Golubev, Georgi; Härdle, Wolfgang

Working Paper

On adaptive estimation in partial linear models

SFB 373 Discussion Paper, No. 2000,21

Provided in Cooperation with:
Collaborative Research Center 373: Quantification and Simulation of Economic Processes, Humboldt University Berlin

On adaptive estimation in partial linear models

G. Golubev, W. Härdle*

Institute Information Transmission Problems
and
Université d'Aix-Marseille

Humboldt-Universität zu Berlin

Abstract

We consider a problem of estimation of parametric component in a partial linear model. Suppose that a finite set $\mathcal{E}$ of linear estimators is given. Our goal is to mimic the estimator in $\mathcal{E}$ that has the smallest risk. Using a second order expansion of the risk of linear estimators we propose a practically feasible adaptive procedure for choice of smoothing parameters based on the principle of unbiased risk estimation.

1 Introduction

Partial linear models represent now a flexible and growing class of models for statistical applications. In the present paper we will deal with the simplest partial linear model

$$Y_i = \theta^T Z_i + m(X_i) + \xi_i, \quad i = 1, \ldots, n, \tag{1}$$

where $\theta \in \mathbf{R}^d$ is a vector of unknown parameters, ξ_i are i.i.d. random variables with zero mean and the finite variance $\sigma^2 = \mathbf{E}\xi_i^2$. The regressors $X_i \in [0,1]$ are assumed to be i.i.d. random variables with a strictly positive density $q(x)$ on $[0,1]$. We will also assume that they do not depend on ξ_i. The nuisance function $m(x)$, $x \in [0,1]$ is unknown but such that the random variables $m(X_i)$ have zero mean.

*The authors were supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 373 "Quantifikation und Simulation Ökonomischer Prozesse", Humboldt-Universität zu Berlin.

It is well known that if $m(x),\ x \in [0,1]$ is sufficiently smooth then θ can be estimated with the ordinary parametric rate Heckman (1986), Speckman (1988), Bhattacharia and Zhao (1997), Mammen and Van de Geer (1997), Chen (1998). Further references and applications of partially linear models can be found in the recent book by Härdle, Liang and Gao (1999). Roughly speaking almost all first order effects in these models can be explained at a heuristic level if one assumes that the nuisance function $m(x)$ is known. The next step in investigation of this problem was made in Golubev and Härdle (2001), where a second order term of quadratic risk of linear estimators was found (see Theorems 1, 2 below).

This paper continues Golubev, Härdle (2001) concentrating on data driven choice of smoothing parameters for penalized least-square estimators. Usually this problem is considered as a problem of minor importance in the modern theory of semi-parametric estimation. But from a practical point of view this problem plays the same role as a data driven choice of the bandwidth in density estimation problem. There is a vast mathematical literature on data driven choice of smoothing parameters (see Efroimovich and Pinsker (1984), Lepski (1991), Lepski (1992), Golubev and Nussbaum (1992), Kneip (1994), Lepski and Spokoiny (1997), Nemirovskii (1998), Tsybakov (1998), Barron, Birgé and Massart (1999), among others). The problem of adaptive choice of smoothing parameters in the framework of the second order minimax theory for distribution function estimation was considered in Golubev and Levit (1996).

To simplify some technical details we will assume that the nuisance function $m(x)$ belongs to the Sobolev ball

$$\mathbf{W}^2_\beta(L) = \Big\{ m : \int_0^1 [m^{(\beta)}(x)]^2 dx \le L, \quad \int_0^1 q(x) m(x)\, dx = 0 \Big\},$$

where the smoothness β is integer.

Our consideration is based on the so-called orthogonal series approach. The cornerstone idea of this approach is to parameterize the functional class $\mathbf{W}^2_\beta(L)$. We do this by constructing an orthonormal system in the Hilbert space $\mathbf{L}^2_q[0,1]$ which is equipped with the norm $\|\cdot\|_q$ and with the inner product $<\cdot,\cdot>_q$ defined by

$$\|f\|_q^2 = \int_0^1 q(x) f^2(x)\, dx, \quad <f,g>_q = \int_0^1 q(x) f(x) g(x)\, dx.$$

Let $\psi_k,\ k = 0,\dots,\beta-1$ be the fist orthonormal polynomials in $\mathbf{L}^2_q[0,1]$. The remainder functions $\psi_k,\ k \ge \beta$ are defined as solutions of the following boundary value problem

$$(-1)^\beta \frac{d^{2\beta}}{dx^{2\beta}} \psi_s(x) = \lambda_s q(x) \psi_s(x), \tag{2}$$
$$\frac{d^k}{dx^k}\psi_s(x)\Big|_{x=0} = \frac{d^k}{dx^k}\psi_s(x)\Big|_{x=1} = 0, \quad k = \beta,\dots,2\beta-1.$$

The eigen functions $\psi_s(x)$ are uniformly bounded in x. The asymptotic behavior of the eigen values λ_k plays a very important role in spline theory and it is well-known that

$$\lambda_s = [1+o(1)](\pi s)^{2\beta}\Big[\int_0^1 q^{1/(2\beta)}(x)\,dx\Big]^{2\beta}, \quad s \to \infty. \tag{3}$$

For details we refer to Utreras (1980), Speckman (1985), Duistermaat (1995). Thus any function $m \in \mathbf{W}^2_\beta(L)$ can be represented as the Fourier series

$$m(t) = \sum_{k=1}^{\infty} \nu_k \psi_k(t), \text{ with } \sum_{k=1}^{\infty} \nu_k^2 \lambda_k \le L, \tag{4}$$

where $\nu_k = <m, \psi_k>_q$.

Now we are ready to construct a penalized least-square estimator. Let Σ be an arbitrary diagonal matrix with the entries $\Sigma_{kk} = \sigma_k > 0$ and matrix Ψ be defined as $\Psi_{ki} = \psi_k(X_i)$. We estimate parameter θ by

$$\widehat{\theta} = \arg\min_{\theta \in R^d} \min_{\nu} \Big\{ \big\| Y - Z^T\theta - \Psi^T \nu \big\|^2 + \big\| \Sigma^{-1}\nu \big\|^2 \Big\}. \tag{5}$$

Our further considerations will be essentially based on the second order theory of semi-parametric estimation. The next two theorems describe the performance of $\widehat{\theta}$ up to the second order terms (see for more details Golubev and Härdle (2001)). Let

$$H = n\Sigma^{-2}(n\Sigma^{-2}+E)^{-1}, \tag{6}$$

where E is identity matrix and

$$IMSE[H,\nu] = \|(E-H)\nu\|^2 + \frac{\sigma^2}{n}\mathrm{tr}\, H^2$$

be the integrated mean-square error of recovering function $m(x)$ in the model (1) provided that θ is known. In other words

$$IMSE[H,\nu] = [1+o(1)]\sum_{k=0}^{\infty}(\nu_k - H_{kk}\bar{\nu}_k)^2,$$

where

$$\bar{\nu}_k = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \theta^T Z_i)\psi_k(X_i).$$

Theorem 1 *Suppose that* $\mathbf{E}\,|\xi_k|^{2+\delta} < \infty$ *for some* $\delta > 0$ *and*

$$\lim_{n\to\infty} \log^{1/2} n \max_{i,k} Z_{ki}^2 / \sum_{i=1}^n Z_{ki}^2 = 0, \quad \max_n \left\|(ZZ^T)^{-1}\right\| \max_k \sum_{i=1}^n Z_{ki}^2 < \infty, \tag{7}$$

$$\lim_{n\to\infty} \mathrm{tr}^2 H \log^{1/2} n/n = 0.$$

Then as $n \to \infty$ *uniformly in* $m \in \mathbf{W}_\beta^2(L)$

$$\mathbf{E}(\widehat{\theta} - \theta)(\widehat{\theta} - \theta)^T = (ZZ^T)^{-1}\{\sigma^2 + [1 + o(1)] IMSE[H, \nu]\}.$$

This theorem demonstrates how the risk of $\widehat{\theta}$ depends on the nuisance function $m(\cdot)$ and the penalization matrix Σ^{-1}. We see the first order term of the risk does not depend on the nonparametric nuisance function. Note the second order term in the risk expansion coincides with the integrated mean square error of recovering $m(x)$ in the model (5) provided θ is known. This is why we are interested in the analysis of the second order theory. One can maximize the second order term over all nuisance functions from $\mathbf{W}_\beta^2(L)$ and than minimize it over all penalizations Σ or equivalently over all H. Thus one obtains the following result about the minimax penalization.

Theorem 2 *Let* $\widehat{\theta}$ *be the estimator defined by (5) with* $\Sigma = H^{1/2}(E - H)^{-1/2}\sigma/\sqrt{n}$, *where* H *is the diagonal matrix*

$$H_{ss} = \left[1 - \omega\sqrt{\lambda_s}\right]_+, \tag{8}$$

where $[x]_+ = \max(x, 0)$, *and* ω *be a root of the equation*

$$\frac{\sigma^2}{n} \sum_{s=1}^{\infty} \lambda_s \left[\frac{1}{\omega\sqrt{\lambda_s}} - 1\right]_+ = L. \tag{9}$$

Under the conditions of Theorem 1 as $n \to \infty$

$$\sup_{m\in\mathbf{W}_\beta^2(L)} \mathbf{E}(\widehat{\theta} - \theta)(\widehat{\theta} - \theta)^T = (ZZ^T)^{-1}\sigma^2\{1 + [1 + o(1)]n^{-1}\mathrm{tr}H\}.$$

If ξ_i *are Gaussian then for any* $B > 0$

$$\inf_{\widetilde{\theta}} \sup_{m\in\mathbf{W}_\beta^2(L)} \sup_{\|\theta\|\le B} \mathbf{E}(\widetilde{\theta} - \theta)(\widetilde{\theta} - \theta)^T = (ZZ^T)^{-1}\sigma^2\{1 + [1 + o(1)]n^{-1}\mathrm{tr}\, H\},$$

where inf is taken over all estimators of the parameter θ.

Thus we see that the optimal regularization matrix Σ strongly depends on the parameter L, which defines the functional class $\mathbf{W}_\beta^2(L)$. In practice this parameter is hardly known. Therefore our next step is to construct a practically feasible data-driven method for adaptive penalization.

2 An adaptive estimator

The goal of adaptation is to chose the regularization matrix Σ in (5) based on the observations in order to minimize the covariance matrix $\mathbf{E}(\widehat{\theta}-\theta)(\widehat{\theta}-\theta)^T$. Theorem 1 plays an essential role in a such choice since it says that the second order risk of $\widehat{\theta}$ is controlled by $IMSE[H,\nu]$. Thus we see that in order to minimize the second order risk we have to minimize $IMSE[H,\nu]$ with respect to H. This functional depends on the nuisance parameters ν, which are of course not known. The ordinary way to perform minimization in this situation is based on the principle of unbiased risk estimation. The main idea is well known and commonly used in nonparametric estimation (see e.g. Akaike (1973), Mallows (1973)). Heuristic arguments for adaptive choice of H are the following. Let $\mu=(\theta_1,\ldots\theta_d,\nu_1,\ldots)^T$. Define an estimator of this vector by

$$\widehat{\mu}(H)=\arg\min_{\mu}\left\{\left\|Y-Q^T\mu\right\|^2+n\left\|H^{-1/2}(E-H)^{1/2}\mu\right\|^2\right\},\quad Q=\begin{pmatrix} Z\\ \Psi\end{pmatrix}. \tag{10}$$

Simple algebra reveals that $\widehat{\mu}(H)$ can be computed as

$$\widehat{\mu}(H)=[QQ^T+nH^{-1}(E-H)]^{-1}QY. \tag{11}$$

With $U(H)=Q^T[QQ^T+nH^{-1}(E-H)]^{-1}Q$ we easily see that

$$\mathbf{E}\left\|Y-Q^T\widehat{\mu}(H)\right\|^2=\mathbf{E}\left\|Q^T[\widehat{\mu}(H)-\mu]\right\|^2-2\sigma^2\mathrm{tr}\,U(H). \tag{12}$$

Noticing that by the law of large numbers

$$QQ^T\approx\begin{pmatrix} ZZ^T & 0\\ 0 & nE\end{pmatrix}$$

we arrive at

$$[QQ^T+nH^{-1}(E-H)]^{-1}QQ^T\approx\begin{pmatrix} E & 0\\ 0 & H\end{pmatrix}$$

and

$$U(H)\approx Z^T(ZZ^T)^{-1}Z+\Psi^T H\Psi/n.$$

So with this in mind we obtain omitting the terms of the order $O(1)$

$$\begin{aligned}\mathbf{E}\left\|Q^T[\widehat{\mu}(H)-\mu]\right\|^2&\approx\mathbf{E}\left\|\Psi^T(E-H)\nu\right\|^2+\sigma^2\mathbf{E}\mathrm{tr}\,U^2(H)\\ &\approx\; n\sum_{k=1}^{\infty}(1-H_{kk})^2\nu_k^2+\sigma^2\mathrm{tr}\,H^2+\sigma^2\mathrm{tr}\,Z^T(ZZ^T)^{-1}Z(1+2\mathrm{tr}H/n)\\ &\approx\; IMSE(H)n.\end{aligned} \tag{13}$$

Thus to minimize the risk of the estimator one could choose (cf. (12))

$$\widehat{H} = \arg\min_{h\in\mathcal{H}} \left\{ \left\| Y - Q^T\widehat{\mu}(H) \right\|^2 + 2\sigma^2 \mathrm{tr}\, U(H) \right\}, \tag{14}$$

where $\mathcal{H}$ is a set in $\mathbf{l}_2(0,\infty)$, which must not be very rich. For instance we will assume that $\mathcal{H}$ is a finite set. We discuss the required properties $\mathcal{H}$ in more detail in the sequel. Examples of commonly used classes $\mathcal{H}$ are projection smoothers with $H_{kk} = \mathbf{1}\{k \le w\}$, for integer parameter $w \in [1,n]$, the Pinsker (1980) or minimax filters (see Theorem 2)

$$H_{kk} = [1 - (k/w)^\beta]_+,\ w \in [1,n] \tag{15}$$

and smoothing spline Wahba (1990)

$$H_{kk} = (1 + \mu\lambda_k)^{-1},\ \mu > 0, \tag{16}$$

where λ_k are the eigen values of the boundary value problem (2). We would like to remind that in the last case the estimator has the form

$$\hat{\theta} = \arg\min_\theta \min_m \Big\{ \sum_{i=1}^n [Y_i - \theta^T Z_i - m(X_i)]^2 + \mu \int_0^1 [m^{(\beta)}(x)]^2\, dx \Big\}$$

and there exist very fast computational algorithms for finding $\hat{\theta}$ (see for more details Green and Silverman (1994) and Schimek (2000)).

Unfortunately in partial linear models the empirical risk $n(\widehat{\theta} - \theta)(\widehat{\theta} - \theta)^T$ is non-degenerate. This leads to some difficulties in evaluation of the performance of $\widehat{\theta}$ with $\widehat{H}$ from (14). To overcome these difficulties we use the idea of splitting by Nemirovskii (1998). Only a part of the whole sample say $Y_1, \ldots, Y_N$, where $N \ll n$ will be used to construct the matrix H. Once H is chosen the whole sample $Y_1, \ldots, Y_n$ is used to compute the estimator. Thus our main idea is to estimate the right-hand side in (14) based the data $Y_1, \ldots, Y_N$. We do this in the following way. Let

$$\widehat{\mu}_N(H) = \arg\min_\mu \left\{ \left\| Y - Q_N^T \mu \right\|_N^2 + N \left\| H^{-1/2}(E - H)^{1/2}\mu \right\|^2 \right\}$$

be the estimator of the parameter μ based on the data $Y_1, \ldots, Y_N$. With (12) and (13) we obtain

$$\begin{aligned}
\mathbf{E} \left\| Y - Q^T\widehat{\mu}(H) \right\|^2 + 2\sigma^2 \mathrm{tr}\, U(H) &\approx\ n \sum_{k=1}^\infty (1 - H_{kk})^2 \nu_k^2 + \sigma^2 \mathrm{tr}\, U^2(H) \\
&=\ \frac{n}{N} \left[N \sum_{k=1}^\infty (1 - H_{kk})^2 \nu_k^2 + \sigma^2 \mathrm{tr}\, U_N^2(H) \right] +\ \left(1 - \frac{n}{N}\right) \sigma^2 \mathrm{tr}\, U_N^2(H) \\
&\approx\ \frac{n}{N} \left[\mathbf{E} \left\| Y_N - Q_N^T \widehat{\mu}_N(H) \right\|_N^2 + 2\sigma^2 \mathrm{tr}\, U_N(H) \right] +\ \left(1 - \frac{n}{N}\right) \sigma^2 \mathrm{tr}\, U_N^2(H).
\end{aligned}$$

Thus we can replace (14) by

$$H^* = \arg\min_{H\in\mathcal{H}} \left\{ \left\| Y - Q_N^T \widehat{\mu}_N(H) \right\|_N^2 + 2\sigma^2 \mathrm{tr}\, U_N(H) - \sigma^2 \mathrm{tr}\, U_N^2(H)\,(1 - N/n) \right\} \qquad (17)$$

and the adaptive estimator of θ is defined now by

$$\theta^* = \arg\min_{\theta}\min_{\nu} \left\{ \left\| Y - Z^T\theta - \Psi^T\nu \right\|^2 + n \left\| (H^*)^{-1/2}(E - H^*)^{1/2}\nu \right\|^2 \right\}. \qquad (18)$$

Our further analysis is essentially based on some properties of the smoothing matrixes H. We will assume the set $\mathcal{H}$ is finite, its cardinality is less then $O(n)$ and uniformly in $H \in \mathcal{H}$:

$$\begin{aligned}
&0 \le H_{kk} \le 1,\\
&H_{kk} = 1, \text{ for } k = 0, \ldots, \beta - 1,\\
&\mathrm{tr}\, H \le C_{\mathcal{H}} \mathrm{tr}\, H^2,\\
&\mathrm{tr}^2 H \le C_{\mathcal{H}} n/\log n,\\
&\sum_{k>n}^{\infty} H_{kk}^2 k^2 \le C_{\mathcal{H}} n,
\end{aligned} \qquad (19)$$

where $C_{\mathcal{H}}$ is some constant.

In order to simplify some technical details we assume also that the regressors are such that for some sufficiently large constant C_Z the following conditions are hold

$$\max_{i\in[1,N]} Z_{ki}^2 \le \frac{C_Z}{N} \sum_{i=1}^{N} Z_{ki}^2 \le \frac{2C_Z}{n} \sum_{i=1}^{n} Z_{ki}^2, \quad \left\| (ZZ^T)^{-1} \right\| \sum_{k,i=1}^{n} Z_{ki}^2 \le C_Z. \qquad (20)$$

The following theorem is the main result of the paper.

Theorem 3 *Let ξ_i be Gaussian and $N = n/\log^{1+\kappa} n$, $\kappa > 0$. Then under conditions (19-20) uniformly in $m \in \mathbf{W}_\beta^2(L)$ and such that $\|m\|_q \le M < \infty$ as $n \to \infty$*

$$\mathbf{E}(\theta^* - \theta)(\theta^* - \theta)^T = (ZZ^T)^{-1}\{\sigma^2 + [1 + o(1)] \inf_{H\in\mathcal{H}} IMSE[H, \nu]\}.$$

This theorem could be interpreted as an oracle inequality (see Nemirovski (1998)) in the following way. Suppose we are allowed to make use of only the estimators $\hat{\theta}$ defined by (5), (6) with $H \in \mathcal{H}$. Let us assume that there is an oracle which says the nuisance function $m(\cdot)$. If we know this function or equivalently ν then according to Theorem 1 we can minimize the risk of $\hat{\theta}$ up to a second order term choosing $H^* = \arg\min_{H\in\mathcal{H}} IMSE[H, \nu]$. It is clear with this family of estimators we cannot do better.

On the other hand Theorem 3 says that we can achieve almost the same performance without oracles.

Remark 1. There is no other way to compute H^* from (17) except the complete search. Thus from a computational point of view it is better to have cardinality of $\mathcal{H}$ as small as possible. But if $\mathcal{H}$ is not sufficiently rich the risk of the adaptive estimator θ^* may increase substantially. So there is a compromise between computational and statistical efficiency. The simplest way to resolve this problem is to use so-called exponential grids. As an example consider the family of minimax smoothers defined by (15). Let us chose sufficiently small number ε, say $\varepsilon = 0.1$ and consider the grid of bandwidths

$$w_s = (1+\varepsilon)^s, \quad s = 0, \dots, \log n/\varepsilon.$$

Let $\mathcal{H}^\varepsilon$ be a corresponding class of smoothers having the elements

$$H_{kk} = [1 - (k/w_s)^\beta]_+, \quad s = 0, \dots, \log n/\varepsilon.$$

Cardinality of $\mathcal{H}^\varepsilon$ is $\log n/\varepsilon$. It is much smaller smaller then cardinality $\mathcal{H}$. On the other hand for any $H \in \mathcal{H}$ we can find $H^\varepsilon \in \mathcal{H}^\varepsilon$ such that uniformly in $\nu \in \mathbf{l}_2(0,\infty)$

$$IMSE[H^\varepsilon, \nu] \le (1+\varepsilon) IMSE[H, \nu].$$

Indeed let $H_{kk} = [1 - (k/w)^\beta]_+$ for some $w \in [1, n]$. Take $H_{kk} = [1 - (k/w^\varepsilon)^\beta]_+$, where $w^\varepsilon = \min\{w_s : w_s \ge w\}$. Since $H^\varepsilon_{kk} \ge H_{kk}$ we evidently have $\|(1-H^\varepsilon)\nu\|^2 \le \|(1-H)\nu\|^2$ and it is easy to see that $\|H^\varepsilon\|^2 \le (1+\varepsilon)\|H^\varepsilon\|^2$. Thus using $\mathcal{H}^\varepsilon$ instead of $\mathcal{H}$ we may have only a little increment of the risk but we improve significantly the computational efficiency. The same remark concerns of course the spline estimator.

Remark 2. In order to consruct θ^* we divided the sample in to two parts. From a practical point of view this idea is of course not very attractive. As a rule we use the estimator

$$\widehat{\theta} = \arg\min_\theta \min_\nu \{\|Y - Z^T\theta - \Psi^T\nu\|^2 + \sigma^2 n \|\widehat{H}^{-1/2}(E - \widehat{H})^{1/2}\nu\|^2\}$$

with $\widehat{H}$ from (14). So it would be very interesting to find out whether this estimator is adaptive in the sense of Theorem 3. Unfortunately, our arguments applied in the proof of this theorem cannot be used to answer this question because of a strong dependence between the estimators of parametric and nonparametric parts in the partial linear model.

3 Proof of Theorem 3

We start the proof of Theorem 3 with some auxiliary results. First using the Taylor formula we will find an asymptotic expansion for the risk of the estimator $\widehat{\theta}$ defined by (5). Rewrite (11) as

$$\widehat{\mu}(H) - \mu = nS^{-1}H^{-1}(E - H) + S^{-1}Q\xi, \tag{21}$$

where $S = A + B$, with

$$A = \begin{pmatrix} ZZ^T & 0 \\ 0 & nE + \sigma^2\Sigma^{-2} \end{pmatrix}, \quad B = \begin{pmatrix} 0 & Z\Psi^T \\ \Psi Z^T & \Psi\Psi^T - nE \end{pmatrix}. \tag{22}$$

In order to compute

$$S^{-1} = A^{-1/2}(E + A^{-1/2}BA^{-1/2})^{-1}A^{-1/2}. \tag{23}$$

we use the Taylor expansion with respect to $A^{-1/2}BA^{-1/2}$ in the right-hand side of the above equation. Denote for brevity $H = (E + \sigma^2\Sigma^{-2}/n)^{-1}$. Thus we have to check that the operator norm of the matrix

$$A^{-1/2}BA^{-1/2} = \begin{pmatrix} 0 & (ZZ^T)^{-1/2}Z\Psi^T H^{1/2}/\sqrt{n} \\ H^{1/2}\Psi Z^T(ZZ^T)^{-1/2}/\sqrt{n} & H^{1/2}(\Psi\Psi^T/n - E)H^{1/2} \end{pmatrix} \tag{24}$$

is sufficiently small. Denote also by $\#\mathcal{H}$ cardinality of $\mathcal{H}$ and $\|Z_s\|_N^2 = \sum_{i=1}^N Z_{si}^2$.

Lemma 1 *For any $x \le n$*

$$\mathbf{P}\Big\{\|A^{-1/2}BA^{-1/2}\|^2 > C(1+x)\mathrm{tr}^2 H/n\Big\} \le \#\mathcal{H}\exp(-x^2/C),$$

where C is a sufficiently large constant.

Proof is similar to Lemma 2 in Golubev, Härdle (2001) and omitted.

The next lemma gives an asymptotic expansion for the risk of $\widehat{\theta}$ defined by (5) with the penalization $\Sigma = H^{1/2}(E - H)^{-1/2}/\sqrt{n}$ depending on the data (Y_i, X_i).

Lemma 2 *Uniformly in $\|m\|_q \le M < \infty$ as $n \to \infty$ we have*

$$\mathbf{E}\,(\widehat{\theta} - \theta)(\widehat{\theta} - \theta)^T = \mathbf{E}\,\Delta_H\Delta_H^T + n^{-1}\|(ZZ^T)^{-1}\|^{-1}O(E), \tag{25}$$

where

$$\begin{aligned} \Delta_H \;=\;& (ZZ^T)^{-1}Z\xi + [E + o(E)](ZZ^T)^{-1}Z\Psi^T(E - H)\nu \\ &+ n^{-1}[E + o(E)](ZZ^T)^{-1}Z\Psi^T H\Psi Z^T(ZZ^T)^{-1}Z\xi \\ &- n^{-1}[E + o(E)](ZZ^T)^{-1}Z\Psi^T H\Psi\xi \\ &+ n^{-1}[E + o(E)](ZZ^T)^{-1}Z\Psi^T H(\Psi\Psi^T/n - E)H\Psi\xi, \end{aligned} \tag{26}$$

$O(E)$ is a $d \times d$-matrix with the bounded operator norm and $o(E)$ is a $d \times d$-matrix with the operator norm tending to 0 as $n \to \infty$.

Proof. Notice that in view of (5) $\|Z^T(\widehat{\theta}-\theta)\| \le \|\xi + \Psi^T\nu\|$. Therefore for any $\delta > 0$ we have

$$\mathbf{E}\|\widehat{\theta}-\theta\|^{2+2\delta} \le C(\delta) n^{1+\delta} \|(ZZ^T)^{-1/2}\|^{-1-\delta}$$

and we get by the Hölder inequality and Lemma 1 with $x = (2C\log(n\#\mathcal{H}))^{1/2}$

$$\begin{aligned}\mathbf{E}\,(\widehat{\theta}-\theta)(\widehat{\theta}-\theta)^T &= \mathbf{E}\,(\widehat{\theta}-\theta)(\widehat{\theta}-\theta)^T \mathbf{1}\left\{\|A^{-1/2}BA^{-1/2}\| \le \varepsilon\right\} \\ &\quad + \mathbf{E}\,(\widehat{\theta}-\theta)(\widehat{\theta}-\theta)^T \mathbf{1}\left\{\|A^{-1/2}BA^{-1/2}\| > \varepsilon\right\} \\ &\le \mathbf{E}\,(\widehat{\theta}-\theta)(\widehat{\theta}-\theta)^T \mathbf{1}\left\{\|A^{-1/2}BA^{-1/2}\| \le \varepsilon\right\} \\ &\quad + n\|(ZZ^T)^{-1/2}\|^{-2}\mathbf{P}^{\delta/(1+\delta)}\left\{\|A^{-1/2}BA^{-1/2}\| > \varepsilon\right\} O(E) \\ &\le \mathbf{E}\,(\widehat{\theta}-\theta)(\widehat{\theta}-\theta)^T \mathbf{1}\left\{\|A^{-1/2}BA^{-1/2}\| \le \varepsilon\right\} + n^{-1}\|(ZZ^T)^{-1}\|^{-1} O(E),\end{aligned} \tag{27}$$

Next one obtains by (23), (24) and by the Taylor formula when $\left\|A^{-1/2}BA^{-1/2}\right\| \le \varepsilon$

$$\begin{aligned}S^{-1} &= \begin{pmatrix} (ZZ^T)^{-1} & 0 \\ 0 & n^{-1}H \end{pmatrix} - \frac{1}{n}\begin{pmatrix} 0 & (ZZ^T)^{-1}Z\Psi^T H \\ H\Psi Z^T(ZZ^T)^{-1} & H(\Psi\Psi^T/n - E) \end{pmatrix} \\ &\quad + (E + \varepsilon O(E)) n^{-1}(ZZ^T)^{-1}Z\Psi^T H \begin{pmatrix} \Psi Z^T(ZZ^T)^{-1} & (\Psi\Psi^T/n - E) \\ * & * \end{pmatrix},\end{aligned} \tag{28}$$

where $*$ denotes a matrix that is not needed in further calculations. Thus once again using Lemma 1, which says that $\mathbf{P}\left\{\|A^{-1/2}BA^{-1/2}\| \ge \varepsilon\right\}$ is exponentially small, we arrive from (21) and (26-28) at the assertion of the lemma. □

In the sequel we will use the following very simple auxiliary fact. Let V be a finite set in $\mathbf{l}_2(0,\infty)$ with cardinality $\#V$ and let η_k be zero mean random variables, such that for any given $v \in V$

$$D(v) = \mathbf{E}\Big(\sum_{k=0}^{\infty} v_k \eta_k\Big)^2 = \sum_{k,l=0}^{\infty} v_k v_l \mathbf{E}\,\eta_k\eta_l < \infty.$$

Suppose now that $v \in V$ are random variables depending on $\eta_k,\ k = 0, \dots$ and we want to evaluate from above $\mathbf{E}\Big(\sum_{k=0}^{\infty} v_k\eta_k\Big)^2$. The next lemma provides a solution of this problem.

Lemma 3 *Assume that for* $\lambda = \sqrt{\log(S\#V)/2}$ *($S \ge 1$) and for any given* $v \in V$

$$\mathbf{E}\exp\Big[\frac{\lambda}{\sqrt{D(v)}}\sum_{k=0}^{\infty} v_k\eta_k\Big] \le \exp(2\lambda^2). \tag{29}$$

Then

$$\mathbf{E}\Big(\sum_{k=0}^{\infty} v_k\eta_k\Big)^2 \le 8\log(S\#V)\mathbf{E}D(v) + 2\max_{v\in V} D(v)/S. \tag{30}$$

Proof. By the Chebyshev inequality for any $Q > 0$ we have

$$\mathbf{E}\Big[\sum_{k=0}^{\infty} v_k\eta_k\Big]^2 = \mathbf{E}\,D(v)\Big[\frac{1}{\sqrt{D(v)}}\sum_{k=0}^{\infty} v_k\eta_k\Big]^2 \le \mathbf{E}\,D(v)\max_{v\in V}\Big[\frac{1}{\sqrt{D(v)}}\sum_{k=0}^{\infty} v_k\eta_k\Big]^2$$
$$\le Q^2\mathbf{E}\,D(v) + \max_{v\in V} D(v)\sum_{v\in V}\mathbf{P}\Big\{\Big|\frac{1}{\sqrt{D(v)}}\sum_{k=0}^{\infty} v_k\eta_k\Big| > Q\Big\}$$
$$\le Q^2\mathbf{E}\,D(v) + 2\#V\exp(-\lambda Q + 2\lambda^2)\max_{v\in V} D(v).$$

Finally choosing $Q = 2\sqrt{2\log(S\#V)}$ we arrive at (30). □

Lemma 4 *Let $\eta_{sk} = \sum_{i=1}^{N} Z_{si}\psi_k(X_i)$. Then uniformly in $\nu \in \mathbf{W}^2_\beta(L)$ and in $H \in \mathcal{H}$*

$$\mathbf{E}\Big[\sum_{k=1}^{\infty}(1-H_{kk})\nu_k\eta_{sk}\Big]^2 \le C\log(n\#\mathcal{H})\|Z_s\|_N^2\,\Big[\mathbf{E}\sum_{k=1}^{\infty}(1-H_{kk})^2\nu_k^2 + \frac{L}{n}\Big]. \tag{31}$$

Proof. Since $|1-H_{kk}| \le 1$ one obtains by the Cauchy-Schwartz inequality that for any $M > 1$

$$\mathbf{E}\Big[\sum_{k=1}^{\infty}(1-H_{kk})\nu_k\eta_{sk}\Big]^2 \le 2\mathbf{E}\Big[\sum_{k=1}^{M}(1-H_{kk})\nu_k\eta_{sk}\Big]^2 + 2CM^{-2\beta+1}L\|Z_s\|_N^2. \tag{32}$$

Next we apply Lemma 3 to the first term in right-hand side of (32). So we put

$$v_k = (1-H_{kk})\nu_k, \quad \eta_k = \sum_{i=1}^{N} Z_{si}\psi_k(X_i), \quad D(v) = \|Z_s\|_N^2\sum_{k=0}^{M}(1-H_{kk})^2\nu_k^2.$$

Check now (29). With $\lambda = \sqrt{\log(n\#\mathcal{H})/2}$ and $M = N/[C_Z\log(n\#\mathcal{H})]$ we get by the Cauchy-Schwartz inequality and (20)

$$\frac{\lambda\max_i|Z_{si}|\sum_{k=1}^{M}(1-H_{kk})|\nu_k|}{\|Z_s\|_N\Big(\sum_{k=1}^{M}(1-H_{kk})^2\nu_k^2\Big)^{1/2}} \le \frac{\lambda\max_i|Z_{si}|M^{1/2}}{\|Z_s\|_N} \le \lambda C_Z^{1/2}M^{1/2}N^{-1/2} < \frac{1}{2}.$$

With this in mind since X_i are independent we have by the Taylor formula

$$\mathbf{E}\exp\Big[\frac{\lambda}{\sqrt{D(v)}}\sum_{k=0}^{\infty} v_k\eta_k\Big] = \prod_{i=1}^{N}\mathbf{E}\exp\Big\{\frac{\lambda Z_{si}}{\sqrt{D(v)}}\sum_{k=1}^{M}(1-H_{kk})\nu_k\psi_k(X_i)\Big\}$$
$$\le\ (1+\sqrt{\mathrm{e}}\lambda^2/2) \le \exp(\lambda^2).$$

Hence using (32) and Lemma 3 we finish the proof. □

Lemma 5 *Let ξ_i be i.i.d. $\mathcal{N}(0,\sigma^2)$. Then*

$$\mathbf{E}\Big\{\sum_{k=1}^{\infty} H_{kk}\Big[\Big(\sum_{i=1}^{N} Z_{si}\psi_k(X_i)\Big)^2 - \|Z_s\|_N^2\Big]\Big\}^2 \le C\log(n\#\mathcal{H})\|Z_s\|_N^4\mathbf{E}\,\mathrm{tr}\,H^2, \quad (33)$$

$$\mathbf{E}\Big\{\sum_{k=1}^{\infty} H_{kk}\Big[\Big(\sum_{i=1}^{N} \xi_i\psi_k(X_i)\Big)^2 - N\sigma^2\Big]\Big\}^2 \le C\log(n\#\mathcal{H})N^2\mathbf{E}\,\mathrm{tr}\,H^2, \quad (34)$$

$$\mathbf{E}\Big\{\sum_{k=1}^{\infty} H_{kk}^2\Big[\Big(\sum_{i=1}^{N} \xi_i\psi_k(X_i)\Big)^2 - N\sigma^2\Big]\Big\}^2 \le C\log(n\#\mathcal{H})N^2\mathbf{E}\,\mathrm{tr}\,H^2, \quad (35)$$

$$\mathbf{E}\Big\{\sum_{k=1}^{\infty} H_{kk}\sum_{i=1}^{N}[\psi_k^2(X_i)-1]\Big\}^2 \le C\log(n\#\mathcal{H})N\mathbf{E}\,\mathrm{tr}\,H, \quad (36)$$

$$\mathbf{E}\Big\{\sum_{k=1}^{\infty} H_{kk}^2\sum_{i=1}^{N}[\psi_k^2(X_i)-1]\Big\}^2 \le C\log(n\#\mathcal{H})N\mathbf{E}\,\mathrm{tr}\,H^2, \quad (37)$$

$$\mathbf{E}\Big\{\sum_{k=0}^{\infty} H_{kk}\sum_{i\neq j}^{N}\psi_k(X_j)\psi_k(X_i)\xi_i\xi_j\Big\}^2 \le C\log(n\#\mathcal{H})N\mathbf{E}\,\mathrm{tr}\,H^2, \quad (38)$$

$$\mathbf{E}\Big\{\sum_{k=0}^{\infty} H_{kk}\sum_{j,i=1}^{N}[\psi_k(X_i)\psi_k(X_j)-\delta_{ij}][Z_N^T(Z_NZ_N^T)^{-1}Z_N]_{ij}\Big\}^2 \quad (39)$$
$$\le\ C\log(n\#\mathcal{H})\mathbf{E}\,\mathrm{tr}\,H^2.$$

Proof. We prove only the first inequality since the others can be checked in the similar way. Once again use Lemma 3. Putting

$$v_k = H_{kk}, \quad \eta_k = \Big(\sum_{i=1}^{N} Z_{si}\psi_k(X_i)\Big)^2 - \|Z_s\|_N^2$$

we have by (19 - 20) that for any given v

$$D(v) = \mathbf{E}\Big(\sum_{k=0}^{\infty} v_k\eta_k\Big)^2 = 2[1+o(1)]\,\mathrm{tr}\,H^2\|Z_s\|_N^4.$$

Since $\mathrm{tr}\,H^2 \le n$ we choose $S = n^2$. Thus to complete the proof of the lemma it remains to check (29). Using the Cauchy-Schwartz inequality one obtains

$$\mathbf{E}\,\exp\Big\{\frac{\lambda}{\sqrt{D(v)}}\sum_{k=0}^{\infty} v_k\eta_k\Big\} \quad (40)$$

$$\leq \mathbf{E}^{1/2} \exp\Big\{\frac{4\lambda}{\sqrt{D(v)}} \sum_{j<i}^{N} Z_{si} Z_{sj} \sum_{k=0}^{\infty} v_k \psi_k(X_i)\psi_k(X_j)\Big\}$$
$$\times\ \mathbf{E}^{1/2} \exp\Big\{\frac{2\lambda}{\sqrt{D(v)}} \sum_{i=1}^{N} Z_{si}^2 \sum_{k=0}^{\infty} v_k[\psi_k^2(X_i) - 1]\Big\}.$$

Notice that in view of (19–20) one obtains for $\lambda = \sqrt{\log(n^2\#\mathcal{H})}$

$$\frac{\lambda|Z_{si}Z_{sj}|}{\sqrt{D(v)}}\Big|\sum_{k=0}^{\infty} v_k\psi_k(X_i)\psi_k(X_j)\Big| = o(1), \quad \frac{\lambda Z_{si}^2}{\sqrt{D(v)}}\Big|\sum_{k=0}^{\infty} v_k[\psi_k^2(X_i)-1]\Big| = o(1).$$

Therefore we can use the Taylor formula to compute the right-hand side of (40). Hence acting in this way we arrive at

$$\mathbf{E}\ \exp\Big\{\frac{\lambda}{\sqrt{D(v)}}\sum_{k=0}^{\infty} v_k\eta_k\Big\} \leq \exp(2\lambda^2).$$

Now (30) implies the inequality (33). □

Lemma 6 *Let* $\eta_{ks} = N^{-1/2}\sum_{i=1}^{N}[\psi_k(X_i)\psi_l(X_i) - \delta_{kl}]$. *For any* $M > 1$ *uniformly in* $\nu \in \mathbf{W}_2^{\beta}(L)$

$$\mathbf{E}\Big|\sum_{k,l=0}^{\infty} \nu_k\nu_l(1-H_{kk})(1-H_{ll})\eta_{ks}\Big| \leq CM\log^{1/2}(M)\Big[\mathbf{E}\sum_{k=1}^{\infty}(1-H_{kk})^2\nu_k^2 + LM^{-2\beta}\Big] \quad (41)$$

and

$$\mathbf{E}\sum_{k,s=0}^{\infty} H_{kk}H_{ss}\eta_{ks}^2 \leq C\log(n)\mathbf{E}\,\mathrm{tr}^2 H. \quad (42)$$

Let $\zeta_{ks} = \sum_{i,j=1}^{N}\psi_k(X_i)\psi_s(X_j)[Z_N^T(Z_NZ_N^T)^{-1}Z_N]_{ij}$. *Then uniformly in* $\nu \in \mathbf{W}_2^{\beta}(L)$

$$\mathbf{E}\Big(\sum_{k,s=0}^{\infty}|\nu_s|H_{kk}\zeta_{ks}\Big)^2 \leq C\log n\Big(\sum_{s=0}^{\infty}|\nu_s|\Big)^2\mathbf{E}\,\mathrm{tr}^2 H + CLn^{-2\beta+1}. \quad (43)$$

Proof. Note that $\psi_k(X_i)\psi_l(X_i) - \delta_{kl}$ are i.i.d. bounded random variables having zero mean. Therefore

$$\mathbf{P}\Big\{|\eta_{ks}| \geq x\Big\} \leq \exp(-Cx^2). \quad (44)$$

On the other hand one obtains by the Cauchy-Schwartz inequality

$$\begin{aligned}
&\mathbf{E}\Big|\sum_{k,l=0}^{\infty}\nu_k\nu_l(1-H_{kk})(1-H_{ll})\eta_{kl}\Big| \\
&\leq \Big(\sum_{k=0}^{M}|\nu_k||1-H_{kk}|\Big)^2\max_{k,l\leq M}|\eta_{kl}| + L\Big(\sum_{k,l>M}k^{-2\beta}l^{-2\beta}\mathbf{E}\,\eta_{kl}^2\Big)^{1/2} \\
&\leq MQ\mathbf{E}\sum_{k=0}^{M}\nu_k^2(1-H_{kk})^2 + CL\sum_{k,l\leq M}\mathbf{P}\Big\{|\eta_{kl}|\geq Q\Big\} + CLM^{-2\beta+1}.
\end{aligned}$$

Putting $Q=\sqrt{C'\log(M)}$, where C' is a sufficiently large constant, we arrive with (44) at (41).

The second inequality (42) can be proved in the similar way. Using the Cauchy-Schwartz inequality, (19) and (44) we have

$$\begin{aligned}
\mathbf{E}\sum_{k,s=0}^{\infty}H_{kk}H_{ss}\eta_{ks}^2 &\leq \mathbf{E}\max_{k,s\leq n}\eta_{ks}^2\sum_{k,s=0}^{n}H_{kk}H_{ss} + \mathbf{E}\sum_{k,s>N}H_{kk}H_{ss}\eta_{ks}^2 \\
&\leq Q^2\mathbf{E}\,\mathrm{tr}^2H + n^2\mathbf{P}\,\{\max_{k,s\leq n}|\eta_{ks}|>Q\} + \mathbf{E}\Big(\sum_{k>n}H_{kk}^2k^2\Big)\Big(\sum_{k,s>n}k^{-2}s^{-2}\eta_{ks}^4\Big)^{1/2} \\
&\leq Q^2\mathbf{E}\,\mathrm{tr}^2H + n^4\exp(-CQ^2) + C.
\end{aligned}$$

We complete the proof of the lemma choosing $Q=2\sqrt{\log n/C}$. The proof of (43) is similar and we omit it. □

Now we are ready to evaluate the performance of the estimator θ^*.

Lemma 7 *Let*

$$\lim_{n\to\infty}\frac{N\log n}{n}=0. \tag{45}$$

Then uniformly in $m\in\mathbf{W}_\beta^2(L)$ as $n\to\infty$

$$\mathbf{E}(\theta^*-\theta)(\theta^*-\theta)^T=(ZZ^T)^{-1}\Big[\sigma^2+(1+o(1))\mathbf{E}\Big(\|(E-H^*)\nu\|^2+\frac{\sigma^2}{n}\mathrm{tr}H^{*2}\Big)\Big]. \tag{46}$$

Proof. In order to compute the right-hand side of (25) we represent ΨZ^T, $\Psi\xi$, $Z\xi$ in the following form

$$\Psi Z^T = \sum_{i=1}^{N}\psi_k(X_i)Z_{si} + \sum_{i=N+1}^{n}\psi_k(X_i)Z_{si} = \Psi Z_0^T + \Psi Z_1^T,$$

$$\Psi\xi = \sum_{i=1}^{N}\psi_k(X_i)\xi_i + \sum_{i=N+1}^{n}\psi_k(X_i)\xi_i = \Psi\xi_0 + \Psi\xi_1,$$
$$Z\xi = \sum_{i=1}^{N} Z_{ki}\xi_i + \sum_{i=N+1}^{n} Z_{ki}\xi_i = Z\xi_0 + Z\xi_1.$$

and note that H^* does not depend on ΨZ_1^T, $\Psi\xi_1$ and $Z\xi_1$. The dependence between H^* and ΨZ_0^T, $\Psi\xi_0$, $Z\xi_0$ can be evaluated by Lemmas 4 – 6. So we have for example by Lemma 4 and (45), (20)

$$\begin{aligned}
&\mathbf{E}\,[Z\Psi^T(E-H^*)\nu][Z\Psi^T(E-H^*)\nu]^T \\
&= \mathbf{E}\,[(Z\Psi_0^T + Z\Psi_1^T)(E-H^*)\nu][(Z\Psi_0^T + Z\Psi_1^T)(E-H^*)\nu]^T \\
&= \mathbf{E}\,[Z\Psi_0^T(E-H^*)\nu][Z\Psi_0^T(E-H^*)\nu]^T + \mathbf{E}\,[Z\Psi_1^T(E-H^*)\nu][Z\Psi_1^T(E-H^*)\nu]^T \\
&\le \Big[\sum_{i=N+1}^{n} Z_{ki}Z_{si} + O(E)\log n \sum_{s=1}^{d}\sum_{i=1}^{N} Z_{si}^2\Big]\mathbf{E}\sum_{k=0}^{\infty}(1-H_{kk}^*)^2\nu_k^2 + \frac{C\log n}{n}\sum_{s=1}^{d}\|Z_s\|_N^2 \\
&\le \Big[ZZ^T + O(E)\frac{N\log n}{n}\sum_{s=1}^{d}\sum_{i=1}^{n} Z_{si}^2\Big]\mathbf{E}\sum_{k=0}^{\infty}(1-H_{kk}^*)^2\nu_k^2 + \frac{C\log n}{n}\sum_{s=1}^{d}\|Z_s\|_N^2 \\
&\le ZZ^T[E+o(E)]\Big[\mathbf{E}\sum_{k=0}^{\infty}(1-H_{kk}^*)^2\nu_k^2 + \frac{o(1)}{n}\Big].
\end{aligned}$$

Similar arguments can be used in order to compute the remainder terms in the right-hand side of (25). □

Proof of Theorem 3. Denote

$$L_N[H] = \big\|Y - Q_N^T\widehat{\mu}_N(H)\big\|_N^2 + 2\sigma^2 \mathrm{tr}\, U_N(H) - (1-N/n)\,\sigma^2 \mathrm{tr}\, U_N^2(H).$$

In view of Lemma 7 it suffices to show that uniformly in $H \in \mathcal{H}$

$$\mathbf{E}L_N[H] = [1+o(1)]N\,\mathbf{E}\Big[\|(E-H)\nu\|^2 + \sigma^2\mathrm{tr}H^2/n\Big] + R, \tag{47}$$

where R is a constant which does not depend on H. Let $S_N = Q_NQ_N^T + NH^{-1}(E-H) = A_N + B_N$, with

$$A_N = \begin{pmatrix} Z_NZ_N^T & 0 \\ 0 & NH^{-1}\end{pmatrix}, \quad B_N = \begin{pmatrix} 0 & Z_N\Psi_N^T \\ \Psi_N Z_N^T & \Psi_N\Psi_N^T - NE\end{pmatrix}.$$

Then by Lemma 1 (see also (28)) we have

$$\begin{aligned}
S_N^{-1} &= \begin{pmatrix} (Z_NZ_N^T)^{-1} & 0 \\ 0 & N^{-1}H\end{pmatrix} \\
&\quad - [E+o(E)]N^{-1}\begin{pmatrix} 0 & (Z_NZ_N^T)^{-1}Z_N\Psi_N^TH \\ H\Psi_NZ_N^T(Z_NZ_N^T)^{-1} & H(\Psi_N\Psi_N^T/N - E)H\end{pmatrix}.
\end{aligned} \tag{48}$$

Therefore

$$S_N^{-1}Q_NQ_N^T = \begin{pmatrix} E & 0 \\ 0 & H \end{pmatrix} \tag{49}$$
$$- [E + o(E)] \begin{pmatrix} (ZZ^T)^{-1}Z\Psi_N^T H\Psi_N Z^T/N & -(ZZ^T)^{-1}Z\Psi_N^T(E-H) \\ 0 & H\Psi_N Z^T(ZZ^T)^{-1}Z\Psi_N^T/N \end{pmatrix}.$$

and

$$Q_NS_N^{-1}Q_N^T = U_N(H) = Z_N^T(Z_NZ_N^T)^{-1}Z_N + N^{-1}\Psi_N H\Psi_N^T \tag{50}$$
$$+ N^{-1}[E+o(E)][Z_N^T(Z_NZ_N^T)^{-1}Z_N\Psi_N H\Psi^T + \Psi_N^T H\Psi_N Z_N^T(Z_NZ_N^T)^{-1}Z_N$$
$$+ \Psi_N^T H(\Psi_N\Psi_N^T/N - E)H\Psi_N].$$

First we estimate the traces of $U_N(H)$ and $U_N^2(H)$. By a simple algebra we obtain

$$N^{-1}\mathbf{E}\ \mathrm{tr}\,\Psi_N^T H\Psi_N = \mathbf{E}\Big\{\mathrm{tr}\,H + \frac{1}{N}\sum_{i=1}^N\sum_{k=0}^\infty H_{kk}[\psi_k^2(X_i)-1]\Big\}, \tag{51}$$

$$N^{-2}\mathbf{E}\,\mathrm{tr}\,\Psi_N^T H\Psi_N\Psi_N^T H\Psi_N \tag{52}$$
$$= N^{-1}\mathbf{E}\ \mathrm{tr}\,\Psi_N^T H^2\Psi_N + N^{-1}\mathbf{E}\ \mathrm{tr}\,\Psi_N^T H(\Psi_N\Psi_N^T/N - E)H\Psi_N$$
$$= (1+N^{-1})\mathbf{E}\,\mathrm{tr}\,H^2 + \mathbf{E}\,\frac{2}{N}\sum_{i=1}^N\sum_{k=0}^\infty H_{kk}^2[\psi_k^2(X_i)-1]$$
$$+ \frac{1}{N}\sum_{k,s=0}^\infty H_{kk}H_{ss}\Big[\frac{1}{\sqrt{N}}\sum_{i=1}^N \psi_k(X_i)\psi_k(X_i) - \delta_{ks}\Big]^2$$

and

$$N^{-1}\mathbf{E}\,\mathrm{tr}\,Z_N^T(Z_NZ_N^T)^{-1}Z_N\Psi_N^T H\Psi_N = N^{-1}\mathrm{tr}\,Z_N^T(Z_NZ_N^T)^{-1}Z_N\mathbf{E}\,\mathrm{tr}\,H \tag{53}$$
$$+\frac{1}{N}\sum_{k=0}^\infty H_{kk}\sum_{j,i=1}^N[\psi_k(X_i)\psi_k(X_j)-\delta_{ij}][Z_N^T(Z_NZ_N^T)^{-1}Z_N]_{ij}.$$

Since $\mathrm{tr}\,Z_N^T(Z_NZ_N^T)^{-1}Z_N \le C$ we obtain by (36), (39) , (42) and (50-53)

$$\mathbf{E}\,\mathrm{tr}\,U_N(H) = [1+o(1)]\,\mathbf{E}\,\mathrm{tr}\,H + \mathrm{tr}\,Z_N^T(Z_NZ_N^T)^{-1}Z_N, \tag{54}$$
$$\mathbf{E}\,\mathrm{tr}\,U_N^2(H) = [1+o(1)]\,\mathbf{E}\,\mathrm{tr}\,H^2 + \mathrm{tr}\,Z_N^T(Z_NZ_N^T)^{-1}Z_N.$$

Next notice

$$\|Y - Q_N^T\widehat{\mu}_N(H)\|_N^2 = \|(E - Q_N^TS_N^{-1}Q_N)\xi_N + (E - Q_N^TS_N^{-1}Q_N)Q_N^T\mu\|_N^2 \tag{55}$$
$$= \|(E - Q_N^TS_N^{-1}Q_N)\xi_N\|_N^2 + \|Q_N^T(E - S_N^{-1}Q_NQ_N^T)\mu\|_N^2$$
$$+ 2\xi_N^T(E - Q_NS_N^{-1}Q_N^T)Q_N^T(E - S_N^{-1}Q_NQ_N^T)\mu.$$

The first term in the right-hand side of (55) is evaluated with (48) as

$$\begin{aligned} \mathbf{E}\|(E-Q_N^TS_N^{-1}Q_N)\xi_N\|_N^2 - N\sigma^2 &= -2\mathbf{E}\,\xi_N^TQ_N^TS_N^{-1}Q_N\xi_N \\ &\quad + \mathbf{E}\,\xi_N^TQ_NS_N^{-1}Q_N^TQ_N^TS_N^{-1}Q_N\xi_N \\ &= -2\mathbf{E}\,\xi_N^TZ_N^T(Z_NZ_N^T)^{-1}Z_N\xi_N - 2[1+o(1)]N^{-1}\mathbf{E}\,\xi_N^T\Psi_N^TH\Psi_N\xi_N \\ &\quad + \mathbf{E}\,\xi_N^T[Z_N^T(Z_NZ_N^T)^{-1}Z_N + [1+o(1)]N^{-1}\Psi_N^TH\Psi_N]^2\xi_N. \end{aligned} \tag{56}$$

Using (35) we get

$$N^{-1}\mathbf{E}\,\xi_N^T\Psi_N^TH\Psi_N\xi_N = \mathbf{E}\sum_{k=0}^{\infty}H_{kk}\frac{1}{N}\sum_{i,j=1}^{N}\psi_k(X_j)\psi_k(X_i)\xi_i\xi_j \tag{57}$$

$$= \sigma^2\mathbf{E}\sum_{k=1}^{\infty}H_{kk} + \mathbf{E}\sum_{k=0}^{\infty}H_{kk}\Big[\Big(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\psi_k(X_i)\xi_i\Big)^2 - \sigma^2\Big] = [1+o(1)]\,\sigma^2\mathbf{E}\sum_{k=0}^{\infty}H_{kk}.$$

Next with (35), (42) we obtain

$$\begin{aligned} N^{-2}\mathbf{E}\,\xi_N^T(\Psi_N^TH\Psi_N)^2\xi_N &= \mathbf{E}\ \mathrm{tr}\,H^2 + \mathbf{E}\sum_{k=0}^{\infty}H_{kk}^2\Big[\Big(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\psi_k(X_i)\xi_i\Big)^2 - \sigma^2\Big] \\ &\quad + \frac{1}{\sqrt{N}}\mathbf{E}\sum_{s,k=0}^{\infty}H_{kk}H_{ss}\Big(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\psi_k(X_i)\xi_i\Big)\Big(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\psi_s(X_i)\xi_i\Big) \\ &\quad \times\Big(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}[\psi_k(X_i)\psi_s(X_i)-\delta_{ks}]\Big) = (1+o(1))\mathbf{E}\ \mathrm{tr}\,H^2 \\ &\quad + \frac{O(1)}{\sqrt{N}}\mathbf{E}\Big[\sum_{s,k=0}^{\infty}H_{kk}\Big(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\psi_k(X_i)\xi_i\Big)^2\Big]^2 \\ &\quad + \frac{O(1)}{\sqrt{N}}\mathbf{E}\sum_{s,k=0}^{\infty}H_{kk}H_{ss}\Big(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}[\psi_k(X_i)\psi_s(X_i)-\delta_{ks}]\Big)^2 = (1+o(1))\mathbf{E}\ \mathrm{tr}\,H^2. \end{aligned} \tag{58}$$

Therefore noticing that $\mathrm{tr}\,Z_N^T(Z_NZ_N^T)^{-1}Z_N \le C$ we arrive with (56)–(58) at

$$\mathbf{E}\,\|(E-Q_N^TS_N^{-1}Q_N)\xi_N\|_N^2 = N\sigma^2 + (1+o(1))\mathbf{E}\,(\mathrm{tr}\,H^2 - 2\,\mathrm{tr}H) \tag{59}$$

$$- \mathrm{tr}\,Z_N^T(Z_NZ_N^T)^{-1}Z_N. \tag{60}$$

Our next step is to estimate $\mathbf{E}\,\|Q_N^T(E-S_N^{-1}Q_NQ_N^T)\mu\|_N^2$ in (55). Starting with the first order term (see (49)) we have by (41) with $M = N/\log^{1/2+\delta}(n)$

$$\mathbf{E}\|\Psi^T(E-H)\nu\|_N^2 = \mathbf{E}\sum_{i=1}^{N}\Big[\sum_{k=1}^{\infty}(1-H_{kk})\nu_k\psi_k(X_i)\Big]^2 \tag{61}$$

$$
\begin{aligned}
&= N\mathbf{E}\sum_{k=0}^{\infty}(1-H_{kk})^2\nu_k^2 + \mathbf{E}\sum_{k,l=\beta}^{\infty}(1-H_{kk})\nu_k\nu_l(1-H_{ll})\sum_{i=1}^{N}[\psi_k(X_i)\psi_l(X_i)-\delta_{kl}] \\
&= [1+o(1)]N\,\mathbf{E}\sum_{k=1}^{\infty}(1-H_{kk})^2\nu_k^2 + \log^{1+\delta}(n)/N.
\end{aligned}
$$

The remainder terms in $\mathbf{E}\,\|Q_N^T(E-S_N^{-1}Q_NQ_N^T)\mu\|_N^2$ associated with the second order terms in (49) are evaluated by the similar arguments. We have by (33) (replacing there Z_{si} by $[Z_N^T(Z_NZ_N^T)^{-1}Z_N]_{si}$ and by $[Z_N\theta]_i$)

$$
\begin{aligned}
&N^{-2}\mathbf{E}\,\|Z_N^T(Z_NZ_N^T)^{-1}Z_N\Psi_NH\Psi_N^TZ_N^T\theta\|_N^2 \\
&\quad= N^{-2}\mathbf{E}\sum_{s=1}^{N}\Big\{\sum_{k=0}^{\infty}H_{kk}\Big(\sum_{i=1}^{N}\psi_k(X_i)[Z_N^T(Z_NZ_N^T)^{-1}Z_N]_{si}\Big)\Big(\sum_{j=1}^{N}\psi_k(X_j)[Z_N\theta]_j\Big)\Big\}^2 \\
&\quad\le 4N^{-2}\mathbf{E}\sum_{s=1}^{N}\Big\{\sum_{k=0}^{\infty}H_{kk}\Big(\sum_{i=1}^{N}\psi_k(X_i)[Z_N^T(Z_NZ_N^T)^{-1}Z_N]_{si}\Big)^2\Big\}^2 \\
&\qquad+ 4N^{-2}\mathbf{E}\sum_{s=1}^{N}\Big\{\sum_{k=0}^{\infty}H_{kk}\Big(\sum_{j=1}^{N}\psi_k(X_j)[Z_N\theta]_j\Big)^2\Big\}^2 \\
&\quad\le CN^{-2}\log(n\#\mathcal{H})\Big\{\|Z_N\theta\|^2+\sum_{i,s=1}^{N}[Z_N^T(Z_NZ_N^T)^{-1}Z_N]_{si}^2\Big\}\mathbf{E}\,\mathrm{tr}^2H = o(1)\mathbf{E}\,\mathrm{tr}H,
\end{aligned}
$$

$$
\begin{aligned}
&\mathbf{E}\,\|Z_N^T(Z_NZ_N^T)^{-1}Z_N\Psi_N^T(E-H)\nu\|_N^2 \\
&\quad= \mathbf{E}\sum_{i=1}^{N}\Big\{\sum_{k=1}^{\infty}(1-H_{kk})\nu_k\Big(\sum_{j=1}^{N}\psi_k(X_j)[Z_N^T(Z_NZ_N^T)^{-1}Z_N]_{ij}\Big)\Big\}^2 \\
&\quad\le \log(n\#\mathcal{H})\sum_{i,j=1}^{N}[Z_N^T(Z_NZ_N^T)^{-1}Z_N]_{ij}^2\Big[\mathbf{E}\sum_{k=1}^{\infty}(1-H_{kk})^2\nu_k^2+\frac{C}{n}\Big] \\
&\quad= o(1)N\Big[\mathbf{E}\sum_{k=0}^{\infty}(1-H_{kk})^2\nu_k^2+\frac{C}{n}\Big]
\end{aligned}
$$

and by (43)

$$
\begin{aligned}
&N^{-2}\mathbf{E}\,\|\Psi_N^TH\Psi_NZ_N^T(Z_NZ_N^T)^{-1}Z_N\Psi_N^T\nu\|_N^2 \\
&\quad= N^{-2}\mathbf{E}\sum_{i=1}^{N}\Big\{\sum_{k=0}^{\infty}\sum_{s=0}^{\infty}H_{kk}\psi_k(X_i)\nu_s\sum_{l,m=1}^{N}[Z_N^T(Z_NZ_N^T)^{-1}Z_N]_{lm}\psi_k(X_l)\psi_s(X_m)\Big\}^2
\end{aligned}
$$

$$\leq \; CN^{-1}\mathbf{E}\Big\{\sum_{k=0}^{\infty}\sum_{s=0}^{\infty}H_{kk}|\nu_s|\sum_{l,m=1}^{N}[Z_N^T(Z_NZ_N^T)^{-1}Z_N]_{lm}\psi_k(X_l)\psi_s(X_m)\Big\}^2$$

$$\leq \; CN^{-1}\log n\Big(\sum_{s=0}^{\infty}|\nu_s|\Big)^2\mathbf{E}\;\mathrm{tr}^2 H + Cn^{-1}N^{-1} = o(1)\mathbf{E}\,\mathrm{tr}\,H.$$

Thus using the above inequalities and (61) we arrive at

$$\mathbf{E}\,\|Q_N^T(E - S_N^{-1}Q_NQ_N^T)\mu\|_N^2 = [1+o(1)]N\,\mathbf{E}\sum_{k=0}^{\infty}(1-H_{kk})^2\nu_k^2 + o(1)\mathbf{E}\sum_{k=0}^{\infty}H_{kk}^2. \qquad (62)$$

The interference term in (55) is evaluated by the similar way. So one gets

$$\mathbf{E}\left|\xi_N^T(E - Q_NS_N^{-1}Q_N^T)Q_N^T(E - S_N^{-1}Q_NQ_N^T)\mu\right| \leq o(1)\mathbf{E}\left[N\sum_{k=0}^{\infty}(1-H_{kk})^2\nu_k^2 + \sum_{k=0}^{\infty}H_{kk}^2\right].$$

The proof of (47) follows now from the above equation (55) and (54), (62). □

References

[1] Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In *Proceedings 2nd International Symposium on Information Theory,* P. N. Petrov and F. Csaki (Eds.) Akademia Kiado, Budapest, 267-281.

[2] Barron, A., Birgé, L., Massart, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory and related Fields* **113** 301–413.

[3] Bhattacharia P.K. and Zhao, P.-L. (1997). Semiparametric inference in a partial linear models. *Ann. Statist.* **25** 244–262.

[4] Chen, H. (1998). Convergence rates for parametric components in a partly linear model. *Ann. Statist.* **16** 136–146.

[5] Golubev, G., Härdle, W. (2001). On the second order minimax estimation in partial linear models. Submitted in *Math. Methods Statist.*

[6] Golubev G., Levit B. (1996). Distribution function estimation: adaptive smoothing. *Math. Methods Statist.* **5** 383–403.

[7] Golubev, G., Nussbaum, M. (1992). An adaptive spline estimates in nonparametric regression model. *Theory of Probab. Appl.* **3** 553–560. (In Russian)

[8] Green, P. J., Silverman, B. W. (1994). *Nonparametric regression and generalized linear models. A roughness penalty approach.* Chapman & Hall.

[9] Härdle, W., Liang, H., Gao, J. (1999). Partially linear models. Electronic version: *http://www.xplore-stat.de/ebooks.html.*

[10] Efroimovich, S.Y. and Pinsker, M.S. (1984). A learning algorithm for nonparametric filtering. *Automat. i Telemekh.* **11** 58–65. (In Russian).

[11] Heckman, N.E. (1986). Spline smoothing in a partly linear models. *Journal of the Royal Statistical Society, Series B,* **48** 244-248.

[12] Kneip, A. (1994). Ordered linear smoothers. *Ann. Statist.* 22, 835–866.

[13] Lepski, O. V. (1991). Asymptotically minimax adaptive estimation I: upper bounds. Optimal adaptive estimates. *Theory Probab. Appl.* **37** 682–697.

[14] Lepski, O. V. (1992). On problems of adaptive estimation in white Gaussian noise. In *Topics in Nonparametric Estimation. Advances in Soviet Math.*, Khasminskii R. Z. ed. Amer. Math. Soc., Providence, R. I. **12** 87–106.

[15] Lepski, O. V., Spokoiny, V. G. (1997). Optimal pointwise adaptive methods in nonparametric estimation. *Ann. Statist.* **25**

[16] Mallows, C. L. (1973). Some comments on C_p. *Technometrics*, **15**, 661–675.

[17] Mammen, E., Van de Geer, S. (1997). Penalized quasi-likelihood estimation in partial linear models. *Annals Statist.* **25**, 1014–1035.

[18] Nemirovskii, A. (1998). *Topics in Noparametric Statistics.* Ecole d'été de probabilités de St. Flour. Lecture Notes.

[19] Pinsker, M. S. (1980). Optimal filtering of square integrable signals in Gaussian in Gaussian noise. *Problems Inform. Transmission,* **16** 120-133.

[20] Robinson, P.M. (1987). Asymptotically efficient estimation in the presence of heteroskedasticity of unknown form. *Econometrica*, **55** 875-891.

[21] Rice, J.A. (1986). Convergence rates for partially splined models. *Statistics and Probability Letters,* **4** 203-208.

[22] Schimek, M. G. (2000). Estimation and inference in partially linear models with splines. *JSPI*, to appear.

[23] Speckman, P. (1988). Kernel smoothing in partial linear models. *Journal of Royal Statistical Society, Series B,* **50** 413-416.

[24] Tsybakov, A. B. (1998). Pointwise and sup-norm sharp adaptive estimation of functions on the Sobolev classes. *Ann. Statist.* **26** 2420–2469.

[25] Wahba, G. (1990). *Spline Models for Observational Data. S.I.A.M.*, Philadelphia.