Functional linear regression with functional response*

David Benatia  
Université de Montréal

Marine Carrasco  
Université de Montréal

Jean-Pierre Florens  
Toulouse School of Economics

May 2015

Abstract

In this paper, we develop new estimation results for functional regressions where both the regressor \( Z(t) \) and the response \( Y(t) \) are functions of an index such as the time or a spatial location. Both \( Z(t) \) and \( Y(t) \) are assumed to belong to Hilbert spaces. The model can be thought as a generalization of the multivariate regression where the regression coefficient is now an unknown operator \( \Pi \). An interesting feature of our model is that \( Y(t) \) depends not only on contemporaneous \( Z(t) \) but also on past and future values of \( Z \).

We propose to estimate the operator \( \Pi \) by Tikhonov regularization, which amounts to apply a penalty on the \( L^2 \) norm of \( \Pi \). We derive the rate of convergence of the mean-square error, the asymptotic distribution of the estimator, and develop tests on \( \Pi \). Often, the full trajectories are not observed but only a discretized version is available. We address this issue in the scenario where the data become more and more frequent (in-fill asymptotics). We also consider the case where \( Z \) is endogenous and instrumental variables are used to estimate \( \Pi \).

Key Words: Functional regression, instrumental variables, linear operator, Tikhonov regularization

*The authors thank the participants of the 6th French Econometrics Conférence and especially their discussant Jan Johannes for helpful comments. Carrasco thanks NSERC for financial support.
1 Introduction

With the increase of storage capability, continuous time data are available in many fields including finance, medicine, meteorology, and microeconometrics. Researchers, companies, and governments look for ways to exploit this rich information. In this paper, we develop new estimation results for functional regressions where both the regressor $Z(t)$ and the response $Y(t)$ are functions of an index such as the time or a spatial location. Both $Z(t)$ and $Y(t)$ are assumed to belong to Hilbert spaces. The model can be thought as a generalization of the multivariate regression where the regression coefficient is now an unknown operator $\Pi$. An interesting feature of our model is that $Y(t)$ depends not only on contemporaneous $Z(t)$ but also on past and future values of $Z$.

We propose to estimate the operator $\Pi$ by Tikhonov regularization, which amounts to apply a penalty on the $L^2$ norm of $\Pi$. The choice of a $L^2$ penalty, instead of $L^1$ used in Lasso, is motivated by the fact that - in the applications we have in mind - there is no reason to believe that the relationship between $Y$ and $Z$ is sparse. We derive the rate of convergence of the mean-square error (MSE) and the asymptotic distribution of the estimator for a fixed regularization parameter $\alpha$ and develop tests on $\Pi$. In some applications, it would be interesting to test whether $Y(t)$ depends only on the past values of $Z$ or only on contemporaneous values of $Z$. If the application is on network and $t$ refers to the spatial location, our model could describe how the behavior of a firm $Y(t)$ depends on the decision of neighboring firms $Z(s)$. Testing properties of $\Pi$ will help to characterize the strategic response of firms.

Often, the full trajectories are not observed but only a discretized version is available. This case raises specific challenges which will be addressed in the scenario where the data become more and more frequent (in-fill asymptotics).

We also consider the case where $Z$ is endogenous and instrumental variables are used to estimate $\Pi$. To the best of our knowledge, the model with functional response and endogenous functional regressor has never been studied before. We derive an estimator based on Tikhonov regularization and show its rate of convergence.

There is a large body of work done on linear functional regression where the response is a scalar variable $Y$ and the regressor is a function. Some recent references include Cardo, Ferraty, and Sarda (2003), Hall and Horowitz (2007), Horowitz and Lee (2007),
Darolles, Fan, Florens and Renault (2011), and Crambes, Kneib, and Sarda (2009). In contrast, only a few researchers have tackled the functional linear regression in which both the predictor $Z$ and the response $Y$ are random functions. The object of interest is the estimation of the conditional expectation of $Y$ given $Z$. In this setting, the unknown parameter is an integral operator. This model is discussed in the monographs by Ramsay and Silverman (2005) and Ferraty and Vieu (2006). Cuevas, Febrero, and Fraiman (2002) consider a fixed design setting and propose an estimator of $\Pi$ based on interpolation. Yao, Müller, and Wang (2005) consider the case where both predictor and response trajectories are observed at discrete and irregularly spaced times. Their estimator is based on spectral cut-off regularized inverse using nonparametric estimators of the principal components. Crambes and Mas (2013) consider again a spectral cut-off regularized inverse and derive the asymptotic mean square prediction error which is then used to derive the optimal choice of the regularization parameter. Our model is also related to the functional autoregressive (FAR) model studied by Bosq (2000), Kargin and Onatski (2008), and Aue, Norinho, and Hormann (2014), among others. The estimation methods used in these papers are based on functional principal components and differ from ours. Antoch, Prchal, Rosa, and Sarda (2010) use a FAR to forecast the electricity consumption. In their model, the weekday consumption curve is explained by the curve from the previous week. The authors use B-spline to estimate the operator.

The paper is organized as follows. Section 2 introduces the model and the estimators. Section 3 derives the rate of convergence of the MSE. Section 4 presents the asymptotic normality of the estimator for a fixed regularization parameter. Issues relative to the choice of the regularization parameter are discussed in Section 5. Discrete observations are addressed in Section 6. Section 7 considers an endogenous regressor. Section 8 presents simulation results. Section 9 presents an application to the electricity market. The proofs are collected in Appendix.

2 The model and estimator

2.1 The model

We consider a regression model where both the predictor and response are random functions. We observe pairs of random trajectories $(y_i, z_i)$ $i = 1, 2, \ldots, n$ with square
integrable predictor trajectories \( z_i \) and response trajectories \( y_i \). They are realizations of random processes \((Y, Z)\) with zero mean functions and unknown covariance operators. The extension to the case, where the mean is unknown but estimated, is straightforward. The arguments of \( Y \) and \( Z \) are denoted \( t \) which may refer to the time, a location or a characteristic such as the age or income of an agent.

We assume that \( Y \) belongs to a Hilbert space \( \mathcal{E} \) equipped with an inner product \( \langle \cdot, \cdot \rangle \) and \( Z \) belongs to a Hilbert space \( \mathcal{F} \) equipped with an inner product \( \langle \cdot, \cdot \rangle \) (to simplify notations, we use the same notation for both inner products even though they usually differ).

The model is

\[
Y = \Pi Z + U \tag{1}
\]

where \( U \) is a zero mean random element of \( \mathcal{E} \) and \( \Pi \) is a nonrandom Hilbert-Schmidt operator from \( \mathcal{F} \) to \( \mathcal{E} \). Moreover, \( Z \) is exogenous so that \( \text{cov}(Z, U) = 0 \). This assumption will be relaxed in Section 7.

For illustration, consider the following example

\[
\mathcal{E} = \left\{ g : \int_S g(t)^2 \, dt < \infty \right\},
\]
\[
\mathcal{F} = \left\{ f : \int_T f(t)^2 \, dt < \infty \right\}
\]

where \( S \) and \( T \) are some intervals of \( \mathbb{R} \). Then, \( \Pi \) can be represented as an integral operator such that

\[
(\Pi \varphi)(s) = \int_T \pi(s, t) \varphi(t) \, dt
\]

for any \( \varphi \in \mathcal{F} \). \( \pi \) is referred to as the kernel of the operator \( \Pi \). Model (1) means that \( Y(t) \) depends not only on \( Z(t) \) but also on all the \( Z(s) \), for \( s \neq t \). The object of interest is the estimation of the operator \( \Pi \).

Below, we describe two examples of applications of Model (1).

**Example 1.** Electricity market.

Let \( Y_i(t) \) be the electricity consumption for the province of Ontario in Canada for day \( i \) at hour \( t \), and \( Z_i(t) \) be the average temperature for the same province for day \( i \) at hour \( t \). We believe that the consumption of electricity over one day may depend on the temperature of the current day but also on that of the previous and following
days. We could model this relationship as a multivariate regression where the dependent variable is the $24 \times 1$ vector of electricity consumptions at different hours of the day and the dependent variable is a $72 \times 1$ vector of temperatures. Estimating this model by ordinary least squares (OLS) would require inverting a $72 \times 72$ matrix which would likely be near singular. The variance of the resulting OLS estimator would be very large. Moreover, if the frequency of the observations increases, it is expected that the successive observations will become more and more correlated. It is therefore natural to assume that the observations result from the discretization of a curve. Model (1) seems to be an attractive alternative to the multivariate regression in this setting. This application will be continued in Section 9.

**Example 2.** Technological spillovers in productivity

This example is inspired from Manresa (2015). Assume that $Y_i(t)$ is the log of the average output of firms with industry code $t$ at time $i$ and $Z_i(t)$ is the log of the average R&D expenditures for firms with industry code $t$ at time $i$. Industry code goes from 1 to 6 digits. A higher number of digits correspond to a thinner grid on the continuum of industry codes. Model (1) would permit to characterize how firms from one sector benefit from R&D advancements done in adjacent sectors. This example will not be pursued here.

### 2.2 The estimator

We denote $V_Z$ the operator from $\mathcal{F}$ to $\mathcal{F}$ which associates to functions $\varphi \in \mathcal{F}$:

$$V_Z \varphi = E [Z \langle Z, \varphi \rangle].$$

Note that, as $Z$ is centered, $V_Z$ is the covariance operator of $Z$. We denote $C_{YZ}$ the covariance operator of $(Y,Z)$. It is the operator from $\mathcal{F}$ to $\mathcal{E}$ such that

$$C_{YZ} \varphi = E [Y \langle Z, \varphi \rangle]$$

Using (1), we have

$$\text{cov} (Y,Z) = \text{cov} (\Pi Z + u, Z) = \Pi \text{cov} (Z,Z) + \text{cov} (u, Z).$$
Hence, we have the following relationships:

\[ C_{YZ} = \Pi V_Z, \]  \hspace{1cm} (2)
\[ C_{ZY} = V_Z \Pi^* \]  \hspace{1cm} (3)

where \( \Pi^* \) is the adjoint of \( \Pi \). \( C_{ZY} \) is defined as the operator from \( \mathcal{E} \) to \( \mathcal{F} \) such that

\[ C_{ZY} \psi = E[Z \langle Y, \psi \rangle] \]

for any \( \psi \) in \( \mathcal{E} \). Note that \( C_{ZY} \) is the adjoint of \( C_{YZ}, C_{YZ}^* \).

First we describe how to estimate \( \Pi^* \) using (3). The unknown operators \( V_Z \) and \( C_{ZY} \) are replaced by their sample counterparts. The sample estimate of \( V_Z \) is

\[ \hat{V}_Z \varphi = \frac{1}{n} \sum_{i=1}^{n} z_i \langle z_i, \varphi \rangle \]

for \( \varphi \in \mathcal{F} \). The sample estimate of \( C_{ZY} \) is

\[ \hat{C}_{ZY} \psi = \frac{1}{n} \sum_{i=1}^{n} z_i \langle y_i, \psi \rangle \]

for \( \psi \in \mathcal{E} \). An estimator of \( \Pi^* \) can not be obtained directly by solving \( \hat{C}_{ZY} = \hat{V}_Z \Pi^* \) because the initial equation \( C_{ZY} = V_Z \Pi^* \) is an ill-posed problem in the sense that \( V_Z \) is invertible only on a subset of \( \mathcal{E} \) and its inverse is not continuous. Note that \( \hat{V}_Z \) has finite rank equal to \( n \) and hence is not invertible. A Moore-Penrose generalized inverse could be used but it would not be continuous. To stabilize the inverse, we need to use some regularization scheme. We adopt Tikhonov regularization (see Kress, 1999 and Carrasco, Florens, and Renault, 2007).

The estimator of \( \Pi^* \) is defined as

\[ \hat{\Pi}^*_\alpha = \left( \alpha I + \hat{V}_Z \right)^{-1} \hat{C}_{ZY} \]  \hspace{1cm} (4)

and that of \( \Pi \) is defined by

\[ \hat{\Pi}_\alpha = \hat{C}_{YZ} \left( \alpha I + \hat{V}_Z \right)^{-1} \]  \hspace{1cm} (5)
where \( \alpha \) is some positive regularization parameter which will be allowed to converge to zero as \( n \) goes to infinity. The estimators (4) and (5) can be viewed as generalization of ordinary least-squares estimators. They also have an interpretation as the solution to an inverse problem.

At this stage, it is useful to make the link with the inverse problem literature. Let \( \mathcal{H} \) be the Hilbert space of linear Hilbert-Schmidt operators from \( \mathcal{F} \) to \( \mathcal{E} \). The inner product on \( \mathcal{H} \) is
\[
\langle \Pi_1, \Pi_2 \rangle_{\mathcal{H}} = tr (\Pi_1 \Pi_2^*).
\]
Dropping the error term in (1), we obtain, for the sample, the equation
\[
\hat{r} = K \Pi
\]
where \( \hat{r} = (y_1, ..., y_n)' \) and \( K \) is the operator from \( \mathcal{H} \) to \( \mathcal{E}^n \) such that \( K \Pi = (\Pi z_1, ..., \Pi z_n)' \).

The inner product on \( \mathcal{E}^n \) is
\[
\langle f, g \rangle_{\mathcal{E}^n} = \frac{1}{n} \sum_{i=1}^{n} \langle f_i, g_i \rangle_{\mathcal{E}}
\]
with \( f = (f_1, ..., f_n)' \) and \( g = (g_1, ..., g_n)' \). Let us check that \( \hat{\Pi}_\alpha \) is a classical Tikhonov regularized inverse of the operator \( K \):
\[
\hat{\Pi}_\alpha = (\alpha I + K^* K)^{-1} K^* \hat{r}.
\]

We need to find \( K^* \). We look for the operator \( B \) from \( \mathcal{F} \) to \( \mathcal{E} \) solution of
\[
\langle K \Pi, f \rangle_{\mathcal{E}^n} = \langle \Pi, B \rangle_{\mathcal{H}} .
\]

Note that
\[
\langle \Pi, B \rangle_{\mathcal{H}} = tr (\Pi B^*) \\
= \sum_j \langle \Pi B^* \varphi_j, \varphi_j \rangle \\
= \sum_j \langle B^* \varphi_j, \Pi^* \varphi_j \rangle
\]
where \( \varphi_j \) is a basis of \( \mathcal{E} \). On the other hand,

\[
\langle K\Pi, f \rangle_{\mathcal{E}^*} = \frac{1}{n} \sum_i \langle \Pi z_i, f_i \rangle_{\mathcal{E}}
\]

\[
= \frac{1}{n} \sum_i \langle z_i, \Pi^* f_i \rangle_{\mathcal{F}}.
\]

Using \( f_i = \sum_j \langle f_i, \varphi_j \rangle \varphi_j \), we obtain

\[
\langle K\Pi, f \rangle_{\mathcal{E}^*} = \frac{1}{n} \sum_i \sum_j \langle f_i, \varphi_j \rangle \langle z_i, \Pi^* \varphi_j \rangle
\]

\[
= \sum_j \left( \frac{1}{n} \sum_i \langle f_i, \varphi_j \rangle z_i, \Pi^* \varphi_j \right).
\]

It follows from (6) that \( B^* \varphi_j = \frac{1}{n} \sum_i \langle f_i, \varphi_j \rangle z_i \) for all \( j \) and hence

\[
B^* \varphi = \frac{1}{n} \sum_i \langle f_i, \varphi \rangle z_i
\]

for all \( \varphi \) in \( \mathcal{E} \). Now, we look for \( B \) the adjoint of \( B^* \). \( B \) is the solution of

\[
\langle B^* \varphi_1, \varphi_2 \rangle_{\mathcal{F}} = \langle \varphi_1, B \varphi_2 \rangle_{\mathcal{E}}.
\]

We have

\[
\langle B^* \varphi_1, \varphi_2 \rangle_{\mathcal{F}} = \frac{1}{n} \sum_i \langle f_i, \varphi_1 \rangle \langle z_i, \varphi_2 \rangle_{\mathcal{F}}
\]

\[
= \left\langle \varphi_1, \frac{1}{n} \sum_i \langle z_i, \varphi_2 \rangle f_i \right\rangle_{\mathcal{E}}.
\]

Hence,

\[
B \varphi = (K^* f) \varphi = \frac{1}{n} \sum_i \langle z_i, \varphi \rangle f_i.
\]

We have

\[
K^* K \Pi = \frac{1}{n} \sum_i \langle z_i, \varphi \rangle \Pi z_i = \Pi \hat{V}_Z
\]
and

\[ K^{*}\hat{r} = \frac{1}{n} \sum_{i} \langle z_{i}, \cdot \rangle y_{i} = \hat{C}_{YZ}. \]

It follows that

\[
\hat{\Pi}_{\alpha} = (\alpha I + K^{*}K)^{-1} K^{*}\hat{r} = \hat{C}_{YZ} \left( \alpha I + \hat{V}_{Z} \right)^{-1}.
\]

The estimator \( \hat{\Pi}_{\alpha} \) is also a penalized least-squares estimator:

\[
\hat{\Pi}_{\alpha} = \arg \min_{\Pi} \| y - \Pi z \|^2 + \alpha \| \Pi \|_{HS}^2
\]

\[
= \arg \min_{\Pi} \sum_{i=1}^{n} \| y_{i} - \Pi z_{i} \|^2 + \alpha \sum \hat{\mu}_{j}^2
\]

where \( \hat{\mu}_{j} \) are the singular values of the operator \( \Pi \).

### 2.3 Identification

It is easier to study the identification from the viewpoint of Equation (3). Let \( \mathcal{H} \) be the space of Hilbert-Schmidt operators from \( \mathcal{E} \) to \( \mathcal{F} \). Let \( T \) be the operator from \( \mathcal{H} \) to \( \mathcal{H} \) defined as

\[ TH = V_{Z}H \text{ for } H \text{ in } \mathcal{H}. \]

According to (3), \( \Pi^{*} \) is identified if and only if \( T \) is injective.

\( V_{Z} \) injective implies \( T \) injective. Indeed, we have

\[ TH = 0 \]

\[ \iff \quad V_{Z}H = 0 \]

\[ \iff \quad V_{Z}H\psi = 0, \quad \forall \psi \]

\[ \iff \quad H\psi = 0, \quad \forall \psi \]

by the injectivity of \( V_{Z} \). Hence \( H = 0 \). It turns out that \( T \) is injective if and only if \( V_{Z} \) is injective. This can be shown by deriving the spectrum of \( T \).
First, we show that $T$ is self-adjoint. The adjoint $T^*$ of $T$ satisfies
\[
\langle TH, K \rangle = \langle H, T^* K \rangle
\]
for arbitrary operators $H$ and $K$ of $\mathcal{H}$. We have
\[
\begin{align*}
\langle TH, K \rangle &= tr(THK^*) \\
&= tr(V_ZHK^*) \\
&= tr(HK^*V_Z)
\end{align*}
\]
because $V_Z$ is self-adjoint. Hence, $T^* K = (K^*V_Z)^* = V_Z K = TK$. Therefore, $T$ is self-adjoint.

The spectrum of $T$ is also closely related to that of $V_Z$. Let $(\mu_j, H_j)_{j=1,2,...}$ denote the eigenvalues and eigenfunctions of $T$ and $(\lambda_j, \varphi_j)_{j=1,2,...}$ be the eigenvalues and eigenfunctions of $V_Z$ so that $V_Z \varphi_j = \lambda_j \varphi_j$. $H_j$ is necessarily of the form, $H_j = \varphi_j \langle \iota, \cdot \rangle$ where $\iota$ is the 1 function in $\mathcal{E}$. Then,
\[
TH_j = V_Z \varphi_j \langle \iota, \cdot \rangle \\
= \lambda_j \varphi_j \langle \iota, \cdot \rangle \\
= \lambda_j H_j.
\]
So that the eigenvalues of $T$ are the same as those of $V_Z$.

In summary, a necessary and sufficient condition for the identification of $\Pi$ is that $V_Z$ is injective.

### 2.4 Computation of the estimator

To show how to compute $\hat{\Pi}_\alpha^*$ explicitly, we multiply the left and right of (4) by $\left( \alpha I + \hat{V}_Z \right)$ to obtain
\[
\begin{align*}
\hat{C}_{ZY} \psi &= \left( \alpha I + \hat{V}_Z \right) \hat{\Pi}_\alpha^* \psi \\
&\Leftrightarrow \\
\frac{1}{n} \sum_{i=1}^{n} z_i \langle y_i, \psi \rangle &= \alpha \hat{\Pi}_\alpha^* \psi + \frac{1}{n} \sum_{i=1}^{n} z_i \left\langle z_i, \hat{\Pi}_\alpha^* \psi \right\rangle.
\end{align*}
\]
Then, we take the inner product with $z_l$, $l = 1, 2, ..., n$ on the left and right hand side of (7), to obtain $n$ equations:

$$
\frac{1}{n} \sum_{i=1}^{n} \langle z_l, z_i \rangle \langle y_i, \psi \rangle = \alpha \langle z_l, \hat{\Pi}_\alpha^* \psi \rangle + \frac{1}{n} \sum_{i=1}^{n} \langle z_l, z_i \rangle \langle z_i, \hat{\Pi}_\alpha^* \psi \rangle, \ l = 1, 2, ..., n, \tag{8}
$$

with $n$ unknowns $\langle z_i, \hat{\Pi}_\alpha^* \psi \rangle$, $i = 1, 2, ..., n$. Let $M$ be the $n \times n$ matrix with $(l, i)$ element $\langle z_l, z_i \rangle / n$, $v$ the $n$–vector of $\langle z_i, \hat{\Pi}_\alpha^* \psi \rangle$ and $w$ the $n$–vector of $\langle y_i, \psi \rangle$. (8) is equivalent to

$$
Mw = (\alpha I + M)v.
$$

And $v = (\alpha I + M)^{-1} Mw = M (\alpha I + M)^{-1} w$. For a given $\psi$, we can compute:

$$
\hat{\Pi}_\alpha^* \psi = \frac{1}{\alpha n} \sum_{i=1}^{n} z_i \left( \langle y_i, \psi \rangle - \langle z_i, \hat{\Pi}_\alpha^* \psi \rangle \right) \tag{9}
$$

$$
= \frac{1}{\alpha n} \hat{z}' (I - M (\alpha I + M)^{-1}) w
$$

$$
= \frac{1}{n} \hat{z}' (\alpha I + M)^{-1} w
$$

where $\hat{z}$ is the $n$–vector of $z_i$.

Now, we explain how to estimate $\Pi \varphi$ for any $\varphi \in \mathcal{F}$. Taking the inner product with $\varphi$ in the left and right hand sides of (9), we obtain

$$
\langle \varphi, \hat{\Pi}_\alpha^* \psi \rangle = \frac{1}{\alpha n} \sum_{i=1}^{n} \langle \varphi, z_i \rangle \left( \langle y_i, \psi \rangle - \langle z_i, \hat{\Pi}_\alpha^* \psi \rangle \right) \Leftrightarrow
$$

$$
\langle \hat{\Pi}_\alpha \varphi, \psi \rangle = \frac{1}{\alpha n} \sum_{i=1}^{n} \langle \varphi, z_i \rangle \langle y_i - \hat{\Pi}_\alpha z_i, \psi \rangle
$$

for all $\psi \in \mathcal{E}$. This implies

$$
\hat{\Pi}_\alpha \varphi = \frac{1}{\alpha n} \sum_{i=1}^{n} \langle \varphi, z_i \rangle \left( y_i - \hat{\Pi}_\alpha z_i \right). \tag{10}
$$
Hence, to compute $\hat{\Pi}_\alpha \varphi$, we need to know $\hat{\Pi}_\alpha z_i$. From (5), we have

$$\alpha \hat{\Pi}_\alpha + \hat{\Pi}_\alpha \hat{V}_Z = \hat{C}_{YZ}.$$  

Applying the l.h.s and r.h.s to $z_i$, $i = 1, 2, ..., n$, we obtain

$$\alpha \hat{\Pi}_\alpha z_i + \hat{\Pi}_\alpha \hat{V}_Z z_i = \hat{C}_{YZ} z_i \Leftrightarrow$$

$$\alpha \left( \hat{\Pi}_\alpha z_i \right) (t) + \frac{1}{n} \sum_{j=1}^{n} \left( \hat{\Pi}_\alpha z_j \right) (t) \langle z_j, z_i \rangle = \frac{1}{n} \sum_{j=1}^{n} y_j (t) \langle z_j, z_i \rangle, \; i = 1, 2, ..., n. \; (11)$$

For each $t$, we can solve the $n$ equations with $n$ unknowns $\left( \hat{\Pi}_\alpha z_j \right) (t)$ given by (11) and deduct $\hat{\Pi}_\alpha \varphi$ from (10).

The prediction of $Y_i$ is given by

$$\hat{y}_i = \hat{\Pi}_\alpha z_i.$$

### 3 Rate of convergence of the MSE

In this section, we study the rate of convergence of the mean square error (MSE) of $\hat{\Pi}_\alpha^*$. Several assumptions are needed.

Assumption 1. $U_i$ is a random process of $\mathcal{E}$ such that $E (U_i) = 0$, $cov(U_i, U_j | Z_1, Z_2, ..., Z_n) = 0$ for all $i \neq j$ and $= V_U$ for $i = j$ where $V_U$ is a trace-class operator.

Assumption 2. $\Pi$ belongs to $\mathcal{H} (\mathcal{F}, \mathcal{E})$ the space of Hilbert-Schmidt operators.

Assumption 3. $V_Z$ is a trace-class operator and $\| \hat{V}_Z - V_Z \|_{HS}^2 = O_p (1/n)$.

Assumption 4. There is a Hilbert-Schmidt operator $R$ from $\mathcal{E}$ to $\mathcal{F}$ and a constant $\beta > 0$ such that $\Pi^* = V_Z^{\beta/2} R$.

An operator $K$ is trace-class if $\sum_j \langle K \phi_j, \phi_j \rangle < \infty$ for any basis $(\phi_j)$. If $K$ is self-adjoint positive definite, it is equivalent to say that the sum of the eigenvalues of $K$ is finite. Given $V_U$ is a covariance operator, $V_U$ is trace-class if and only if $E (\|U_i\|^2) < \infty$.

The notation $\| \|_{HS}$ refers to the Hilbert-Schmidt norm of operators. An operator $K$ is Hilbert-Schmidt (noted HS) if $\| K \|_{HS}^2 \equiv \sum_j \langle K \phi_j, K \phi_j \rangle < \infty$ for any basis $(\phi_j)$. If $K$ is self-adjoint positive definite, it is equivalent to the condition that the eigenvalues of $K$ are square summable. A sufficient condition for $\| \hat{V}_Z - V_Z \|_{HS}^2 = O_p (1/n)$ is that
$Z_i$ is a i.i.d. random process and $E \left( \|V_i\|^4 \right) < \infty$, see Proposition 5 of Dauxois, Pousse, and Romain (1982).

Assumption 4 is a source condition needed to characterize the rate of convergence of the MSE. Moreover, it guarantees that $\Pi^*$ belongs to the orthogonal of the null space of $V_Z$ denoted $\mathcal{N}(V_Z)$. Given this condition, there is no need to impose $\mathcal{N}(V_Z) = \{0\}$ to get the identification.

The MSE is defined by

$$ E \left( \|\hat{\Pi}_\alpha - \Pi\|^2_{HS} | Z_1, \ldots, Z_n \right). $$

**Proposition 1** Under Assumption 3, $\hat{\Pi}_\alpha$ belongs to $\mathcal{H}(\mathcal{F}, \mathcal{E})$ for all $\alpha > 0$.

Proof: See Appendix.

Replacing $y_i$ by $\Pi z_i + u_i$ in the expression of $\hat{C}_{ZY}$, we obtain

$$ \hat{C}_{ZY} = \frac{1}{n} \sum_i z_i \langle y_i, . \rangle $$

$$ = \frac{1}{n} \sum_i z_i \langle u_i, . \rangle + \frac{1}{n} \sum_i z_i \langle \Pi z_i, . \rangle $$

$$ = \hat{C}_{ZU} + \hat{V}_Z \Pi^*. $$

We decompose $\hat{\Pi}_\alpha^* - \Pi^*$ in the following manner:

$$ \hat{\Pi}_\alpha^* - \Pi^* = \left( \alpha I + \hat{V}_Z \right)^{-1} \hat{C}_{ZV} - \Pi^* $$

$$ = \left( \alpha I + \hat{V}_Z \right)^{-1} \hat{C}_{ZU} + \left( \alpha I + \hat{V}_Z \right)^{-1} \hat{V}_Z \Pi^* - (\alpha I + V_Z)^{-1} V_Z \Pi^* $$

$$ + (\alpha I + V_Z)^{-1} V_Z \Pi^* - \Pi^*. $$

To study the rate of convergence of the MSE, we will study the rates of the three terms (12), (13), and (14).
Proposition 2 Assume Assumptions 1 to 4 hold.

If $\beta > 1$, then $\text{MSE} = O_p \left( \frac{1}{n\alpha} + \alpha^{\beta/2} \right)$. If $\beta < 1$, then $\text{MSE} = O_p \left( \frac{\alpha^{\beta/2}}{n\alpha} + \alpha^{2\beta/2} \right)$.

Proposition 2 shows that the MSE exhibits the usual trade-off between the variance decreasing in $\alpha$ and the bias increasing in $\alpha$.

4 Asymptotic normality for fixed $\alpha$ and tests

Assumption 5. $(U_i, Z_i)$ are iid and $E(U_i|Z_i) = 0$.

Under Assumption 5 and some extra moment conditions (see Dauxois, Pousse, and Romain (1982) and Mas (2006)), we have

$$\sqrt{n} \left( \hat{V}_Z - V_Z \right) \xrightarrow{d} \mathcal{N}(0, K_Z),$$

$$\sqrt{n} \hat{C}_{ZU} \xrightarrow{d} \mathcal{N}(0, K_{ZU})$$

where $K_Z$ and $K_{ZU}$ are covariance operators and the convergence is either in the space of Hilbert-space operators (Dauxois et al. 1982) or in the space of trace-class operators (Mas, 2006). Moreover, $\sqrt{n} \left( \hat{V}_Z - V_Z \right)$ and $\sqrt{n} \hat{C}_{ZU}$ are asymptotically independent.

In this section, we consider the case where $\alpha$ is fixed. In that case, $\Pi_\alpha^*$ is not consistent and keeps an asymptotic bias. It is useful to define $\Pi_\alpha^*$ the regularized version of $\Pi^*$:

$$\Pi_\alpha^* = (\alpha I + V_Z)^{-1} V_Z \Pi^*.$$
We have

\[ \hat{\Pi}_\alpha^* - \Pi_\alpha^* = \left( \alpha I + \hat{V}_Z \right)^{-1} \hat{C}_{ZU} \]

\[ + \left( \alpha I + \hat{V}_Z \right)^{-1} \hat{V}_Z \Pi^* - (\alpha I + V_Z)^{-1} V_Z \Pi^* \]

\[ = (\alpha I + V_Z)^{-1} \hat{C}_{ZU} \]

\[ + \left[ \left( \alpha I + \hat{V}_Z \right)^{-1} - (\alpha I + V_Z)^{-1} \right] \hat{C}_{ZU} \]

\[ + \alpha \left( \alpha I + \hat{V}_Z \right)^{-1} \left( V_Z - \hat{V}_Z \right) (\alpha I + V_Z)^{-1} \Pi^* \]

\[ = (\alpha I + V_Z)^{-1} \hat{C}_{ZU} \]

\[ + \alpha (\alpha I + V_Z)^{-1} \left( \hat{V}_Z - V_Z \right) (\alpha I + V_Z)^{-1} \Pi^* \] \hspace{1cm} (15)

\[ + O_p \left( \frac{1}{n} \right) . \]

As \( n \) goes to infinity, \( \hat{\Pi}_\alpha^* - \Pi_\alpha^* \) converges to zero and is \( \sqrt{n} \)--asymptotically normal. The first two terms of the r.h.s are \( O_p \left( 1/\sqrt{n} \right) \) and will affect the asymptotic distribution. This distribution is not simple. We are going to characterize it below.

From Equations (15) and (16), neglecting the \( O_p(1/n) \) term, we have

\[ \hat{\Pi}_\alpha^* - \Pi_\alpha^* = (\alpha I + V_Z)^{-1} \hat{C}_{ZU} + \alpha (\alpha I + V_Z)^{-1} (\hat{V}_Z - V_Z)(\alpha I + V_Z)^{-1} \Pi^* \]

\[ = (\alpha I + V_Z)^{-1} \frac{1}{n} \sum_i (u_i \otimes z_i) + \alpha (\alpha I + V_Z)^{-1} \frac{1}{n} \sum_i (z_i \otimes z_i - V_Z)(\alpha I + V_Z)^{-1} \Pi^* \]

\[ = \frac{1}{n} \sum_i \left( u_i \otimes (\alpha I + V_Z)^{-1} z_i + \alpha \Pi (\alpha I + V_Z)^{-1} z_i \otimes (\alpha I + V_Z)^{-1} z_i \right) \]

\[ - \alpha (\alpha I + V_Z)^{-1} V_Z (\alpha I + V_Z)^{-1} \Pi^* \]

\[ = \frac{1}{n} \sum_i \left( u_i \otimes (\alpha I + V_Z)^{-1} z_i + \alpha \Pi (\alpha I + V_Z)^{-1} z_i \otimes (\alpha I + V_Z)^{-1} z_i \right) \]

\[ - \alpha \mathbb{E}[\Pi (\alpha I + V_Z)^{-1} Z \otimes (\alpha I + V_Z)^{-1} Z] \]

\[ = \frac{1}{n} \sum_i \left( u_i \otimes \tilde{z}_i + \alpha \Pi \tilde{z}_i \otimes \tilde{z}_i - \alpha \mathbb{E}[\Pi \tilde{Z} \otimes \tilde{Z}] \right) \]

\[ = \frac{1}{n} \sum_i \left( (u_i + \alpha \Pi \tilde{z}_i) \otimes \tilde{z}_i - \alpha C_{\tilde{Z} \Pi \tilde{Z}} \right). \]
where the first equality makes use of the definition of the empirical covariance operators using tensor products. The second line uses the elementary properties $K(Y \otimes X) = Y \otimes KX$ and $(Y \otimes X)K = K^*Y \otimes X$ for $X \in \mathcal{F}, X \in \mathcal{E}$ and $K \in \mathcal{H}$. The third line uses the definition of $V_Z$ and the fourth introduces the notation $\tilde{Z} \equiv (\alpha I + V_Z)^{-1}Z$. The interchange of the expectation operator and $(\alpha I + V_Z)^{-1}$ is allowed since the latter is a bounded linear operator, by Banach inverse theorem the inverse of a bounded linear operator is itself linear and bounded (see, for instance, Rudin, 1991). The last equality holds since the functional tensor product, denoted $\otimes$, distributes over addition.

The covariance operator of $\tilde{\Pi}_\alpha - \Pi_\alpha$ is an operator which maps the space of Hilbert-Schmidt operators from $\mathcal{E}$ to $\mathcal{F}$, denoted $\mathcal{H}$, into itself. Such an operator may be difficult to write explicitly. Fortunately, the properties of tensor products of infinite-dimensional Hilbert-Schmidt operators defined on separable Hilbert spaces are well-known,\(^1\) and may be used like in Dauxois, Pousse and Romain (1982) to write explicitly the covariance operator of an infinite-dimensional Hilbert-Schmidt random operator. The tensor product $\tilde{\Pi}_1 \tilde{\otimes} \Pi_2$ for $(\Pi_1, \Pi_2) \in \mathcal{H}^2$ is a mapping from $\mathcal{H}$ into itself, hence $\tilde{\Pi}_1 \tilde{\otimes} \Pi_2$ is an element of the Hilbert space of Hilbert-Schmidt operators from $\mathcal{H}$ to $\mathcal{H}$ equipped with the Hilbert-Schmidt inner product. For $T = \varphi \otimes \psi \in \mathcal{H}$, $\Pi_1 = X \otimes Z \in \mathcal{H}$ and $\Pi_2 = Y \otimes W \in \mathcal{H}$, this tensor product is equivalently defined as:

\[(i) \quad (\Pi_1 \tilde{\otimes} \Pi_2)T = \langle T, \Pi_1 \rangle_{\mathcal{H}} \Pi_2 \in \mathcal{H},\]
\[(ii) \quad \left((X \otimes Z) \tilde{\otimes} (Y \otimes W)\right)(\varphi \otimes \psi) = \left((X \otimes Y)\varphi\right) \otimes \left((Z \otimes W)\psi\right),\]
\[\forall \varphi, X, Y \in \mathcal{F}, \psi, Z, W \in \mathcal{E},\]

Based upon definition (i), the covariance operator of $\Pi_1$ and $\Pi_2$ naturally writes as

\[E \left[ \langle \cdot, \Pi_1 - E[\Pi_1] \rangle_{\mathcal{H}} (\Pi_2 - E[\Pi_2]) \right] = E \left[ (\Pi_1 - E[\Pi_1]) \tilde{\otimes} (\Pi_2 - E[\Pi_2]) \right]. \quad (17)\]

Furthermore, to show asymptotic normality we shall use the classical central limit theorem for i.i.d. processes in separable Hilbert spaces. The following is stated as Theorem 2.7 in Bosq (2000) and is reproduced here for clarity.

\(^1\)See, for instance, Vilenkin (1968, p.59-65).
Theorem 3 (Bosq, 2000) Let \((Z_i, i \geq 1)\) be a sequence of i.i.d. \(\mathcal{F}\)-valued random variables, where \(\mathcal{F}\) is a separable Hilbert space, such that \(E\|Z_i\|^2 < \infty\), \(E(Z_i) = \bar{Z}\) and \(V_Z = V\), then one has

\[
\frac{1}{\sqrt{n}} \sum_i (Z_i - \bar{Z}) \overset{d}{\to} \mathcal{N}(0, V),
\]

We are now geared to derive the asymptotic covariance operator of interest in its general form, under some standard assumptions.

Proposition 4 Assume \((U_i, Z_i)\) i.i.d., \(\Omega_{\alpha} < \infty\), \(E\|Z_i\|^4 < \infty\), \(E\|U_i\|^2\|Z_i\|^2 < \infty\), then

\[
\sqrt{n}(\hat{\Pi}_{\alpha}^* - \Pi_{\alpha}^*) \overset{d}{\to} \mathcal{N}(0, \Omega_{\alpha}),
\]

where the asymptotic covariance operator\(^2\) \(\Omega_{\alpha}\) for fixed \(\alpha\) is given by

\[
\Omega_{\alpha} = E\left[\left((U + \alpha \Pi \bar{Z}) \otimes \bar{Z}\right) \otimes \left((U + \alpha \Pi \bar{Z}) \otimes \bar{Z}\right)\right] - \alpha^2 C_{\Pi \bar{Z}} \otimes C_{\Pi \bar{Z}},
\]

which simplifies to

\[
\Omega_0 = E\left[(U \otimes V^{-1}_Z Z) \otimes (U \otimes V^{-1}_Z Z)\right],
\]

when \(\alpha \to 0\).

Proof. Under the assumptions \((U_i, Z_i)\) i.i.d., \(E\|Z_i\|^4 < \infty\), and \(E\|U_i\|^2\|Z_i\|^2 < \infty\), Theorem 3 ensures the root-n asymptotic normality of \(\frac{1}{\sqrt{n}} \sum_i \left( \frac{U_i \otimes Z_i}{Z_i \otimes Z_i} \right)\). By the continuous mapping theorem, one has (18) using the continuous transformation \(\begin{pmatrix} A \\ B \end{pmatrix} \mapsto (\alpha I + V_Z)^{-1} A + \alpha (\alpha I + V_Z)^{-1} B (\alpha I + V_Z)^{-1} \Pi^*\). The covariance operator of \(\sqrt{n}(\hat{\Pi}_{\alpha}^* - \Pi_{\alpha}^*)\)

\(^2\)The kernel of \(\Omega_{\alpha}\) has four dimensions and its kernel may be written as

\[
\omega_{\alpha}(s, t, r, \tau) = E\left[(U(s) + \alpha \Pi \bar{Z}(s)) Z(t) (U(r) + \alpha \Pi \bar{Z}(r)) Z(\tau)\right] - \alpha^2 E\left[\Pi \bar{Z}(s) \bar{Z}(t)\right] E\left[\Pi \bar{Z}(r) \bar{Z}(\tau)\right].
\]
may be written using (17) as

$$
\Omega_\alpha = E \left[ \left( \frac{1}{\sqrt{n}} \sum_i (u_i + \alpha \Pi \tilde{z}_i) \otimes \tilde{z}_i - \alpha C \tilde{z}_i \Pi \right) \otimes \left( \frac{1}{\sqrt{n}} \sum_i (u_i + \alpha \Pi \tilde{z}_i) \otimes \tilde{z}_i - \alpha C \tilde{z}_i \Pi \right) \right]
$$

$$
= E \left[ \left( (U + \alpha \Pi \tilde{Z}) \otimes \tilde{Z} - \alpha C \tilde{Z} \Pi \right) \otimes \left( (U + \alpha \Pi \tilde{Z}) \otimes \tilde{Z} - \alpha C \tilde{Z} \Pi \right) \right],
$$

where the second line is obtained from the i.i.d. assumption. Straightforward developments yield (19). Now letting $\alpha \to 0$ gives (20).

Furthermore, under the strict exogeneity assumption, $E[U_i|Z_i] = 0$, the asymptotic covariance operator in (19) is simplified into

$$
\Omega_\alpha = E \left[ \left( U \otimes \tilde{Z} \right) \otimes \left( U \otimes \tilde{Z} \right) \right] + \alpha^2 E \left[ \left( \Pi \tilde{Z} \otimes \tilde{Z} \right) \otimes \left( \Pi \tilde{Z} \otimes \tilde{Z} \right) \right] - \alpha^2 C \tilde{Z} \Pi \otimes C \tilde{Z} \Pi.
$$

In econometrics, we are often interested in testing the significance of estimates and produce confidence bands. However, there is no obvious meaningful way to perform standard significance tests using the derived asymptotic covariance. Indeed, for fixed $\alpha$ the estimated residuals will be biased and one must specify $\Pi^*$. On the other hand, if we assume $\alpha \to 0$, an estimator of (20) may be uninformative since $V^{-1}_Z$ does not necessarily exist. A more practical approach would be to keep $\alpha$ fixed to obtain an estimate of $(\alpha I + V_Z)^{-1}$ and use it to derive an estimator of (20). Other statistical tests may involve applying a test operator to $\Omega_\alpha$.

We want to test the null hypothesis: $H_0 : \Pi = \Pi_0$ where $\Pi_0$ is known. A simple way to test this hypothesis is to look at $\hat{C}_{ZY} - \hat{V}_Z \Pi_0^*$. Under $H_0$, this operator equals $\hat{C}_{ZU}$ and should be close to zero. Moreover, under $H_0$,

$$
\sqrt{n} \left( \hat{C}_{ZY} - \hat{V}_Z \Pi_0^* \right) \overset{d}{\to} N(0, K_{ZU})
$$

where

$$
K_{ZU} = E \left[ (u \otimes Z) \otimes (u \otimes Z) \right]
$$

and $(x \otimes y)(f) = \langle x, f \rangle y$ and $(\Pi_1 \otimes \Pi_2) T = \langle T, \Pi_1^* \rangle_{H_1} \Pi_2$ (see Dauxois, Pousse, and Romain, 1982)
Let \( \{\phi_j, \psi_j : j = 1, 2, ..., q\} \) be a set of test functions, then

\[
\begin{bmatrix}
\sqrt{n} \langle (\hat{C}_{ZY} - \hat{V}_Z \Pi_0^*) \phi_1, \psi_1 \rangle \\
\vdots \\
\sqrt{n} \langle (\hat{C}_{ZY} - \hat{V}_Z \Pi_0^*) \phi_q, \psi_q \rangle
\end{bmatrix}
\]

converges to a multivariate normal distribution with mean \( 0_q \) and covariance matrix the \( q \times q \) matrix \( \Sigma \) with \((j,l)\) element:

\[
\Sigma_{jl} = E \left[ \langle \sqrt{n} \hat{C}_{ZU} \phi_j, \psi_j \rangle \langle \sqrt{n} \hat{C}_{ZU} \phi_l, \psi_l \rangle \right] = \langle \phi_j, V_Z \psi_l \rangle \langle \phi_l, V_U \psi_l \rangle.
\]

This covariance matrix can be easily estimated by replacing \( V_Z \) and \( V_U \) by their sample counterpart. The appropriately rescaled quadratic form converges to a chi-square distribution with \( q \) degrees of freedom which can be used to test \( H_0 \). The test functions could be cumulative normals as in Conley, Hansen, Luttmer, and Scheinkman (1997) or could be normal densities with same small variance but centered at different means.

5 Data-driven selection of \( \alpha \)

The estimator involves a tuning parameter, \( \alpha \), which needs to be selected. It can be chosen as the solution to

\[
\min_{\alpha} \frac{1}{\alpha} \left\| \hat{V}_Z \hat{\Pi}_\alpha^* - \hat{C}_{ZY} \right\|_{HS}^2.
\]

See Engl, Hanke, and Neubauer (2000, p.102).

Another possibility is to use leave-one-out cross-validation

\[
\min_{\alpha} \frac{1}{n} \sum_j \left\| y_i - \hat{\Pi}_\alpha^{(-i)} z_i \right\|^2
\]

where \( \hat{\Pi}_\alpha^{(-i)} \) has been computed using all observations except for the \( i \)th one. Centorrino (2014) studies the properties of the leave-one-out cross-validation for nonparametric IV regression and shows that this criterion is rate optimal in mean squared error. This method is also used in a binary response model by Centorrino and Florens (2014).
Various data-driven selection techniques are compared via simulations in Centorrino, Féve, and Florens (2013).

An alternative approach would be to use a penalized minimum contrast criterion as in Goldenshluger and Lepski (2011). This could lead to a minimax-optimal estimator (Comte and Johannes, 2012).

6 Discrete observations

In this section, to simplify the exposition, we will refer to the arguments of \((y_i, z_i), t,\) as time even though it could refer to a location or other characteristic. Suppose that the data \((y_i, z_i)\) are not observed in continuous time but at discrete (not necessarily equally spaced) times. We use some smoothing to construct pairs of curves \((y_i^m, z_i^m),\) \(i = 1, 2, ..., n\) such that \(y_i^m \in \mathcal{E}\) and \(z_i^m \in \mathcal{F}.\) This smoothing can be obtained by approximating the curves by step functions or kernel smoothing for instance. The subscript \(m\) corresponds to the smallest number of discrete observations across \(i = 1, 2, ..., n.\) \(m\) grows with the sample size \(n.\)

Using the smoothed observations, we compute the corresponding estimators of \(V_Z\) and \(C_{ZY}\) denoted \(\hat{V}_Z^m, \hat{C}_{ZY}^m\) and the estimator of \(\Pi^*\) denoted \(\hat{\Pi}^{m*}_\alpha:\)

\[
\hat{\Pi}^{m*}_\alpha = \left(\alpha I + \hat{V}_Z^m\right)^{-1} \hat{C}_{ZY}^m.
\]

To assess the rate of convergence of \(\hat{\Pi}^{m*}_\alpha,\) we add the following conditions which guarantee that the discretization error is negligible with respect to the estimation error.

Assumption 6. \(||z_i^m - z_i|| = O_p(f(m))\) and \(||y_i^m - y_i|| = O_p(f(m)).\)

Assumption 7.

\[
\frac{f(m)}{\alpha n} = o\left(\alpha^{\beta/2}\right).
\]

Proposition 5 Under Assumptions 1 to 4, 6, and 7, the MSE of \(\hat{\Pi}^{m*}_\alpha - \Pi^*\) has the same rate of convergence as that of the MSE of \(\hat{\Pi}_\alpha - \Pi^*\) in Proposition 2.
7 Case where Z is endogenous

Now, assume Z is endogenous but we observe instrumental variables W such that cov(U, W) = 0. Hence, E ((Y - ΠZ) (W, .)) = 0. It follows that

\[ C_{YW} = ΠC_{ZW} \] (21)

where \( C_{YW} = E (Y \langle W, . \rangle) \) and \( C_{ZW} = E (Z \langle W, . \rangle) \). Similarly, we have

\[ C_{WY} = C_{WZ}Π^* \] (22)

where \( C_{WZ} = E (W \langle Z, . \rangle) \).

We need the following identification conditions:

Assumption 8. \( C_{WZ} \) is injective.

Under this assumption, Π is uniquely defined from (21). To see this, assume that there are two solutions Π₁ and Π₂ to (21). It follows that \((Π₁ - Π₂) C_{ZW} = 0\) or equivalently \(C_{WZ} (Π₁^* - Π₂^*) = 0\). Hence the range of \((Π₁^* - Π₂^*)\) belongs to the null space of \(C_{WZ}\). However, under Assumption 6, the null space of \(C_{WZ}\) is reduced to zero and thus the range of \((Π₁^* - Π₂^*)\) is equal to zero. It follows that \(Π₁^*φ - Π₂^*φ = 0\) for all φ, hence \(Π₁^* = Π₂^*\).

To construct an estimator of \(Π^*\), we first apply the operator \(C_{ZW}\) on the l.h.s and r.h.s of Equation (22) to obtain

\[ C_{ZW}C_{WY} = C_{ZW}C_{WZ}Π^*. \]

Note that \(C_{ZW} = C_{WZ}^*\) and therefore the operator \(C_{ZW}C_{WZ}\) is self-adjoint. The operators \(C_{ZW}, C_{WZ},\) and \(C_{WY}\) can be estimated by their sample counterparts. The estimator of \(Π^*\) is defined by

\[ \hat{Π}_α^* = \left(αI + \hat{C}_{ZW}\hat{C}_{WZ}\right)^{-1}\hat{C}_{ZW}\hat{C}_{WY}. \] (23)

Similarly, the estimator of Π is given by

\[ \hat{Π}_α = \hat{C}_{YW}\hat{C}_{WZ} \left(αI + \hat{C}_{ZW}\hat{C}_{WZ}\right)^{-1}. \]
Now, we explain how to compute \( \hat{\Pi}_\alpha^* \) is practice. From (23), we have

\[
\left( \alpha I + \hat{\mathcal{C}}_{ZW} \hat{\mathcal{C}}_{WZ} \right) \hat{\Pi}_\alpha^* \psi = \hat{\mathcal{C}}_{ZW} \hat{\mathcal{C}}_{WY} \psi.
\]

Note that

\[
\hat{\mathcal{C}}_{ZW} \hat{\mathcal{C}}_{WY} \psi = \frac{1}{n^2} \sum_{i,j} \langle y_j, \psi \rangle \langle w_i, w_j \rangle z_i,
\]

\[
\hat{\mathcal{C}}_{ZW} \hat{\mathcal{C}}_{WZ} \hat{\Pi}_\alpha^* \psi = \frac{1}{n^2} \sum_{i,j} \langle z_j, \hat{\Pi}_\alpha^* \psi \rangle \langle w_i, w_j \rangle z_i.
\]

Taking the inner product with \( z_l \) yields \( n \) equations

\[
\alpha \left( \langle z_l, \hat{\Pi}_\alpha^* \psi \rangle \right) + \frac{1}{n^2} \sum_{i,j} \langle z_j, \hat{\Pi}_\alpha^* \psi \rangle \langle w_i, w_j \rangle \langle z_l, z_i \rangle
\]

\[
= \frac{1}{n^2} \sum_{i,j} \langle z_j, \hat{\Pi}_\alpha^* \psi \rangle \langle w_i, w_j \rangle \langle z_l, z_i \rangle, \quad l = 1, 2, \ldots, n
\]

with \( n \) unknowns \( \langle z_j, \hat{\Pi}_\alpha^* \psi \rangle, \quad j = 1, 2, \ldots, n \). Then, for each \( \psi, \hat{\Pi}_\alpha^* \psi \) can be computed from

\[
\hat{\Pi}_\alpha^* \psi = \frac{1}{\alpha} \left[ \hat{\mathcal{C}}_{ZW} \hat{\mathcal{C}}_{WY} \psi - \hat{\mathcal{C}}_{ZW} \hat{\mathcal{C}}_{WZ} \hat{\Pi}_\alpha^* \psi \right].
\]

The computation of \( \hat{\Pi}_\alpha \varphi \) can be done using the same approach as in Section 2.

Assumption 9. \( C_{ZW} C_{WZ} \) is a trace-class operator and \( \left\| \hat{\mathcal{C}}_{ZW} \hat{\mathcal{C}}_{WZ} - C_{ZW} C_{WZ} \right\|_{HS}^2 = O_p (1/n) \).

Assumption 10. There is a Hilbert-Schmidt operator \( R \) from \( \mathcal{E} \) to \( \mathcal{F} \) and a constant \( \beta > 0 \) such that \( \Pi^* = (C_{ZW} C_{WZ})^{\beta/2} R \).

We decompose \( \hat{\Pi}_\alpha^* - \Pi^* \) in the following manner:

\[
\hat{\Pi}_\alpha^* - \Pi^* = \left( \alpha I + \hat{\mathcal{C}}_{ZW} \hat{\mathcal{C}}_{WZ} \right)^{-1} \hat{\mathcal{C}}_{ZW} \hat{\mathcal{C}}_{WY} - \Pi^* \quad (24)
\]

\[
= \left( \alpha I + \hat{\mathcal{C}}_{ZW} \hat{\mathcal{C}}_{WZ} \right)^{-1} \hat{\mathcal{C}}_{ZW} \hat{\mathcal{C}}_{WU} \quad (25)
\]

\[
+ \left( \alpha I + \hat{\mathcal{C}}_{ZW} \hat{\mathcal{C}}_{WZ} \right)^{-1} \hat{\mathcal{C}}_{ZW} \hat{\mathcal{C}}_{WZ} \Pi^* - (\alpha I + C_{ZW} C_{WZ})^{-1} C_{ZW} C_{WZ} \Pi^* \quad (26)
\]

\[
+ (\alpha I + C_{ZW} C_{WZ})^{-1} C_{ZW} C_{WZ} \Pi^* - \Pi^*. \quad (27)
\]
Proposition 6 Under Assumptions 1, 2, 8, 9, and 10, the MSE of $\hat{\Pi}_n^* - \Pi^*$ has the same rate of convergence as in Proposition 2.

8 Simulations

This section consists of a simulation study of the estimator presented earlier. Let $\mathcal{E} = \mathcal{F} = L^2[0,1]$ and $\mathcal{S} = \mathcal{T} = [0,1]$. $\Pi$ is an integral operator from to $L^2[0,1]$ to $L^2[0,1]$ with kernel $\pi(s,t) = 1 - |s-t|^2$.\(^3\) We consider an Ornstein-Uhlenbeck process with zero mean and mean reversion rate equal to one to represent the error function. It is described by the differential equation $dU(s) = -U(s)ds + \sigma_u dG_u(s)$, for $s \in [0,1]$ and where $G_u$ is a Wiener process and $\sigma_u$ denotes the standard deviation of its increments $dG_u$. Note that this error function is stationary.

We study the model $Y_i = \Pi Z_i + U_i, \quad i = 1, ..., n$

in two different settings. First, we consider design functions uncorrelated to the error functions ($\text{cov}(U,Z) = 0$), then investigate the case where $Z$ is endogenous ($\text{cov}(U,Z) \neq 0$).

8.1 Exogenous predictor functions

We consider the design function

$Z_i(t) = \frac{\Gamma(\alpha_i + \beta_i)}{\Gamma(\alpha_i) + \Gamma(\beta_i)} t^{\alpha_i-1}(1-t)^{\beta_i-1} + \eta_i$

for $t \in [0,1]$, with $\alpha_i, \beta_i \sim iid U[2,5]$ and $\eta_i \sim iid N(0,1)$, for all $i = 1, ..., n$. These predictor functions are probability density functions of some random beta distributions over the interval $[0,1]$, with an additive gaussian term.

The numerical simulation is performed as follows:

1. Construct both a pseudo-continuous interval of $[0,1]$, denoted $\mathcal{T}$, consisting of 1000 equally-spaced discrete steps, and a discretized interval of $[0,1]$, denoted $\tilde{\mathcal{T}}$, consisting of only 100 equally-spaced discrete steps.

\(^3\)Simulations have also been performed using different kernels. In particular, we have considered multiple kernels, allowing to include multiple functional predictors in a single functional model. Results suggest that the performance of the estimator is analogous in "multivariate" functional linear regression.
2. Generate $n$ predictor functions $z_i(t)$ and error functions $u_i(s)$, where $t, s \in T$ so as to obtain pseudo-continuous functions.

3. Generate the $n$ response functions $y_i(s)$ using the specified model where $s \in T$.

4. Generate the sample of $n$ discretized pairs of functions $(\tilde{z}_i, \tilde{y}_i)$ by extracting the corresponding values of the pairs $(z_i, y_i)$ for all $t, s \in \tilde{T}$.

5. Estimate $\Pi$ using the regularization method on the sample of $n$ pairs of functions $(\tilde{z}_i, \tilde{y}_i)$ and a fixed smoothing parameter $\alpha = .01$.

6. Repeat steps 2-5 100 times and calculate the $MSE$ by averaging the quantities

$$
||\hat{\Pi}_\alpha - \Pi||^2_{HS} = \int \int (\hat{\pi}_\alpha(s, t) - \pi(s, t))^2 dt ds
$$

over all repetitions.

All numerical integrations are performed using the trapezoidal rule (i.e. piecewise linear interpolation) although it is possible to use other quadrature rules (such as another Newton–Cotes rule or adaptive quadrature). In addition, the simulations of the stochastic processes for the error terms are constructed using the Euler-Maruyama method for approximating numerical solutions to stochastic differential equations.

Figure 1 shows 10 discretized predictor functions $(z_i)$, Ornstein-Uhlenbeck error functions for $\sigma_u = 1 (u_i)$, response functions $(y_i)$ and an example of a response function for various values of $\sigma_u$.

Table 1 reports the $MSE$ for 4 different sample sizes ($n = 50, 100, 500, 1000$) and 5 values of the standard deviation parameter ($\sigma_u = 0.1, 0.25, 0.5, 1, 2$). Naturally, the use of a fixed smoothing parameter $\alpha = .01$ that is independent of the sample size prevents the $MSE$ from converging towards zero. In fact, the $MSE$ converges to $||\Pi - \Pi_\alpha||^2_{HS}$, which is a measure of the squared bias introduced by the regularization method. The last two columns of Table 1 report the true global ($R^2$) and extended local ($\tilde{R}^2$) functional

---

4In practice, the nature of the functions of interest should provide guidance for the researcher with regards to the selection of the appropriate integration method. As we study square integrable functions in this setup, the trapezoidal rule allows reducing the discretization bias with respect to the rectangular rule.

5The magnitude of this bias depends on both the design functions and the value of $\alpha$ since $\Pi_\alpha = (\alpha I + V_2)^{-1}V_2\Pi$. We perform Monte-Carlo simulations to approximate the regularized operator $\Pi_\alpha$ using 100 random samples of 1000 $z_i$’s.
Figure 1: Examples of simulated functions (top left: discretized $y_i$; top right: discretized $u_i$ for $\sigma_U = 1$, bottom left: discretized $z_i$, bottom right: a single $y_i$ for various $\sigma_u$ ).
coefficients of determination, defined as

$$R^2 = \frac{\int \text{var}(E[Y(s)|Z])ds}{\int \text{var}(Y(s))ds} = \frac{\int \text{var}(\Pi Z(s))ds}{\int \text{var}(Y(s))ds}$$

$$\tilde{R}^2 = \frac{\int \text{var}(E[Y(s)|Z])ds}{\int \text{var}(Y(s))ds} = \frac{\int \text{var}(\Pi Z(s))ds}{\int \text{var}(Y(s))ds},$$

which are directly related to those proposed in Yao, Muller and Wang (2005).\(^6\)

**Table 1: Simulation results: Mean-Square Errors over 100 replications**

<table>
<thead>
<tr>
<th>Std errors</th>
<th>Empirical MSE</th>
<th>Squared bias</th>
<th>Coef. of d.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n = 50</td>
<td>n = 100</td>
<td>n = 500</td>
</tr>
<tr>
<td>$\sigma_u = 0.1$</td>
<td>.0154</td>
<td>.0135</td>
<td>.0126</td>
</tr>
<tr>
<td></td>
<td>(.0027)</td>
<td>(.0017)</td>
<td>(.0008)</td>
</tr>
<tr>
<td>$\sigma_u = 0.25$</td>
<td>.0291</td>
<td>.0205</td>
<td>.0138</td>
</tr>
<tr>
<td></td>
<td>(.0098)</td>
<td>(.0063)</td>
<td>(.0022)</td>
</tr>
<tr>
<td>$\sigma_u = 0.5$</td>
<td>.0773</td>
<td>.0438</td>
<td>.0194</td>
</tr>
<tr>
<td></td>
<td>(.0363)</td>
<td>(.0193)</td>
<td>(.0057)</td>
</tr>
<tr>
<td>$\sigma_u = 1$</td>
<td>.2909</td>
<td>.1354</td>
<td>.0371</td>
</tr>
<tr>
<td></td>
<td>(.1789)</td>
<td>(.0659)</td>
<td>(.0161)</td>
</tr>
<tr>
<td>$\sigma_u = 2$</td>
<td>.9128</td>
<td>.4755</td>
<td>.1245</td>
</tr>
<tr>
<td></td>
<td>(.5495)</td>
<td>(.2607)</td>
<td>(.0660)</td>
</tr>
</tbody>
</table>

Note: Standard deviations are reported in parentheses.

Simulations results are in line with the theoretical results. We observe that, for a fixed $\alpha$, the $MSE$ decreases as the sample size grows. Further, the coefficients of determination decrease as the error function’s standard deviation parameter increases, since the estimation is made more difficult. As a result, the MSE grows with $\sigma_u$.

For illustration purposes, we provide two sets of surface plots. Figure 2 shows 3D-plots of the actual kernel (top-left), the regularized kernel (top-right), their superposition

\(^6\)These true coefficients are approximated by their mean values using 1000 random functions over 100 simulations. In practice (when the true $\Pi$ is unknown) it is possible to use a consistent estimators of those coefficients by using $\tilde{\Pi}_\alpha$ and the sample counterpart of variance operators.
Figure 2: True kernel vs. regularized kernel (top left: True; top right: Regularized, bottom left: True vs. regularized, bottom right: Bias). (bottom-left) and the bias computed as their difference (bottom-right). The Tikhonov regularization appears to introduce most of the bias on the edges of the kernel.

Figure 3 shows the mean estimated kernel for \( n = 500 \) and Ornstein-Uhlenbeck errors with \( \sigma_u = 1 \) (top-left), against the true kernel (bottom-left), against the regularized kernel (top-right), and its mean errors with respect to the true kernel (bottom-right). One may observe that the mean estimate is relatively close to the regularized kernel. However it does not perform well on the edges when compared to the true kernel.

Let us now turn to the case where \( Z \) is endogenous.
Figure 3: True kernel vs. mean estimate (100 runs with $n = 500$, $\sigma_u = 1$) (top left: Mean estimate, top right: Regularized vs. mean estimate, bottom left: True vs. mean estimate, bottom right: Mean errors)
8.2 Endogenous predictor functions

We consider the design function

\[ Z_i(t) = bW_i(t) + \xi_i(t), \]

where \( \xi_i(t) = aU_i(t) + c\varepsilon_i(t) \) and the instrument \( w_i \) is defined as

\[ W_i(t) = \frac{\Gamma(\alpha_i + \beta_i)}{\Gamma(\alpha_i) + \Gamma(\beta_i)} t^{\alpha_i-1}(1 - t)^{\beta_i-1} + \eta_i \]

for \( t \in [0, 1], \alpha_i, \beta_i \sim iid \mathcal{U}[2, 5] \) and \( \eta_i \sim iid \mathcal{N}(0, 1) \), for all \( i = 1, \ldots, n \). Moreover, \( U_i \) and \( \varepsilon_i \) are Ornstein-Uhlenbeck processes with standard deviation parameters \( \sigma_u = \sigma_\varepsilon = 1 \). It is easily shown that \( \xi_i \) is also an Ornstein-Uhlenbeck process with unit mean-reversion rate described by the differential equation

\[ d\xi(t) = -\xi(t) + \sqrt{a^2\sigma_u^2 + c^2\sigma_\varepsilon^2} dG(t). \]

We further assume \( a = 1, b \in [0, 1] \) and \( c \) such that \( \int_s^s \text{var}(Y(s))ds \) is unchanged as \( b \) varies.\(^7\) Hence, the choice of \( b \) amounts to that of the instrument’s strength.

The numerical simulation design is slightly modified so as to incorporate the generation of the instruments \( W \) and the dependence between \( Z \) and \( U \):

1. Construct both a pseudo-continuous interval of \([0, 1] \), denoted \( T \), consisting of 1000 equally-spaced discrete steps, and a discretized interval of \([0, 1] \), denoted \( \tilde{T} \), consisting of only 100 equally-spaced discrete steps.
2. Generate \( n \) instrument functions \( w_i(t) \) and error functions \( u_i(s) \) and \( \varepsilon_i(s) \), where \( t, s \in T \) so as to obtain pseudo-continuous functions.
3. Generate \( n \) predictor functions \( z_i(t) \) using the design specified above, where \( t, s \in T \) so as to obtain pseudo-continuous functions.
4. Generate the \( n \) response functions \( y_i(s) \) using the specified model where \( s \in T \).
5. Generate the sample of \( n \) discretized pairs of functions \( (\tilde{w}_i, \tilde{z}_i, \tilde{y}_i) \) by extracting the corresponding values of the pairs \( (w_i, z_i, y_i) \) for all \( t, s \in \tilde{T} \).

\(^7\)This assumption allows to keep the variance of \( Y \) stable when varying instrument strength. It implies \( c = \sqrt{1 + (1 - b^2) \int_{s}^{\text{var}(FW(s))ds}} \int_{s}^{\text{var}(FW(s))ds}. \)
6. Estimate $\Pi$ using the regularization method on the sample of $n$ triplets of functions $(\hat{w}_i, \hat{z}_i, \hat{y}_i)$ and a fixed smoothing parameter $\alpha = .01$.

7. Repeat steps 2-5 100 times and calculate the $MSE$ by averaging the quantities $\|\hat{\Pi}_\alpha - \Pi\|^2_{HS} = \int \int (\hat{\pi}_\alpha(s,t) - \pi(s,t))^2 dt ds$ over all repetitions.

Table 2 reports the $MSE$ for 4 different sample sizes ($n = 50, 100, 500, 1000$) and 4 values of $b$ when estimating the model without accounting for the endogeneity of $Z$. Unsurprisingly, the estimation errors are important. The squared bias is smaller to that of the previous design and decreases with $b$. The last two columns report $R^2$ and $\tilde{R}^2$ for the full model. They are relatively stable since $\int_s \var(Y(s)) ds$ is fixed.

<table>
<thead>
<tr>
<th>Table 2: Non-IV estimator: Mean-Square Errors over 100 replications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Instr. strength</td>
</tr>
<tr>
<td>------------------</td>
</tr>
<tr>
<td>$b = 0.25$</td>
</tr>
<tr>
<td>$(c = 2.3)$</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>$b = 0.5$</td>
</tr>
<tr>
<td>$(c = 1.96)$</td>
</tr>
<tr>
<td>$b = 0.75$</td>
</tr>
<tr>
<td>$(c = 1.55)$</td>
</tr>
<tr>
<td>$b = 1$</td>
</tr>
<tr>
<td>$(c = 1)$</td>
</tr>
</tbody>
</table>

Note: Standard deviations are reported in parentheses.

We now turn to the simulations results for the IV estimator. Table 3 reports the $MSE$'s along with $R^2$ and the squared regularization biases. Squared biases are fairly small in this setup. This is related to the covariance operator of the predictor functions. $R^2_{FS}$ denotes the first-stage regression’s coefficient of determination. It shows how $b$ relates to the instrument’s strength. Naturally, weaker instruments are associated with larger $MSE$’s, although the spread seems to vanish rather quickly in this setup.
Table 3: IV estimator: Mean-Square Errors over 100 replications

<table>
<thead>
<tr>
<th>Instr. str.</th>
<th>Empirical MSE</th>
<th>Squared bias</th>
<th>Coef. of d.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$n = 50$</td>
<td>$n = 100$</td>
<td>$n = 500$</td>
</tr>
<tr>
<td>$b = 0.25$</td>
<td>.2383</td>
<td>.1710</td>
<td>.0752</td>
</tr>
<tr>
<td>$b = 0.5$</td>
<td>.1040</td>
<td>.0619</td>
<td>.0315</td>
</tr>
<tr>
<td>$b = 0.75$</td>
<td>.0682</td>
<td>.0444</td>
<td>.0242</td>
</tr>
<tr>
<td>$b = 1$</td>
<td>.0466</td>
<td>.0330</td>
<td>.0211</td>
</tr>
<tr>
<td>$b = 1.5$</td>
<td>.0244</td>
<td>.0138</td>
<td>.0029</td>
</tr>
</tbody>
</table>

Note: Standard deviations are reported in parentheses.

For comparisons with the exogenous case, we provide a final set of surface plots. Figure 4 shows 3D-plots of the mean IV estimated kernel (top-left), the mean non-IV (top-right), the superposition of the mean IV and the true kernels (bottom-left) and the mean estimation errors computed as the difference between the true kernel and the mean IV estimate (bottom-right). Note that the mean IV estimate is relatively close to the actual kernel, whereas the estimate when neglecting endogeneity exhibits a large bias.

9 Application

9.1 Introduction

This section presents an empirical application of the functional linear regression model with functional response to the study of the dynamics between daily electricity consumption and temperature patterns. There is a tradition of applications to electricity data in functional data analysis (Ferraty and Vieu, 2003; Antoch et al., 2008; Andersson and Lillestol, 2010; Liebl, 2013), although particularly focused on statistical predictions. Rather, we propose an application illustrating the usefulness of our estimator in structural econometrics.

The growing deployment of smart-metering technologies in electricity systems creates a need for micro-econometric models able to capture the dynamic behaviors of end-users
Figure 4: True kernel vs. mean IV estimate (100 runs with $n = 500$, $\sigma_u = 1$ and $b = 0.75$) (top left: Mean estimated IV; top right: Mean estimated non-IV, bottom left: True vs. mean IV estimate, bottom right: Mean IV errors)
with respect to market fundamentals. Such models allow to take advantage of the large amount of data so as to provide new insights to practitioners and policymakers about the behavior of consumers. In particular, information with regards to the end-users’ behavior with respect to changes in weather or prices is valuable to local distribution companies since it can contribute to improve their demand-side bidding strategies in day-ahead markets, as proposed by Patrick and Wolak (1997).

We develop a discrete-time model with partial adjustment dynamics based on the literature on dynamic linear rational expectations (Muth, 1961; Hansen & Sargent, 1980). The model generates a decision rule that coincides with the functional regression of interest. We abstract from uncertainty and estimate the corresponding Euler equation using actual future values, as in Kennan (1979). The remaining of the section provides details on data construction before performing a preliminary data analysis. Finally, estimation results are interpreted using contour plots and report goodness-of-fit statistics.

9.2 The model

A representative forward-looking agent with preferences representable by a time-additive utility function is assumed to have perfect foresight over future weather realizations. At each period $s$, the agent chooses how much electrical heating $y_s$ to consume so as to minimize her discomfort in the next period, as well as her consumption level for other purposes, denoted $c_s$, which generates instantaneous utility gains only. The level of discomfort is defined as the quadratic difference between the desired heat stock $x_s^*$ and actual heat stock $x_s$. Considering the agent to live in a representative home, $x_s^*$ would be the ideal level of indoor temperature while $x_s$ would be actual indoor temperature. Following McLaughlin et al. (2011) and Yu et al. (2012), the evolution of $x_s$ is assumed to depend on the previous period outside air temperature $z_{s-1}$. We further assume that the heat stock is subject to time-dependent adjustment constraints. Those may be related to the disutility of abrupt temperature variations, or production capacity limits.

---

8It is also possible to estimate a model like $Y(s) = \int_0^s \pi_0(s,t)Z(t)dt + \int_0^s \pi_1(s,t)E[Z(t)|I_s]dt + U(s)$, which explicitly accounts for expectations, although it requires a significant departure from the original model presented in this paper.
on the heating system. The agent is considered to solve the dynamic recursive problem

\[ V_s(x_{s-1}, x_s) = \max_{y_s, c_s} \left\{ u_s(c_s) - \frac{1}{2}(x_s - x_s^*)^2 - \frac{\gamma_s}{2}(x_s - x_{s-1})^2 - p_s(c_s + y_s) + \beta V_{s+1}(x_s, x_{s+1}) \right\} \]

s.t. \( x_{s+1} - x_s = \delta(z_s - x_s) + \eta y_s \),

where \( u_s(.) \) is a strictly increasing and concave utility function associated with the consumption of \( c_s \). For the sake of simplicity and since electricity consumption is relatively price-inelastic, we neglect budget considerations and assume a constant price \( p_s = p \).

Moreover, the entire trajectories of \( z_s \) and \( x_s^* \) are known. The first order condition with respect to \( c_t \) gives the usual optimality condition

\[ u'_s(c_s) = p, \tag{28} \]

which states that the agent will consume electricity until its marginal utility equates the energy price. Making use of the first order condition with respect to \( q_t \) and the envelope theorem, one obtains the Euler equation which determines the optimal path for the heat stock given by

\[ x_{s+1} - \gamma_s + \beta(1 + \gamma_{s+1}) \frac{\beta}{\beta \gamma_{s+1}} x_s + \frac{\gamma_s}{\beta \gamma_{s+1}} x_{s-1} = \frac{\delta}{\eta \beta \gamma_{s+1}} p - \frac{x_s^*}{\gamma_{s+1}}. \tag{29} \]

This second-order difference equation has varying coefficients depending on the stringency of the adjustment constraint in both the current and future periods. Iterative substitutions of the expressions of past and future values of the heat stock yields a general solution of the form

\[ x(s) = \kappa_0(s) + \sum_{t=1}^{\infty} \kappa_1(s, t) x^*(t), \tag{30} \]

where \( \kappa_0(s) \) and \( \kappa_1(s, t) \) are aggregates of the model parameters, which crucially depend on the trajectory of adjustment costs. Let us define the desired electrical heating consumption level as \( y_s^* = \frac{1}{\eta}(x_{s+1}^* - (1 - \delta)x_s^* - \delta z_s) \). Therefore, multiplying both sides of (30) by \( (1 - (1 - \delta)L)L^{-1} \), where \( L \) is the lag operator, yields a general solution for \( y_s \).
given by

\[ y(s) = \kappa_0'(s) + \sum_{t=1}^{\infty} \kappa_1'(t) + \delta z(t)), \]  

(31)

which depends on outdoor temperature through its effect on the motion of the heat stock. Finally, specifying \( x_s^* = \xi_{1,s} + \xi_{2,s} z_s + \varepsilon_s \), where \( \xi_{1,s}, \xi_{2,s} \) are known functions of time and \( \varepsilon_s \) is a random component unobservable to the econometrician, we obtain \( \eta y_s^* = (\xi_{1,s+1} - (1 - \delta) \xi_{1,s}) + \xi_{2,s+1} x_s + (1 - \delta) z_s + \varepsilon_s + (1 - \delta) \varepsilon_s \). Substituting this expression in (31) allows to write the electrical heating consumption decision in terms of the path of temperature and a serially correlated error term \( \epsilon(s) \) as

\[ y(s) = \kappa_0''(s) + \sum_{t=1}^{\infty} \kappa_1''(s, t) z(t) + \epsilon(s), \]  

(32)

for any \( s \). Since \( y_s \) is hardly observable in practice, the estimation of the above equation appears difficult. Instead one can use (28) and specify the inverse marginal utility function, such as \( u_s^{-1}(p) = v(s) \forall p \). Now, if one only observes aggregate electricity consumption \( Y(s) = y_s + c_s \), each of the parameters \( \kappa_1''(s, t) \) can be estimated using

\[ Y(s) = \kappa_0''(s) + \sum_{t=1}^{\infty} \kappa_1''(s, t) z(t) + \epsilon(s), \]  

(33)

as long as \( v(s) \) is independent of outdoor temperature in all periods. The decision rule for the analogous continuous-time model is therefore of the form

\[ Y(s) = \pi_0(s) + \int_0^{\infty} \pi_1(s, t) z(t) dt + u(s). \]  

(34)

This model can be estimated using the procedure described above if one has an \( i.i.d. \) sample of observations \( \{Y_i, z_i\}_{i=1,...,n} \), and if \( u_i \) can be presumed a mean-zero \( i.i.d. \) random functional error process. We propose to estimate the model using daily patterns extracted from annual aggregate electricity consumption and temperature trajectories at the level of the province of Ontario. Unfortunately, this approach introduces correlation across observations by construction. A more suited data set would consist of measurements from individual households’ smart meters.
9.3 Preliminary data analysis

Prior to presenting the data set construction, let us present some facts about the electricity market in Ontario. There are two types of consumers. First, small consumers (residential end-users and small businesses) are billed for electricity usage by their local distribution company. The vast majority pays fixed time-of-use rates which are updated from season to season. Second, large consumers (large businesses and the public sector) are subject to the wholesale market price. The wholesale market price is determined on an hourly basis in a uniform-price multi-unit auction subject to operational constraints. It is considered as being quite volatile. Therefore, some large consumers choose to go with retail contractors to avoid market risk exposure, although the bulk of electricity trade goes through the wholesale market. Furthermore, all large consumers must also pay the monthly *Global Adjustment* which represents other charges related to market, transport and regulatory operations. Consequently, it is difficult to evaluate the extent to which aggregate electricity consumption depends on the wholesale price and the time-of-use price. In our theoretical model, we assumed the price to be constant since short-term demand for electricity is typically perceived as being very price inelastic.\(^9\) Note that it would be possible to relax this assumption by using the time-of-use prices and the wholesale price. We could estimate the price elasticity of the electricity demand using the instrumental variable version of the Tikhonov estimator. A natural instrument for the wholesale electricity price in Ontario is wind power production, which depreciates prices significantly and is as good as randomly assigned. We do not pursue this approach here.

The original data set consists of hourly observations of real-time aggregate electricity consumption and weighted average temperature in Ontario from January 1, 2010 to September 30, 2014. Hourly power data for Ontario are publicly available on the system operator’s website,\(^11\) whereas hourly province-wide temperature data have been constructed from hourly measurements at 77 weather stations in Ontario,\(^12\) publicly

---

\(^9\)Roughly speaking, businesses are considered large when their electricity bills exceed $2,000 per month.

\(^10\)For example, the recent implementation of time-of-use pricing has been proved to have had limited effects on electricity consumption in Ontario (Faruqui et al., 2013)

\(^11\)http://www.ieso.ca

\(^12\)The complete data set contains 139 weather stations although once matched to neighboring cities, only 77 are found relevant.
available on Environment Canada’s website.\textsuperscript{13}

Let $Z(t)$ denote our measure of temperature in hour $t$ for the entire province. It is constructed in three steps. First, we match a set of 41 Ontarian cities (of above 10,000 inhabitants)\textsuperscript{14} to their three nearest weather stations, then compute a weighted average using a distance metric. Finally, we obtain $Z(t)$ as a weighted average of cities’ temperatures, where weights are defined as each city’s relative population. The constructed province-wide hourly temperature variable is formally defined by

$$Z(t) = \sum_c \gamma_c \{ \sum_{w(c)} \rho_{w(c)} Z_{w(c)}(t) \}, \quad \forall i, h$$

where $\gamma_c = \frac{Pop_c}{\left( \sum_j Pop_j \right)}$ is city $c$’s weight, $\rho_{w(c)} = \frac{(\text{lat}_c - \text{lat}_{w(c)})^6 + (\text{lon}_c - \text{lon}_{w(c)})^6 - 1}{\sum_{l(c)} ((\text{lat}_c - \text{lat}_{l(c)})^6 + (\text{lon}_c - \text{lon}_{l(c)})^6)^{-1}}$ is station $w$’s weight for city $c$’s temperature average, $\text{lat}$ denotes latitude, $\text{lon}$ longitude, and $Z_{w(t)}$ is the temperature measurement at station $w$ in hour $t$.\textsuperscript{15} Finally, we use robust locally weighted polynomial regression on the constructed temperature series in order to smooth implausible jumps, which are most likely due to measurement errors. Table 1 reports descriptive statistics for hourly electricity consumption and our constructed measure of temperature. Unsurprisingly, we observe some correlation between annual consumption peaks and maximum temperatures.

\textsuperscript{13}\url{http://climat.meteo.gc.ca/}
\textsuperscript{14}Those cities represent 85.3\% of the province’s population as of 2011.
\textsuperscript{15}Distance is calculated as the difference in geographic coordinates to the sixth power. The exponent is chosen so as to put arbitrarily more weight on nearby weather stations with respect to those located further away from the cities.
The prices charged to individual consumers vary depending on the time of the day and the two main seasons (winter and the rest of the year) but follow a fixed pattern. The winter period spans from November 1 to April 30 and has two daily peak periods: one in the morning (7am-11am) and the other after worktime (5pm-7pm), a mid-peak period (11am-5pm) and an off-peak period (7pm-7am). The remaining part of the year is considered as the summer period and mid-peak and peak periods are reversed with respect to winter. Given this fact and the model developed above, dividing up the sample into two corresponding periods appears natural. We define the periods consequently as shown in Figure 5 (blue for winter and red for summer) though decide to discard the spring, from May 1 to May 31, and autumn, from Sep 15 to Oct 31, periods (black line) given the limited heating or and air conditioning needs in these periods. Figure 6 displays the relationship between electricity consumption and temperature in winter, summer and the discarded months.

The plots show evidence of a relatively linear relationship between the load and temperature for winter and summer months. On the other hand, the relation is much flatter in October and May. The figures also suggest that power usage is more sensitive to warm weather than cold weather. Probably because electricity represents a small
share of heating fuel for residential users in Ontario,\textsuperscript{16} or because cooling can be more energy-intensive than heating.

The functional data samples are finally constructed with daily trajectories of 25 discrete observations for the dependent variable and a three-day window of 73 observations for the predictor variable. This window is chosen so that the dependent will always be regressed on at least 24 lagged hours and 24 future hours. We discard weekends, Mondays and Fridays as well as holidays.\textsuperscript{17} For ease of interpretation, the temperature variable is transformed into heating and cooling degrees for each season using

\[
Z_h(t) = \max_{Z \in \text{Winter}} (Z(t)) - Z(t) \\
Z_c(t) = Z(t) - \min_{Z \in \text{Summer}} (Z(t)),
\]

where \( \min_{Z \in \text{Summer}} (Z(t)) = 6^\circ C \) and \( \max_{Z \in \text{Winter}} (Z(t)) = 19^\circ C \). The samples are presented in

\textsuperscript{16}Only 14\% as of 2011 according to Statistics Canada (2011)

\textsuperscript{17}In particular, Good Fridays and the holiday season for the winter period, and Canada day and Labour day for the summer period.
Figure 7 and Figure 8, the bold lines show the sample mean trajectories.

9.4 The functional model and estimation results

Based on the model developed above and previous data considerations, we consider $\mathcal{E} = L^2[0, 24]$ and $\mathcal{F} = L^2[-24, 48]$ with $S = [0, 24]$ and $T = [-24, 48]$. The functional linear regression model of interest is given by
\[ Y_i(s) = \pi_0(s) + \int_{-24.48} \pi_1(s, t)Z_i(t)dt + U_i(s), \quad (35) \]

where \( Y_i(s) \) is total electricity consumption in hour \( s \), \( \pi_0(s) \) is a constant function, \( Z_i(s) \) is temperature in hour \( s \in [-24, 48] \), and \( U_i(t) \) is a zero-mean error term. The object of interest is the kernel \( \pi_1 \), which characterizes the dynamic relation between electricity consumption for heating/AC needs and temperature patterns. Since we are interested in daily electricity consumption patterns, the cross-sectional unit \( i \) denotes a daily observation. We use Tuesdays, Wednesdays, and Thursdays. Using successive days may induce autocorrelation in the error which is not taken into account in our theory. However, we believe that our estimators are still consistent in this setting.

Figures 9 and 10 display contour plots of the estimated kernels using the samples detailed previously. The regularization parameter was chosen using leave-one-out cross validation. This data-driven selection criterion suggests \( \alpha = 1.4 \) for the winter sample and \( \alpha = 0.7 \) for the summer period.

In order to ease interpretation of the results, we plot indicative dashed lines to separate out the daily windows and add a diagonal so as to emphasize the contemporaneous relation between the functions of interest. One should remain cautious with regards to the interpretation of the estimated kernel. Estimated correlations may be read both horizontally and vertically. The effects of the entire temperature pattern on electricity consumption at a given hour of the day is observed horizontally, whereas the effect of temperature at a specific time upon the daily electricity consumption pattern is read vertically. The magnitudes of the correlation are indicated using colors from dark red to white with corresponding values given in the legend. The seemingly \( S \)-shape of the kernel around the indicative diagonal is likely to be an artifact of the regularization scheme, as observed in our simulations, and consequently, interpretative attempts of the results should account of this feature.

We observe from Figure 9 that the contemporaneous correlation between temperature in winter is the largest during the morning and evening peaks, whereas it is low between 10 am and 4 pm for the summer period. These high demand periods also seem to exhibit
Figure 9: Contour plots of estimation results for winter months with $\alpha = 1.4$

Figure 10: Contour plots of estimations results for summer months with $\alpha = 0.7$
some serial correlation with previous and future peaks in winter but not in summer. Generally, the effect of temperature on demand is found to be of larger magnitude in summer than in winter. Furthermore, correlations between current consumption and past and future temperatures appear to span more smoothly around the diagonal than in the winter period. Estimates could finally be used to recover the parameters of the structural model developed earlier and perform counterfactual analyses.

![Figure 11: Leave-one out cross validation criterion](image)

Figure 11 displays the leave-one-out cross-validation criterion as a function of $\alpha$. Figure 13 and Figure 12 display comparisons of predictions against actual trajectories for the year 2011.

Additionally, plots of the residuals and goodness-of-fit statistics are included in the appendix.\footnote{We find significant serial correlation in the residuals but no concern with regards to the stationarity of the series. Various specifications of the Augmented Dickey-Fuller tests were performed on the hourly residuals and all rejected the presence of a unit-root.} Figures 14 and 15 display $\hat{u}_i(t)/Y_i(t)$ where $i$ is on the $x$-axis. The 24 curves correspond to the 24 hours of the days, namely different values of $t$. The estimated
Figure 12: Actual vs Prediction (Summer)

Figure 13: Actual vs Prediction (Winter)
kernel provides an overall good fit of the series, as suggested by the large coefficients of determination of about .8 for both periods.

In conclusion, the estimation results suggest some correlations between electricity consumption and past, current and future temperatures. The correlation with future values may be due to anticipatory behavior of agents who look at the temperature forecast before deciding on whether to leave the heater on that day or not for instance. It may also be due to a measurement error in the temperature variable. Indeed, we measure the temperature as a weighted average over the province, so it is necessarily a noisy measure of the temperature experienced by each individual agent. In general, the estimated coefficients vary greatly across hours of the day and seasons. The model allows to obtain a good fit to the consumption data with temperature as a single predictor function. It would also be possible to include several predictor functions, each defined over possibly varying time-intervals. Furthermore, one could estimate a constrained version of the estimator by letting the time-interval to depend on s for instance. A direct adaptation of our model to this setup can easily be done by transforming predictor functions appropriately using binary variables. The study of the estimator’s predictive power as well as the introduction of a more thorough treatment of the temporal dimensions are left for future research.
A Appendix

Figure 14: Residuals (Winter)
Figure 15: Residuals (Summer)

Figure 16: Coefficients of determination
B Proofs

Proof of Proposition 1.

\[ \left\| \hat{\Pi}_\alpha \right\|_{HS}^2 = \left\| \hat{C}_{YZ} \left( \alpha I + \hat{V}_Z \right)^{-1} \right\|_{HS}^2 \]

\[ \leq \left\| \hat{C}_{YZ} \right\|_{HS}^2 \left\| \left( \alpha I + \hat{V}_Z \right)^{-1} \right\|_{op} \]

using the fact that, if \( A \) is a HS operator and \( B \) is a bounded operator, \( \|AB\|_{HS} \leq \|A\|_{HS} \|B\|_{op} \) where \( \|B\|_{op} \equiv \sup_{\|\phi\| \leq 1} \|B\phi\| \) is the operator norm. Then, we have

\[ \left\| \hat{\Pi}_\alpha \right\|_{HS}^2 \leq \frac{1}{\alpha} \left\| \hat{C}_{YZ} \right\|_{HS}^2 . \]

It remains to show that \( \hat{C}_{YZ} \) is a HS operator. \( \hat{C}_{YZ} \) is an integral operator with degenerate kernel \( \frac{1}{n} \sum_{i=1}^{n} y_i(s) z_i(t) \). A sufficient condition for \( \hat{C}_{YZ} \) to be HS is that its kernel is square integrable which is true because \( Y_i \) and \( Z_i \) are elements of Hilbert spaces. The result of Proposition 1 follows.

Proof of Proposition 2. To prove Proposition 2, we need two preliminary lemmas.

Lemma 7 Let \( A = B + C \) where \( B \) is a zero mean random operator and \( C \) is a non-random operator. Then,

\[ E \left( \|A\|_{HS}^2 \right) = E \left( \|B\|_{HS}^2 \right) + \|C\|_{HS}^2 . \]
Proof of Lemma 7.

\[ E(\|A\|_{HS}^2) = E \left( \sum_j \langle A\phi_j, A\phi_j \rangle \right) = E \left( \sum_j \langle A^*A\phi_j, \phi_j \rangle \right) = E \left( \sum_j \langle (B+C)^* (B+C) \phi_j, \phi_j \rangle \right) = E \left( \sum_j \langle B^*B\phi_j, \phi_j \rangle \right) + E \left( \sum_j \langle C^*B\phi_j, \phi_j \rangle \right) + E \left( \sum_j \langle B^*C\phi_j, \phi_j \rangle \right) + E \left( \sum_j \langle C^*C\phi_j, \phi_j \rangle \right). \]

The second and third terms on the r.h.s are equal to zero because \( E(B) = 0 \) and \( C \) is deterministic. We obtain \( E(\|A\|_{HS}^2) = E(\|B\|_{HS}^2) + \|C\|_{HS}^2 \).

**Lemma 8** Let \( A \) be a random operator from \( \mathcal{E} \) to \( \mathcal{F} \).

\[ E(\|A\|_{HS}^2) = trE(A^*A). \]

**Proof of Lemma 8.** We have

\[ E(\|A\|_{HS}^2) = E \left( \sum_j \langle A^*A\phi_j, \phi_j \rangle \right) = \sum_j \langle E(A^*A) \phi_j, \phi_j \rangle = trE(A^*A). \]

We turn to the proof of Proposition 2. Applying Lemma 7 on the decomposition
(12), (13), and (14), we have

\[
E \left( \left\| \hat{\Pi}_\alpha - \Pi \right\|_{HS}^2 \big| Z_1, Z_2, \ldots, Z_n \right) \\
= E \left( \| (12) \|_{HS}^2 \big| Z_1, Z_2, \ldots, Z_n \right) + \| (13) + (14) \|_{HS}^2 \\
\leq E \left( \| (12) \|_{HS}^2 \big| Z_1, Z_2, \ldots, Z_n \right) + 2 \| (13) \|_{HS}^2 + 2 \| (14) \|_{HS}^2.
\]

We study the first term of the r.h.s. By Lemma 8,

\[
E \left( \| (12) \|_{HS}^2 \big| Z_1, Z_2, \ldots, Z_n \right) \\
= E \left( \left\| \left( \alpha I + \hat{V}_Z \right)^{-1} \hat{C}_{ZU} \right\|_{HS}^2 \big| Z_1, Z_2, \ldots, Z_n \right) \\
= tr E \left( \left( \alpha I + \hat{V}_Z \right)^{-1} \hat{C}_{ZU} \hat{C}_{ZU}^* \left( \alpha I + \hat{V}_Z \right)^{-1} \big| Z_1, Z_2, \ldots, Z_n \right) \\
= tr \left\{ \left( \alpha I + \hat{V}_Z \right)^{-1} E \left( \hat{C}_{ZU} \hat{C}_{ZU}^* \big| Z_1, Z_2, \ldots, Z_n \right) \left( \alpha I + \hat{V}_Z \right)^{-1} \right\}.
\]

Note that

\[
\hat{C}_{ZU} \hat{C}_{ZU}^* \varphi = \frac{1}{n^2} \sum_{i,j} z_i \langle z_j, \varphi \rangle \langle u_i, u_j \rangle,
\]

\[
E \left( \hat{C}_{ZU} \hat{C}_{ZU}^* \varphi \big| Z_1, Z_2, \ldots, Z_n \right) = \frac{1}{n} \sum_i z_i \langle z_i, \varphi \rangle E \left[ \langle u_i, u_i \rangle \big| Z_1, Z_2, \ldots, Z_n \right] \\
= \frac{1}{n} \sum_i z_i \langle z_i, \varphi \rangle tr \left( V_U \right) \\
= \frac{1}{n} tr \left( V_U \right) \hat{V}_Z \varphi
\]

because the \( u_i \) are uncorrelated. To see that \( E \left[ \langle u, u \rangle \right] = tr V_U \), decompose \( u \) on the basis formed by the eigenfunctions \( \psi_j \) of \( V_U \) so that \( u = \sum_j \langle u, \psi_j \rangle \psi_j \). It follows that \( \langle u, u \rangle = \sum_j \langle u, \psi_j \rangle^2 \) and \( E \langle u, u \rangle = \sum_j \langle V_U \psi_j, \psi_j \rangle = tr \left( V_U \right) \). Hence,

\[
E \left( \| (12) \|_{HS}^2 \big| Z_1, Z_2, \ldots, Z_n \right) = \frac{1}{n} tr \left( V_U \right) tr \left( \left( \alpha I + \hat{V}_Z \right)^{-1} \hat{V}_Z \left( \alpha I + \hat{V}_Z \right)^{-1} \right) \\
\leq \frac{C}{n \alpha}
\]
where $C$ is a generic constant. It follows that $E \left( \| (12) \|_{HS}^2 \right) \leq \frac{C}{n\alpha}$.

Now, we turn toward the term (13). We have

$$
\left( (\alpha I + \hat{V}_Z)^{-1} \hat{V}_Z \Pi^* - (\alpha I + V_Z)^{-1} V_Z \Pi^* \right)
= \left[ - \left( \Pi - (\alpha I + \hat{V}_Z)^{-1} \hat{V}_Z \right) + (I - (\alpha I + V_Z)^{-1} V_Z) \right] \Pi^*.
$$

Using $I = (\alpha I + \hat{V}_Z)^{-1} (\alpha I + \hat{V}_Z)$, we obtain

$$
I - (\alpha I + \hat{V}_Z)^{-1} \hat{V}_Z = \alpha (\alpha I + \hat{V}_Z)^{-1}.
$$

Hence,

$$
(13) = \left[ -\alpha (\alpha I + \hat{V}_Z)^{-1} + \alpha (\alpha I + V_Z)^{-1} \right] \Pi^*
= (\alpha I + \hat{V}_Z)^{-1} (V_Z - \hat{V}_Z) \alpha (\alpha I + V_Z)^{-1} \Pi^*
$$

where the last equality follows from $A^{-1} - B^{-1} = A^{-1} (A - B) B^{-1}$.

Now, we have

$$
\left\| (\alpha I + \hat{V}_Z)^{-1} (V_Z - \hat{V}_Z) \alpha (\alpha I + V_Z)^{-1} \Pi^* \right\|_{HS}^2
\leq \left\| (\alpha I + \hat{V}_Z)^{-1} \right\|_{op}^2 \left\| (V_Z - \hat{V}_Z) \right\|_{op}^2 \left\| \alpha (\alpha I + V_Z)^{-1} \Pi^* \right\|_{HS}^2
$$

where $\left\| (\alpha I + \hat{V}_Z)^{-1} \right\|_{op}^2 \leq 1/\alpha^2$, $\left\| (V_Z - \hat{V}_Z) \right\|_{op}^2 = O_p(1/n)$ by Assumption 2 and $\left\| \alpha (\alpha I + V_Z)^{-1} \Pi^* \right\|_{HS}^2 = O \left( \alpha^{\beta n^2} \right)$.

If $\beta > 1$ then the term corresponding to (13) is negligible with respect to (12). If $\beta < 1$, then (12) is negligible with respect to (13).
Now, we turn our attention toward the term (14). We have

\[
(\alpha I + V_Z)^{-1} V_Z \Pi^* - \Pi^* \\
= (\alpha I + V_Z)^{-1} (V_Z - \alpha I - V_Z) \Pi^* \\
= \alpha (\alpha I + V_Z)^{-1} \Pi^* \\
= \alpha (\alpha I + V_Z)^{-1} V_Z^{\beta/2} R
\]

by Assumption 4. Let \( \{\lambda_j, \varphi_j\} \) be the eigenvalues and orthonormal eigenfunctions of \( V_Z \).

\[
\| \alpha (\alpha I + V_Z)^{-1} V_Z^{\beta/2} R \|_{HS}^2 \\
= \alpha^2 \sum_j \langle (\alpha I + V_Z)^{-1} V_Z^{\beta/2} R \varphi_j, (\alpha I + V_Z)^{-1} V_Z^{\beta/2} R \varphi_j \rangle \\
= \alpha^2 \sum_j \frac{\lambda_j^\beta}{(\lambda_j + \alpha)^2} \langle R \varphi_j, R \varphi_j \rangle^2 \\
\leq \alpha^2 \sup_\lambda \frac{\lambda^\beta}{(\lambda + \alpha)^2} \sum_j \langle R \varphi_j, R \varphi_j \rangle^2 \\
= O(\alpha^{\beta/2}).
\]

The last equality follows from the fact that \( \sum_j \langle R \varphi_j, R \varphi_j \rangle^2 = \| R \|_{HS}^2 < \infty \) and, using the notation \( \lambda = \mu^2 \), we have

\[
\sup_\lambda \frac{\alpha^2 \lambda^\beta}{(\lambda + \alpha)^2} = \sup_\mu \frac{\alpha^2 \mu^{2\beta}}{(\mu^2 + \alpha)^2} = O(\alpha^{\beta/2})
\]

by Carrasco, Florens, and Renault (2007, Proposition 3.11). Consequently,

\[
\| \alpha (\alpha I + V_Z)^{-1} V_Z^{\beta/2} R \|_{HS}^2 = O(\alpha^{\beta/2}).
\]

This concludes the proof of Proposition 2.

**Proof of Proposition 5.**

We have

\[
\hat{\Pi}_{\alpha}^{m*} - \Pi^* = \hat{\Pi}_{\alpha}^{m*} - \hat{\Pi}_{\alpha}^* + \hat{\Pi}_{\alpha}^* - \Pi^*.
\]
We focus on the term $\hat{\Pi}_{\alpha}^{m*} - \hat{\Pi}_{\alpha}^*$. 

\[
\hat{\Pi}_{\alpha}^{m*} - \hat{\Pi}_{\alpha}^* = \left( \alpha I + \hat{V}_Z^m \right)^{-1} \hat{C}_{ZY}^m - \left( \alpha I + \hat{V}_Z \right)^{-1} \hat{C}_{ZY} \\
= \left( \alpha I + \hat{V}_Z^m \right)^{-1} \left( \hat{C}_{ZY}^m - \hat{C}_{ZY} \right) + \left[ \left( \alpha I + \hat{V}_Z^m \right)^{-1} - \left( \alpha I + \hat{V}_Z \right)^{-1} \right] \hat{C}_{ZY}.
\]

\[
\left\| \hat{C}_{ZY}^m - \hat{C}_{ZY} \right\|_{HS}^2 = \left\| \frac{1}{n} \sum_{i=1}^{n} z_i^m \langle y_i^m, \cdot \rangle - \frac{1}{n} \sum_{i=1}^{n} z_i \langle y_i, \cdot \rangle \right\|_{HS}^2 \\
= \left\| \frac{1}{n} \sum_{i=1}^{n} \left\{ (z_i^m - z_i) \langle y_i, \cdot \rangle + z_i^m \langle y_i^m - y_i, \cdot \rangle \right\} \right\|_{HS}^2 \\
\leq \frac{2}{n^2} \sum_{i=1}^{n} \left\{ \left\| (z_i^m - z_i) \langle y_i, \cdot \rangle \right\|_{HS}^2 + \left\| z_i^m \langle y_i^m - y_i, \cdot \rangle \right\|_{HS}^2 \right\}.
\]

\[
\left\| (z_i^m - z_i) \langle y_i, \cdot \rangle \right\|_{HS}^2 = \sum_j \left\langle (z_i^m - z_i) \langle y_i, \phi_j \rangle, (z_i^m - z_i) \langle y_i, \phi_j \rangle \right\rangle \\
= \| z_i^m - z_i \|^2 \sum_j \langle y_i, \phi_j \rangle^2 \\
= O_p \left( f(m)^2 \right).
\]

\[
\left\| z_i^m \langle y_i^m - y_i, \cdot \rangle \right\|_{HS}^2 = \sum_j \left\langle z_i^m \langle y_i^m - y_i, \phi_j \rangle, z_i^m \langle y_i^m - y_i, \phi_j \rangle \right\rangle \\
= \| z_i^m \|^2 \sum_j \langle y_i^m - y_i, \phi_j \rangle^2 \\
= \| z_i^m \|^2 \| y_i^m - y_i \|^2 \\
= O_p \left( f(m)^2 \right).
\]

Hence,

\[
\left\| \left( \alpha I + \hat{V}_Z^m \right)^{-1} \left( \hat{C}_{ZY}^m - \hat{C}_{ZY} \right) \right\|_{HS}^2 = O_p \left( \frac{f(m)^2}{\alpha^2 n^2} \right).
\]
\[
\left[ \left( \alpha I + \hat{V}^m_Z \right)^{-1} - \left( \alpha I + \hat{V}_Z \right)^{-1} \right] \hat{C}_{ZY} \\
= \left( \alpha I + \hat{V}^m_Z \right)^{-1} \left( \hat{V}_Z - \hat{V}_Z^m \right) \left( \alpha I + \hat{V}_Z \right)^{-1} \hat{C}_{ZY} \\
= \left( \alpha I + \hat{V}^m_Z \right)^{-1} \left( \hat{V}_Z - \hat{V}_Z^m \right) \hat{\Pi}_a.
\]

Hence,
\[
\left\| \left( \alpha I + \hat{V}^m_Z \right)^{-1} \left( \hat{V}_Z - \hat{V}_Z^m \right) \hat{\Pi}_a^* \right\|_{HS}^2 = O_p \left( \frac{f(m)^2}{\alpha^2 n^2} \right).
\]

This concludes the proof of Proposition 5.

**Proof of Proposition 6.** Using the fact that \( \|a + b + c\|_{HS}^2 \leq 3 \left( \|a\|_{HS}^2 + \|b\|_{HS}^2 + \|c\|_{HS}^2 \right) \), we can evaluate the terms (25), (26), and (27) separately. The proof follows closely that of Proposition 2. Let \( Z \) and \( W \) be the sets \((Z_1, Z_2, \ldots, Z_n)\) and \((W_1, W_2, \ldots, W_n)\).

\[
\begin{align*}
E \left( (25) \right|_{HS}^2 |Z, W) &= tr \left\{ \left( \alpha I + \hat{C}_{ZW} \hat{C}_{WZ} \right)^{-1} \hat{C}_{ZW} E \left( \hat{C}_{WU} \hat{C}_{UW} |Z, W) \hat{C}_{WZ} \left( \alpha I + \hat{C}_{ZW} \hat{C}_{WZ} \right)^{-1} \right\}. \\
&= tr \left\{ \left( \alpha I + \hat{C}_{ZW} \hat{C}_{WZ} \right)^{-1} \hat{C}_{ZW} E \left( \hat{C}_{WU} \hat{C}_{UW} |Z, W) \hat{C}_{WZ} \left( \alpha I + \hat{C}_{ZW} \hat{C}_{WZ} \right)^{-1} \right\}. \\
&= \frac{1}{n} tr \left( \hat{V}_U \right) \hat{V}_W,
\end{align*}
\]

Using
\[
E \left( \hat{C}_{WU} \hat{C}_{UW} |Z, W) = \frac{1}{n} tr \left( \hat{V}_U \right) \hat{V}_W,
\]
we obtain
\[
E \left( (25) \right|_{HS}^2 \leq \frac{C}{n \alpha}
\]
for some constant \( C \).

The proof regarding the rates of convergence of (26) and (27) is similar to that of Proposition 2 and is not repeated here.
References


Carrasco, M. and J. P. Florens, 2000, Generalization of GMM to a continuum of moment conditions. Econometric Theory, 16, 797-834.


Centorrino, S., 2014, Data-driven selection of the regularization parameter in additive nonparametric instrumental regressions, mimeo, Stony Brook University.


Centorrino, S., F. Fève, and J. P. Florens, 2013, Implementation simulations and bootstrap in nonparametric instrumental variable estimation, mimeo, Stony Brook University.


Horowitz, J. and S. Lee, 2007, Nonparametric Instrumental Variables Estimation of a
Quantile Regression Model, Econometrica, Vol. 75, No. 4, 1191-1208.