Functional Sliced Inverse Regression in a Reproducing Kernel Hilbert Space: a Theoretical Connection to Functional Linear Regression

Jinan University and City University of Hong Kong

Supplementary Material

Proofs

As we mentioned in the text, the proof for the main result is based on a connection to the FLR problem. Since we assumed the more general conditions for the eigenvalues of T than Cai and Yuan (2012) (in condition (A3)), we provide a proof of the convergence rates for the estimator of the FLR problem for completeness. Our proof also differs from that used in Cai and Yuan (2012) and is slightly simpler. To ease notation, we do not emphasize the uniformity of the upper bound over $\{\beta \in \mathcal{H}_K : \|\beta\|_{\mathcal{H}_K} = 1\}$ below, but it can be easily checked step by step that all the bounds we obtain below are uniform over this set.

Proposition 1. *For a FLR problem* $Y = \langle \beta, X\rangle + \epsilon$ *with* $EX\epsilon = 0$*. Suppose the model satisfies conditions (A1)-(A3) (except* Y *does not have to be dis-*

crete here), and $\beta \in \mathcal{H}_K$. Given i.i.d. data (X_i, Y_i), the RKHS-based estimator $\hat{f}$ of $f = K^{-1/2}\beta$ is as explained in Section 2. With $\lambda \to 0, n\lambda \to \infty$, we have

$$\|T^{1/2}(\hat{f} - f)\|^2 = O_p(\lambda + \frac{1}{n}\sum_j \frac{s_j^2}{(s_j + \lambda)^2}).$$

Proof of Proposition 1. In the proofs we use C to denote a generic positive constant. With $\hat{\beta} = K^{1/2}\hat{f}$ and $\beta = K^{1/2}f$, using $\hat{f} = (T_n + \lambda I)^{-1}(\sum_i K^{1/2}X_iY_i/n)$ and noting that $T_n = \sum_i(K^{1/2}X_i \otimes K^{1/2}X_i)/n$, we have

$$\begin{aligned}
& T^{1/2}(\hat{f} - f) \\
= \; & T^{1/2}(T_n + \lambda I)^{-1}\frac{\sum_i K^{1/2}X_i(\langle f, K^{1/2}X_i\rangle + \epsilon_i)}{n} - T^{1/2}f \\
= \; & T^{1/2}(T_n + \lambda I)^{-1}T_n f - T^{1/2}f + T^{1/2}(T_n + \lambda I)^{-1}\frac{\sum_i K^{1/2}X_i\epsilon_i}{n} \\
= \; & T^{1/2}(T_n(T_n + \lambda I)^{-1} - I)f + T^{1/2}(T_n + \lambda I)^{-1}\frac{\sum_i K^{1/2}X_i\epsilon_i}{n}.
\end{aligned}$$

Using the identity that for two operators A and B, $A^{-1} - B^{-1} = B^{-1}(B - A)A^{-1}$, we have

$$\begin{aligned}
& T_n(T_n + \lambda I) \\
= \; & -\lambda(T_n + \lambda I)^{-1} \\
= \; & -\lambda(T + \lambda I)^{-1} - \lambda(T + \lambda I)^{-1}(T - T_n)(T_n + \lambda I)^{-1},
\end{aligned}$$

and thus we have

$$
\begin{aligned}
& T^{1/2}(\hat{f}-f) \\
= \ & -\lambda T^{1/2}(T+\lambda I)^{-1}f-\lambda T^{1/2}(T+\lambda I)^{-1}(T-T_n)(T_n+\lambda I)^{-1}f \\
& \quad +T^{1/2}(T_n+\lambda I)^{-1}\frac{\sum_i K^{1/2}X_i\epsilon_i}{n} \\
=: \ & A_1+A_2+A_3.
\end{aligned}
$$

For A_1, using that $\|T^{1/2}(T+\lambda I)^{-1/2}\|_{op}\leq 1$, $\|\sqrt{\lambda}(T+\lambda I)^{-1/2}\|_{op}\leq 1$, we have $\|A_1\|^2=O_p(\lambda)$.

For A_2, we have

$$
\begin{aligned}
& \|A_2\|^2 \\
\leq \ & \|T^{1/2}(T+\lambda I)^{-1}(T-T_n)\|_{op}^2\|\lambda(T_n+\lambda I)^{-1}\|_{op}^2 \\
\leq \ & \|T^{1/2}(T+\lambda I)^{-1}(T-T_n)\|_{HS}^2,
\end{aligned}
$$

where $\|.\|_{HS}$ is the Hilbert-Schmidt norm, and we used the property that the operator norm is upper bounded by the Hilbert-Schmidt norm. We have

$$
\begin{aligned}
& E\|(T-T_n)(T+\lambda I)^{-1}T^{1/2}\|_{HS}^2 \\
= \ & E\sum_{j,k}\langle (T-T_n)(T+\lambda I)^{-1}T^{1/2}\psi_j,\psi_k\rangle^2 \\
= \ & E\sum_{j,k}\langle (T-T_n)\frac{s_j^{1/2}}{(s_j+\lambda)}\psi_j,\psi_k\rangle^2. \qquad \text{(S0.1)}
\end{aligned}
$$

Direct calculation reveals that

$$\begin{aligned}
& E\langle (T-T_n)\psi_j, \psi_k\rangle^2 \\
= \; & E\langle s_j\psi_j - \frac{1}{n}\sum_i ((\sum_l \xi_{il}\psi_l)\otimes(\sum_m \xi_{im}\psi_m))\psi_j, \psi_k\rangle^2 \\
= \; & E\langle s_j\psi_j - \sum_i \frac{\sum_l \xi_{il}\xi_{ij}\psi_l}{n}, \psi_k\rangle^2 \\
= \; & E(s_j\delta_{jk} - \frac{\sum_i \xi_{ij}\xi_{ik}}{n})^2.
\end{aligned}$$

Noting that $E[\xi_{ij}\xi_{ik}] = s_j\delta_{jk}$, the above is equal to $Var(\sum_i \xi_{ij}\xi_{ik}/n) \leq E(\sum_i \xi_{ij}\xi_{ik}/n)^2$. Using assumption (A2), we have $E\langle (T-T_n)\psi_j, \psi_k\rangle^2 \leq Cs_js_k/n$, which combined with (S0.1) implies

$$E\|(T-T_n)(T+\lambda I)^{-1}T^{1/2}\|_{HS}^2 \leq \frac{C}{n}\sum_{j,k}\frac{s_j^2 s_k}{(s_j+\lambda)^2} = O\left(\frac{1}{n}\sum_j \frac{s_j^2}{(s_j+\lambda)^2}\right). \tag{S0.2}$$

Thus $\|A_2\|^2 = O_p(\frac{1}{n}\sum_j \frac{s_j^2}{(s_j+\lambda)^2})$.

For A_3, we write $A_3 = T^{1/2}(T+\lambda I)^{-1}\frac{\sum_i K^{1/2}X_i\epsilon_i}{n} + T^{1/2}(T+\lambda I)^{-1}(T-T_n)(T_n+\lambda I)^{-1}\frac{\sum_i K^{1/2}X_i\epsilon_i}{n}$. Writing $K^{1/2}X_i = \sum_j \xi_{ij}\psi_j$, we have

$$\begin{aligned}
& E\|T^{1/2}(T+\lambda I)^{-1}\frac{\sum_i K^{1/2}X_i\epsilon_i}{n}\|^2 \\
= \; & \frac{\sigma_\epsilon^2}{n}E\|T^{1/2}(T+\lambda I)^{-1}K^{1/2}X_1\|^2 \\
= \; & \frac{\sigma_\epsilon^2}{n}E\|T^{1/2}(T+\lambda I)^{-1}\sum_j \xi_{1j}\psi_j\|^2,
\end{aligned}$$

where $\sigma_\epsilon^2 = E[\epsilon^2]$. Since $(T+\lambda I)^{-1}\psi_j = (s_j+\lambda)^{-1}\psi_j$ and $T^{1/2}\psi_j = \sqrt{s_j}\psi_j$, we have $E\|T^{1/2}(T+\lambda I)^{-1}\sum_j \xi_{1j}\psi_j\|^2 = E\|\sum_j \frac{\sqrt{s_j}\xi_{1j}}{s_j+\lambda}\psi_j\|^2 = \sum_j \frac{s_j^2}{(s_j+\lambda)^2}$.

Furthermore, using (S0.2), letting $\mathcal{A} := T^{1/2}(T+\lambda I)^{-1}(T-T_n)$ and $\mathcal{A}^T$ the adjoint operator of $\mathcal{A}$ for simplicity of notation, we have

$$
\begin{aligned}
& E[\|T^{1/2}(T+\lambda I)^{-1}(T-T_n)(T_n+\lambda I)^{-1}\frac{\sum_i K^{1/2}X_i\epsilon_i}{n}\|^2|X_1,\ldots,X_n] \\
= & \frac{\sigma_\epsilon^2}{n^2}\sum_i \|\mathcal{A}(T_n+\lambda I)^{-1}K^{1/2}X_i\|^2 \\
= & \frac{\sigma_\epsilon^2}{n^2}\sum_i \langle \mathcal{A}(T_n+\lambda I)^{-1}K^{1/2}X_i, \mathcal{A}(T_n+\lambda I)^{-1}K^{1/2}X_i\rangle \\
= & \frac{\sigma_\epsilon^2}{n^2}\sum_i \langle (T_n+\lambda I)^{-1}\mathcal{A}^T\mathcal{A}(T_n+\lambda I)^{-1}K^{1/2}X_i, K^{1/2}X_i\rangle.
\end{aligned}
$$

Next we introduce the definition of the trace, which is given by $\mathrm{tr}(\mathcal{F}) := \sum_j \langle \mathcal{F}e_j, e_j\rangle$ for any orthonormal basis $\{e_j\}$. Using the properties $\langle f, g\rangle = \mathrm{tr}(f\otimes g)$ for any $f, g \in L_2[0,1]$, $\mathrm{tr}(\mathcal{F}\mathcal{G}) = \mathrm{tr}(\mathcal{G}\mathcal{F})$ and $\mathrm{tr}(\mathcal{F}^T\mathcal{F}) = \|\mathcal{F}\|_{HS}^2$ (see section 18 of Conway (2000) for these basic properties of trace and Hilbert-Schmidt norm), we get

$$
\begin{aligned}
& \frac{1}{n^2}\sum_i \langle (T_n+\lambda I)^{-1}\mathcal{A}^T\mathcal{A}(T_n+\lambda I)^{-1}K^{1/2}X_i, K^{1/2}X_i\rangle \\
= & \frac{1}{n^2}\sum_i \mathrm{tr}\left((T_n+\lambda I)^{-1}\mathcal{A}^T\mathcal{A}(T_n+\lambda I)^{-1}K^{1/2}X_i\otimes K^{1/2}X_i\right) \\
= & \frac{1}{n}\mathrm{tr}((T_n+\lambda I)^{-1}\mathcal{A}^T\mathcal{A}(T_n+\lambda I)^{-1}T_n) \\
= & \frac{1}{n}\mathrm{tr}(T_n^{1/2}(T_n+\lambda I)^{-1}\mathcal{A}^T\mathcal{A}(T_n+\lambda I)^{-1}T_n^{1/2}) \\
= & \frac{1}{n}\|\mathcal{A}(T_n+\lambda I)^{-1}T_n^{1/2}\|_{HS}^2.
\end{aligned}
$$

Using $\|\mathcal{F}\mathcal{G}\|_{HS} \le \|\mathcal{F}\|_{HS}\|\mathcal{G}\|_{op}$ and the bound for $\|\mathcal{A}\|_{HS}$ which was already obtained in (S0.1), the above is bounded by

$$\begin{aligned} & O_p\left(\frac{1}{n}\|(T_n+\lambda I)^{-1}T_n^{1/2}\|_{op}^2\|\mathcal{A}\|_{HS}^2\right) \\ = \ & O_p(\frac{1}{n\lambda})O_p(\frac{1}{n}\sum_j \frac{s_j^2}{(s_j+\lambda)^2}) \\ = \ & o_p(\frac{1}{n}\sum_j \frac{s_j^2}{(s_j+\lambda)^2}), \end{aligned}$$

since $n\lambda \to \infty$.

Thus we have $\|A_3\|^2 = O_p(\frac{1}{n}\sum_j \frac{s_j^2}{(s_j+\lambda)^2})$. The theorem is proved by combining the bounds for $\|A_1\|^2$, $\|A_2\|^2$ and $\|A_3\|^2$. □

The above proposition demonstrated the convergence rate based on the prediction risk $\|T^{1/2}(\hat{f}-f)\|$. Since $T^{1/2}$ has eigenvalues converging to zero, this does not even imply the consistency of $\hat{f}$ itself. The following proposition shows $\|\hat{f}-f\| = O_p(1)$, which suffices for our purpose later (in particular, this is used in Step 5 in the proof of Theorem 1 below).

Proposition 2. *Under the same setup for FLR as in Proposition 1, and choose λ to be the solution of (3.2), we have $\|\hat{f}-f\| = O_p(1)$. In particular, $\|\hat{f}\| = O_p(1)$.*

Proof of Proposition 2. The proof follows similar lines of Proposition 1.

We now have

$$\begin{aligned} & \hat{f} - f \\ = \ & -\lambda(T+\lambda I)^{-1}f - \lambda(T+\lambda I)^{-1}(T-T_n)(T_n+\lambda I)^{-1}f \\ & +(T_n+\lambda I)^{-1}\frac{\sum_i K^{1/2}X_i\epsilon_i}{n} \\ =: \ & B_1 + B_2 + B_3. \end{aligned}$$

Since $\|\lambda(T+\lambda)^{-1}\|_{op} \leq 1$, obviously $\|B_1\| = O_p(1)$.

For B_2, we have $\|B_2\|^2 \leq \|(T-T_n)(T+\lambda I)^{-1}\|_{HS}^2$. Unlike (S0.2) here $T^{1/2}$ does not appear. However, we can follow the same lines to get

$$E\|(T-T_n)(T+\lambda I)^{-1}\|_{HS}^2 = O\left(\frac{1}{n}\sum_j \frac{s_j}{(s_j+\lambda)^2}\right).$$

For B_3 we similarly have $\|B_3\|^2 = O_p\left(\frac{1}{n}\sum_j \frac{s_j}{(s_j+\lambda)^2}\right)$.

Thus we only need to show

$$\frac{1}{n}\sum_j \frac{s_j}{(s_j+\lambda)^2} = O_p(1). \tag{S0.3}$$

Demonstration of (S0.3) is similar to the discussions following the statement of Theorem 1. Let $J = \lfloor \phi^{-1}(\lambda) \rfloor$. By splitting the sum over j into $j \leq J$ and $j > J$, we have

$$\frac{1}{n}\sum_j \frac{s_j}{(s_j+\lambda)^2} \leq \frac{J}{ns_J} + \frac{\sum_{j\geq J+1} s_j}{n\lambda^2}.$$

Since λ satisfies $\phi^{-1}(\lambda) = n\lambda$, and $J \leq \phi^{-1}(\lambda)$, $J+1 > \phi^{-1}(\lambda)$, we have

$$\frac{J}{ns_J} = \frac{J}{n\phi(J)} \leq \frac{\phi^{-1}(\lambda)}{n\phi(\phi^{-1}(\lambda))} = 1,$$

and

$$\frac{\sum_{j\geq J+1} s_j}{n\lambda^2} \leq \frac{(J+2)s_{J+1}}{n\lambda^2} \leq \frac{J}{n\lambda} \leq 1,$$

where we used that $\sum_{j\geq J+1} s_j \leq (J+2)s_{J+1}$ obtained from Lemma 1 of Cardot et al. (2007). This established (S0.3) and thus $\|\hat{f} - f\| = O_p(1)$. $\square$

Proof of Theorem 1. For clarity, the proof is splitted into several steps. For simplicity of notation, we only consider the convergence rate of the first eigenfunction associated with eigenvalue α_1. That is we focus on $\|\Gamma^{1/2}(\hat{\beta}_1 - \beta_1)\| = \|T^{1/2}(\hat{f}_1 - f_1)\|$, and we omit the subscript 1 in the following. Convergence rates for other eigenfunctions can be proved in exactly the same way.

STEP 1. There exists H numbers $a_1, \ldots, a_H$ such that $f = \sum_{h=1}^H a_h T^{-1} K^{1/2} X_h$ and $a = (a_1, \ldots, a_H)^T$ satisfies the eigenvalue equation $\mathcal{P}\mathcal{C}a = \alpha a$, where $X_h = E[X|Y = y_h]$, $\mathcal{P} = \text{diag}(p_1, \ldots, p_H)$ and $\mathcal{C}$ is an $H \times H$ matrix with entries given by $\mathcal{C}_{h,h'} = \langle T^{-1}K^{1/2}X_h, K^{1/2}X_{h'}\rangle$.

Proof. Here we use the fact that $Var(E[K^{1/2}X|Y]) = \sum_h p_h K^{1/2}X_h \otimes K^{1/2}X_h$ is of finite-rank. $T^{-1}Var(E[K^{1/2}X|Y])f = \alpha f$ can be equivalently written as

$$\sum_h p_h \langle K^{1/2}X_h, f\rangle T^{-1}K^{1/2}X_h = \alpha f. \tag{S0.4}$$

Since we assumed $\alpha > 0$, f is a linear combination of $T^{-1}K^{1/2}X_h, h = 1, \ldots, H$. We write $f = \sum_h \tilde{a}_h T^{-1}K^{1/2}X_h$. Note that in general $K^{1/2}X_h, h = 1, \ldots, H$ are not linearly independent (for example we know $\sum_h p_h X_h = EX = 0$). We will pick a particular $\tilde{a}_h$ soon.

Plugging this expression of f into (S0.4), we get

$$\sum_{h=1}^{H}(\sum_{h'=1}^{H} p_h \tilde{a}_{h'} \mathcal{C}_{hh'})T^{-1}K^{1/2}X_h = \sum_h \alpha \tilde{a}_h T^{-1}K^{1/2}X_h, \tag{S0.5}$$

where $\mathcal{C}_{hh'} = \langle T^{-1}K^{1/2}X_h, K^{1/2}X_{h'}\rangle$. Using the notations introduced above, this is same as

$$(\mathcal{PC}\tilde{a} - \alpha\tilde{a})^T(T^{-1}K^{1/2}X_{1:H}) = 0, \tag{S0.6}$$

where $\tilde{a} = (\tilde{a}_1, \ldots, \tilde{a}_H)^T$ and $T^{-1}K^{1/2}X_{1:H} = (\Gamma^{-1}K^{1/2}X_1, \ldots, \Gamma^{-1}K^{1/2}X_H)^T$ is a vector of elements in $L_2[0,1]$. Let $b = \mathcal{PC}\tilde{a} - \alpha\tilde{a}$ and (S0.6) says $b^T T^{-1}K^{1/2}X_{1:H} = 0$, and thus $\mathcal{C}b = 0$ by the definition of $\mathcal{C}$. Using $\mathcal{C}b = 0$ and $b = \mathcal{PC}\tilde{a} - \alpha\tilde{a}$, we have $\mathcal{PC}(\tilde{a} + b/\alpha) = \alpha(\tilde{a} + b/\alpha)$, and thus $a := \tilde{a} + b/\alpha$ satisfies the statement at the beginning of the step.

STEP 2. There exists H numbers $\hat{a}_1, \ldots, \hat{a}_H$ such that $\hat{f} = \sum_{h=1}^{H} \hat{a}_h (T_n + \lambda I)^{-1}K^{1/2}\hat{X}_h$ and $\hat{a} = (\hat{a}_1, \ldots, \hat{a}_H)^T$ satisfies the eigenvalue equation $\hat{\mathcal{P}}\hat{\mathcal{C}}\hat{a} = \hat{\alpha}\hat{a}$, where $\hat{\mathcal{P}} = \mathrm{diag}(\hat{p}_1, \ldots, \hat{p}_H)$ and $\hat{\mathcal{C}}$ is an $H \times H$ matrix with entries given by $\hat{\mathcal{C}}_{h,h'} = \langle (T_n + \lambda I)^{-1}K^{1/2}\hat{X}_h, K^{1/2}\hat{X}_{h'}\rangle$.

Proof. Given that $\widehat{Var}(E[K^{1/2}X|Y]) = \sum_h \hat{p}_h K^{1/2}\hat{X}_h \otimes K^{1/2}\hat{X}_h$ and $\hat{f}$ satisfies $(T_n + \lambda I)^{-1}\widehat{Var}(E[K^{1/2}X|Y])\hat{f} = \hat{\alpha}\hat{f}$, the proof is the same as for Step 1.

STEP 3. Let $Z^{(h)} = I\{Y = y_h\} - p_h$ where $I\{.\}$ is the indicator function. Then $Z^{(h)}$ can be expressed as $Z^{(h)} = \langle \beta^{(h)}, X\rangle + \epsilon$ for some ϵ with $E[X\epsilon] = 0$, and some $\beta^{(h)} \in \mathcal{H}_K$.

Proof. We use the arguments put forward in Section 2 of Cardot et al. (2003) (see page 575 in that paper). More specifically, consider the minimization problem $f^{(h)} = \arg\min_f E[(Z^{(h)} - \langle f, K^{1/2}X\rangle)^2]$. For this minimization problem, the covariance operator is T with eigenvalues and eigenfunctions $\{s_j, \psi_j\}$. We have $E[Z^{(h)}K^{1/2}X] = E[I\{Y = y_h\}K^{1/2}X] = p_h E[K^{1/2}X|Y = y_h] \in T\mathcal{S}^*_{Y|X}$. Thus $E[Z^{(h)}K^{1/2}X]$ is a linear combination of $f_1, \ldots, f_M$ which spans $\mathcal{S}^*_{Y|X}$. Since

$$\sum_j \frac{\langle Tf_m, \psi_j\rangle^2}{s_j^2} = \sum_j \frac{\langle f_m, T\psi_j\rangle^2}{s_j^2} = \sum_j \frac{\langle f_m, s_j\psi_j\rangle^2}{s_j^2} = \sum_j \langle f_m, \psi_j\rangle^2 = \|f_m\|^2 < \infty,$$

we also have

$$\sum_j \frac{\langle E[Z^{(h)}K^{1/2}X], \psi_j\rangle^2}{s_j^2} < \infty,$$

which verifies condition 1 of Cardot et al. (2003) and thus $f^{(h)}$ exists and is unique. We can thus write $Z^{(h)} = \langle f^{(h)}, K^{1/2}X\rangle + \epsilon$ with $E[X\epsilon] = 0$. Now we only need to let $\beta^{(h)} = K^{1/2}X$.

STEP 4. For $1 \le h \le H$, $\|\hat{p}_h T^{1/2}(T_n+\lambda I)^{-1}K^{1/2}\hat{X}_h - p_h T^{-1/2}K^{1/2}X_h\|^2 = O_p(\lambda + \frac{1}{n}\sum_j \frac{s_j^2}{(s_j+\lambda)^2})$.

Proof. Note that $p_h K^{1/2}X_h = EK^{1/2}XZ^{(h)}$ and $\hat{p}_h K^{1/2}\hat{X}_h = \sum_i K^{1/2}X_i Z_i^{(h)}/n$ (note we assumed $\sum_i X_i/n = 0$ to make arguments simpler), where $Z_i^{(h)} = I\{Y_i = y_h\} - p_h$. By the representation of $Z^{(h)}$ presented in Step 3, $\|T^{1/2}(T_n+\lambda I)^{-1}\sum_i K^{1/2}X_i Z_i^{(h)}/n - p_h T^{-1/2}K^{1/2}X_h\|^2$ is exactly the risk for the FLR problem $Z^{(h)} = \langle \beta^{(h)}, X\rangle + \epsilon$ and thus is order $O_p(\lambda + \frac{1}{n}\sum_j \frac{s_j^2}{(s_j+\lambda)^2})$.

STEP 5. For $1 \le h, h' \le H$, $((\hat{\mathcal{P}}\hat{\mathcal{C}})_{h,h'} - (\mathcal{P}\mathcal{C})_{h,h'})^2 = O_p(\lambda + \frac{1}{n}\sum_j \frac{s_j^2}{(s_j+\lambda)^2})$, *where* $(\mathcal{P}\mathcal{C})_{h,h'}$ *is the* (h,h')*-entry of* $\mathcal{P}\mathcal{C}$*, for example.*

Proof. We have

$$\begin{aligned}
& [(\hat{\mathcal{P}}\hat{\mathcal{C}})_{h,h'} - (\mathcal{P}\mathcal{C})_{h,h'}]^2 \\
= \; & [\langle \hat{p}_h(T_n+\lambda I)^{-1}K^{1/2}\hat{X}_h, K^{1/2}\hat{X}_{h'}\rangle - \langle p_h T^{-1}K^{1/2}X_h, K^{1/2}X_{h'}\rangle]^2 \\
\le \; & 2\langle \hat{p}_h(T_n+\lambda I)^{-1}K^{1/2}\hat{X}_h - p_h T^{-1}K^{1/2}X_h, K^{1/2}X_{h'}\rangle^2 \\
& +2\langle \hat{p}_h(T_n+\lambda I)^{-1}K^{1/2}\hat{X}_h, K^{1/2}\hat{X}_{h'} - K^{1/2}X_{h'}\rangle^2. \qquad \text{(S0.7)}
\end{aligned}$$

Noting $K^{1/2}X_h' = T\beta$ for some $\beta \in \mathcal{S}_{Y|X}$, we have $\langle \hat{p}_h(T_n+\lambda I)^{-1}K^{1/2}\hat{X}_h - p_hT^{-1}K^{1/2}X_h, K^{1/2}X_{h'}\rangle^2 = \langle \hat{p}_hT^{1/2}(T_n+\lambda I)^{-1}K^{1/2}\hat{X}_h - p_hT^{-1/2}K^{1/2}X_h, T^{1/2}\beta\rangle^2$. Using Step 4, the first term in (S0.7) is thus $O_p(\lambda + \frac{1}{n}\sum_j \frac{s_j^2}{(s_j+\lambda)^2})$. Furthermore, for the second term in (S0.7), we have

$$\hat{p}_h(T_n+\lambda I)^{-1}K^{1/2}\hat{X}_h = (T_n+\lambda I)^{-1}\sum_i K^{1/2}X_iZ_i^{(h)}/n.$$

By Proposition 2, noting the right hand side above is the estimator of $f^{(h)}$ for the FLR problem $Z^{(h)} = \langle f^{(h)}, K^{1/2}X\rangle + \epsilon$, we see it is of order $O_p(1)$.

Finally, it is easy to show that $\|K^{1/2}\hat{X}_{h'} - K^{1/2}X_{h'}\|^2 = O_p(n^{-1})$. Thus the second term of (S0.7) is also $O_p(\lambda + \frac{1}{n}\sum_j \frac{s_j^2}{(s_j+\lambda)^2})$.

STEP 6. Finally we can combine the claims above to prove the theorem. Based on Steps 1,2,5 and the perturbation theory for matrices, one can get $\|\alpha - \widehat{\alpha}\|^2 = O_p(\lambda + \frac{1}{n}\sum_j \frac{s_j^2}{(s_j+\lambda)^2})$ *and* $\|\hat{a} - a\| = O_p(\lambda + \frac{1}{n}\sum_j \frac{s_j^2}{(s_j+\lambda)^2})$ *(when the sign of the eigenvector is appropriately chosen). Using* $T^{1/2}\hat{f} = \sum_{h=1}^H \hat{a}_hT^{1/2}(T_n+\lambda I)^{-1}K^{1/2}\hat{X}_h$ *and* $T^{1/2}f = \sum_{h=1}^H a_hT^{-1/2}K^{1/2}X_h$*, combined with Step 4, we have* $\|T^{1/2}(\hat{f} - f)\|^2 = O_p(\lambda + \frac{1}{n}\sum_j \frac{s_j^2}{(s_j+\lambda)^2})$. □

Proof of Theorem 2. Consider the model

$$Y_i = \int \beta(s)X_i(s)ds + \epsilon_i,$$

with $\|\beta\|_{\mathcal{H}_K} \leq 1$. We need a modification of the proof of Theorem 1 in Cai and Yuan (2012) due to the more general assumption on the eigenvalues of T. Let $\eta_j = \sqrt{c\lambda/(Js_j)}$ for some $0 < c \leq 1$ to be determined later. We apply Theorem 2.5 of Tsybakov (2009) using the following collection of 2^J functions

$$f_\theta = \sum_{k=1}^{J} \theta_k \eta_k K^{1/2}\psi_k,$$

where $\theta = (\theta_1, \ldots, \theta_J) \in \{0, 1\}^J$.

First, using that $\|K^{1/2}\psi_j, K^{1/2}\psi_k\|_{\mathcal{H}_K} = \langle \psi_j, \psi_k\rangle = \delta_{jk}$,

$$\|f_\theta\|^2_{\mathcal{H}_K} = \sum_{k=1}^{J} \theta_k^2 \eta_k^2 \leq \sum_{k=1}^{J} \eta_k^2 = \frac{c\lambda}{J}\sum_{k=1}^{J}\frac{1}{s_k} \leq \frac{c\lambda}{J}\frac{J}{s_J} \leq c \leq 1,$$

since $s_J \geq \lambda$ by $s_J = \phi(J)$ and the definition $J = \lfloor \phi^{-1}(\lambda)\rfloor$.

By the Varshamov-Gilbert bound (Lemma 2.9 in Tsybakov (2009)), there is a subset $\Theta = \{\theta^0, \ldots, \theta^N\} \subset \{0, 1\}^J$ such that $\theta^0 = (0, \ldots, 0)$, $N \geq 2^{J/8}$ and $\sum_{k=1}^{J}(\theta_k - \theta'_k)^2 \geq J/8$ whenever $\theta \neq \theta' \in \Theta$.

We have

$$\|\Gamma^{1/2}(f_\theta - f_{\theta'})\|^2 = \sum_{k=1}^{J} \eta_k^2(\theta_k - \theta'_k)^2 s_k \geq \frac{c\lambda}{J}\frac{J}{8} = c\lambda/8,$$

verifying condition (i) in Theorem 2.5 of Tsybakov (2009). Furthermore, the Kullback-Leibler distance between P_θ and $P_{\theta'}$ (P_θ is the joint distribution of training data when $\beta = f_\theta$) can be found to be

$$K(P_\theta|P_{\theta'}) = \frac{n}{2\sigma^2}\sum_{k=1}^{J}\eta_k^2(\theta_k - \theta_k')^2 s_k \le \frac{nc\lambda}{2\sigma^2},$$

and thus

$$\frac{1}{N}\sum_{j=1}^{N}K(P_\theta|P_{\theta'}) \le \frac{nc\lambda}{2\sigma^2} = \frac{c\phi^{-1}(\lambda)}{2\sigma^2} \le \frac{c}{2\sigma^2}(J+1) \le \alpha\log N,$$

for some $0 < \alpha < 1/8$ if c is chosen small enough, verifying condition (ii) in Theorem 2.5 of Tsybakov (2009). The lower bound is proved by applying Theorem 2.5 of Tsybakov (2009). $\square$

Bibliography

Cai, T. and Yuan, M. (2012). Minimax and adaptive prediction for functional linear regression, *Journal of the American Statistical Association* **107**(499): 1201–1216.

Cardot, H., Ferraty, F. and Sarda, P. (2003). Spline estimators for the functional linear model, *Statistica Sinica* **13**(3): 571–591.

Cardot, H., Mas, A. and Sarda, P. (2007). Clt in functional linear regression models, *Probability Theory and Related Fields* **138**(3): 325–361.

Conway, J. (2000). *A course in operator theory*, American Mathematical Society.

Tsybakov, A. (2009). *Introduction to Nonparametric Estimation*, Springer, New York.