ is a linear integral equation in function $\varphi_0$:

  \[
  \int f(x \mid z) \varphi_0(x) dx = \int y f(y \mid z) dy , \quad z \in Z \tag{2}
  \]

- There exist "large" highly oscillating deviations $\Delta \varphi = \varphi - \varphi_0$ which are hard to detect since $\int f(x \mid z) \Delta \varphi(x) dx \sim 0$

- Small errors in the estimation of the RHS of (2) may imply large errors in the estimation of $\varphi_0$

  $\Rightarrow$ "Naive" estimators are inconsistent!
Review of the literature: Estimation methodology

- Newey and Powell (NP, 2003), Ai and Chen (AC, 2003)
  
  Propose a consistent minimum distance estimator which is the non-parametric analog of 2SLS

  Regularisation of ill-posedness by introducing a bound on the norm of \( \varphi \) and of \( \nabla \varphi \) to force compactness of parameter space

- Darolles, Florens and Renault (DFR, 2003), Hall and Horowitz (HH, 2005)
  
  Solve the empirical analog of the linear integral equation implied by the conditional moment restriction

  Regularisation technique resulting in a kind of ridge regression

- Florens (2003), Blundell and Powell (2003), Carrasco, Florens and Renault (2005), Horowitz (2005) present further background
Review of the literature: Applications


- Asset pricing models with functional specification of preferences: Chen and Ludvigson (2004)


Aims and contributions of the paper (I)

(1) To introduce a new minimum distance estimator for a functional parameter identified by a conditional moment restriction

\[ \hat{\varphi} = \arg \min_{\varphi} Q_T(\varphi) + \lambda_T G(\varphi) \]

Minimum distance \( \min distance \) + Penalty \( \uparrow \) penalty

Penalty function \( G(\varphi) \) involves \( L^2 \) norm of \( \varphi \) and \( \nabla \varphi \) (Sobolev norm)

\( \lambda_T \geq 0 \) tunes the amount of regularisation

**Basic intuition**: Penalty term damps out highly oscillating components of \( \hat{\varphi} \) otherwise enhanced by ill-posedness [Tikhonov (1963)]

**Appealing features:**

- Applies to linear and non-linear conditional moment restrictions
- Regularisation parameter \( \lambda_T \) is allowed to be data-dependent
- May feature faster rate of convergence than existing estimators
- Closed form in linear case
Aims and contributions of the paper (II)

(2) To study the asymptotic properties of our estimator

We provide

- Consistency with data-dependent regularisation parameter $\lambda_T$
- Asymptotic expansion of the MISE
  $\Rightarrow$ data-driven selection of the regularisation parameter $\lambda_T$
- Optimal rates of convergence
- MSE and pointwise asymptotic normality

(3) To investigate the attractiveness from an applied point of view

Our estimator benefits from:

- Numerical tractability (unconstrained optimization, quadratic penalty)
- Good finite-sample performance
- Reliable data-driven selection procedure for regularisation parameter
Nonparametric minimum distance estimators

- **Parameter of interest**: a function \( \varphi_0 \) defined on \( \mathcal{X} = [0, 1] \) satisfying the conditional moment restriction

\[
E [Y - \varphi_0(X) | Z] = 0
\]  
(3)

- **Minimum distance approach**: Function \( \varphi_0 \) minimizes

\[
Q_\infty(\varphi) = E \left[ m(\varphi, Z)' \Omega_0(Z)m(\varphi, Z) \right]
\]

where \( m(\varphi, z) = E [Y - \varphi(X) | Z = z] \) and \( \Omega_0(z) \) is a p.d. matrix

\( \Rightarrow \) Estimate \( \varphi_0 \) by minimizing empirical analog of \( Q_\infty(\varphi) \)

- **Ill-posedness**: (3) is an integral equation

\[
\int f(x|z)\varphi_0(x)dx = \int y f(y|z)dy
\]

\( \downarrow \)

\[
(A\varphi_0)(z) \quad r(z)
\]

\( \Rightarrow \) \( Q_\infty(\varphi) \) is flat along some directions!
Tikhonov Regularised (TiR) estimator (I)

- **Empirical minimum distance criterion:**

\[
Q_T(\varphi) = \frac{1}{T} \sum_{t=1}^{T} \tilde{m}(\varphi, Z_t) \quad \Omega_T(Z_t) \quad \tilde{m}(\varphi, Z_t)
\]

where

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

\(\hat{f}(y, x|z)\) is kernel estimator of the density of \((Y, X)\) given \(Z = z\)

- **Penalty term:** involves the Sobolev norm \(\|\varphi\|_H\) defined by

\[
\|\varphi\|_H^2 = \|\varphi\|^2 + \|\nabla \varphi\|^2
\]

where \(\|\varphi\|^2 = \int \varphi(x)^2 dx\)

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

\[
\hat{\varphi} = \arg \inf_{\varphi \in \Theta} Q_T(\varphi) + \lambda_T \|\varphi\|_H^2
\]

where \(\lambda_T\) is a stochastic sequence such that \(\lambda_T \geq 0\) and \(\lambda_T \to 0\) \(P\text{-a.s.}\)
Tikhonov Regularised (TiR) estimator (II)

Intuition:

$$\hat{\varphi} = \arg\inf_{\varphi \in \Theta} \ Q_T(\varphi) + \lambda_T \cdot \|\varphi\|^2_H$$

- Minimum distance criterion is flat along some directions spanned by highly oscillating functions because of ill-posedness
- Penalty term damps out these highly oscillating components
- Sequence $\lambda_T$ tunes the amount of regularisation

First order condition: 

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

TiR estimator: defined on the function space

$$\hat{\varphi} = \left(\lambda_T + \tilde{A}^* \tilde{A}\right)^{-1} \tilde{A}^* \hat{r}$$
Implementation of the TiR estimator

- **Numerical approximation:**

\[
\varphi(x) \simeq \sum_{j=0}^{5} \theta_j P_j(x) =: \theta' P(x) \quad , \quad x \in [0, 1]
\]

where the \( P_j \) are the shifted Chebyshev polynomials of the first kind

- **Sobolev norm:** \( \| \varphi \|_{H^2} \simeq \theta'D\theta \) (quadratic in \( \theta \! \!) \) with

\[
D = \begin{pmatrix}
\frac{1}{\pi} & 0 & -\sqrt{2} & 0 & -\sqrt{2} & 0 \\
0 & \frac{3}{\pi} & 0 & \frac{15}{5\pi} & 0 & 0 \\
\vdots & 0 & \frac{21}{5\pi} & 0 & \frac{1182}{35\pi} & 0 \\
\frac{26}{3\pi} & 0 & \frac{38}{5\pi} & \frac{3898}{35\pi} & 0 & \frac{5090}{63\pi} \\
\vdots & \frac{315}{3\pi} & 0 & 67894 & 0 & \frac{82802}{231\pi} \\
\end{pmatrix}
\]

- **Closed form estimator for the linear case:**

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

matrices \( \hat{P} \) and \( \hat{\tilde{r}} \) have rows \( \int P(x)' \tilde{f}(x \mid Z_t) dx \) resp. \( \int y \tilde{f}(y \mid Z_t) dy \)
Links with the literature: 
Advantages of the TiR approach (I)

Regularisation by compactness (NP and AC)

- Induces compactness of the parameter space by imposing the inequality constraint $\|\varphi\|^2_H \leq \bar{B}$
- $\lambda_T$ is interpreted as a Kuhn-Tucker multiplier and is determined by the slackness condition: either $\|\hat{\varphi}\|^2_H = \bar{B}$ or $\lambda_T = 0$ P-a.s.

Advantages of the TiR estimator

- Features a faster rate of convergence than estimators with fixed $\bar{B}$ since sequence $\lambda_T$ may be optimally selected
- Allows for data-driven regularisation parameter $\lambda_T$, whereas tuning parameter $\bar{B}$ is fixed in the theory of NP and AC
- Is defined by an unconstrained optimization problem and admits a closed form expression in the linear case
Links with the literature: Advantages of the TiR approach (II)

Regularisation with $L^2$ norm (DFR and HH)

- Approach by DFR and HH can be seen as Tikhonov regularisation with $L^2$ penalty $G(\varphi) = \|\varphi\|^2$ (without any derivative $\nabla \varphi$)

Advantages of the TiR estimator

- May feature a faster rate of convergence
- Clear-cut superior finite-sample performance in our two examples
- Applies to linear and non-linear conditional moment restrictions
Consistency of TiR estimator

**Theorem 1:** Let

\[ \hat{\varphi} = \operatorname{arg\ inf}_{\varphi \in \Theta} Q_T(\varphi) + \lambda_T G(\varphi) \]

and assume regularity conditions. Then, if \( \lambda_T > 0, \lambda_T \to 0, (\lambda_T T)^{-1} \to 0 \) P-a.s., estimator \( \hat{\varphi} \) is consistent: \( \| \hat{\varphi} - \varphi_0 \|_p \to 0 \).

**Remarks:**

- Theorem 1 is a general consistency result for penalized extremum estimators holding for any function \( G \) and possibly stochastic \( \lambda_T \).
- When \( G(\varphi) = \| \varphi \|^2_H \), Theorem 1 implies the consistency of the TiR estimator with data-driven regularisation parameter \( \lambda_T \).
Proposition 2:

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

up to terms asymptotically negligible w.r.t. RHS, where \( \{\phi_j\} \) are orthonormal eigenfunctions of \( A^*A \) to eigenvalues \( \nu_j \), \( A^* \) is adjoint of \( A \) and

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

Remarks:

- Bias \( b(\lambda_T) \) is induced by regularisation
- Ill-posedness implies \( \nu_j \to 0 \) as \( j \to \infty \): the variance term converges to zero slower than \( 1/T \)!
- Similar formula for the MISE with \( L^2 \) regularisation
TiR estimator (solid) and OLS (dashed), \( \varphi_0 = \sin, T=400, \rho=0.5 \)
$L^2$ regularis. (solid) and OLS (dashed), $\varphi_0 = \sin$, $T=400$, $\rho=0.5$
TiR estimator (solid) and OLS (dashed), $\varphi_0 = \sin, T=400, \rho=0$
$L^2$ regularisation (solid) and OLS (dashed), $\varphi_0 = \sin$, $T=400$, $\rho=0$
Optimal rates of convergence (I)

Optimal sequence of regularisation parameters: \( \lambda^*_T = \arg \min_{\lambda > 0} M_T(\lambda) \)
Optimal MISE of the TiR: \( M^*_T = M_T(\lambda^*_T) \)

**Assumption G**: For \( j = 1, 2, \ldots \) and \( \lambda > 0 \)
(i) \( \nu_j = C_1 \exp(-\alpha j) \), \( \alpha > 0 \) \quad \text{(Geometric decay of the spectrum of } A^*A) \)
(ii) \( \| \phi_j \|^2 = C_2 j^{-\beta} \), \( \beta > 0 \) \quad \text{ (iii) } b(\lambda) = C_3 \lambda^\delta \), \( \delta > 0 \)

**Proposition 3**: Under Assumption G, up to negligible terms
(i) \( \log \lambda^*_T = \log c - \frac{1}{1+2\delta} \log T \)
(ii) \( M^*_T = \bar{c} T^{-\frac{2\delta}{1+2\delta}} (\log T)^{-\frac{2\delta\beta}{1+2\delta}} \)

**Optimal rates of convergence**: general picture

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

Case 1: Beta

Case 2: Sin
Data-driven selection of $\lambda_T$: Algorithm

**Idea:** Estimate the asymptotic MISE and minimize it w.r.t. $\lambda$!

**Algorithm:**

(i) Spectral decomposition of matrix $D^{-1}\hat{P}'\hat{P}/T$: eigenvalues $\hat{\nu}_j$ and eigenvectors $\hat{w}_j$, $\hat{w}'_jD\hat{w}_j = 1$, $j = 1, \ldots, 6$

(ii) First-step TiR estimator $\bar{\theta}$ using small pilot regularisation parameter $\bar{\lambda}$

(iii) Estimate the MISE:

$$
\bar{M}(\lambda) = \frac{1}{T} \sum_{j=1}^{6} \frac{\hat{\nu}_j}{(\lambda + \hat{\nu}_j)^2} \hat{w}_j' B \hat{w}_j 
+ \bar{\theta}' \left[ \frac{1}{T} \hat{P}' \hat{P} \left( \lambda D + \frac{1}{T} \hat{P}' \hat{P} \right)^{-1} - I \right] B \left[ \frac{1}{T} \hat{P}' \hat{P} \left( \lambda D + \frac{1}{T} \hat{P}' \hat{P} \right)^{-1} - I \right] \bar{\theta}
$$

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

(iv) Compute second-step TiR estimator $\hat{\theta}$ using regularisation parameter $\hat{\lambda}$
## Data-driven selection of $\lambda_T$: Monte-Carlo, $T = 1000$

<table>
<thead>
<tr>
<th></th>
<th>Beta</th>
<th>Sin</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\bar{\lambda}$</td>
<td>$0.0005$</td>
<td>$0.0001$</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>$0.0005$</td>
<td>$0.0001$</td>
</tr>
<tr>
<td>Optimal $\lambda^*_T$</td>
<td>$0.0013$</td>
<td>$0.0007$</td>
</tr>
<tr>
<td>Mean $\hat{\lambda}$</td>
<td>$0.0028$</td>
<td>$0.0027$</td>
</tr>
<tr>
<td>25% quartile $\hat{\lambda}$</td>
<td>$0.0014$</td>
<td>$0.0007$</td>
</tr>
<tr>
<td>Median $\hat{\lambda}$</td>
<td>$0.0020$</td>
<td>$0.0014$</td>
</tr>
<tr>
<td>75% quartile $\hat{\lambda}$</td>
<td>$0.0033$</td>
<td>$0.0029$</td>
</tr>
<tr>
<td>Optimal $M^*_T$</td>
<td>$0.0099$</td>
<td>$0.0121$</td>
</tr>
<tr>
<td>MISE of data-driven TiR</td>
<td>$0.0120$</td>
<td>$0.0156$</td>
</tr>
</tbody>
</table>
MSE and asymptotic normality of the TiR

Proposition 4:

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

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

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

Proposition 5:

\[ \sqrt{T/\sigma_T^2(x)} (\hat{\varphi}(x) - \varphi_0(x) - B_T(x)) \xrightarrow{d} N(0, 1) \]
**Empirical application: Engel curve estimation**

**Engel curve**: based on $E[Y - \varphi_0(X) \mid Z] = 0$, $X = \Phi(X^*)$

- $Y =$ food expenditure share
- $X^* =$ logarithm of total expenditures
- $Z =$ logarithm of annual income from wages and salaries

**Data**: $T = 785$ households from 1996 US Consumer Expenditure Survey

---

**Estimated Engel curve**

![Estimated Engel curve](image1.png)

**Estimated Engel curve**

![Estimated Engel curve](image2.png)
Concluding remarks

- We introduced a new estimator of a functional parameter identified by conditional moment restrictions

- Ill-posedness is addressed with Tikhonov regularisation by penalizing the Sobolev norm of the estimator

- Our approach proves to be:

  (i) numerically tractable (closed form in linear case)

  (ii) well-behaved in finite sample

  (iii) amenable to in-depth asymptotic analysis (MISE, rates of convergence, asymptotic normality)

  ⇒ A route towards:
  - numerous empirical applications!
  - further theoretical developments: asymptotics for data-driven estimators, estimation of functional derivatives, semiparametric models, etc.