On adaptive estimation in partial linear models

G. Golubev and W. Härdle [1]

Institute for Problems of Information Transmission
Bolshoi Karetny 19
101447 Moscow, Russia
e-mail: glbv@ippi.ac.msk.su

Humboldt-Universität zu Berlin
Spandauer Strasse 1
10178 Berlin, Germany
e-mail: haerdle@wiwi.hu-berlin.de

Abstract

The problem of estimation of the finite dimensional parameter in a partial linear model is considered. We derive upper and lower bounds for the second minimax order risk and show that the second order minimax estimator is a penalized maximum likelihood estimator. It is well known that the performance of the estimator is depending on the choice of a smoothing parameter. We propose a practically feasible adaptive procedure for the penalization choice.

1 Introduction

In the partial linear model we estimate an unknown parameter $\theta \in \mathbf{R}^d$ based on the observations

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

where ξ_i are i.i.d. random variables with zero mean and finite variance $\sigma^2 = \mathbf{E}\xi_i^2$. It is assumed that the regressors X_i are i.i.d. random variables taking values in $[0,1]$ and do not depend on ξ_i. The function $m(x),\ x \in [0,1]$ here is the unknown

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

1991 Mathematics Subject Classification. Primary 62G05; secondary 62G20.

Key words and phrases. Adaptive estimation, second order minimax risk, penalized likelihood.

non-parametric nuisance function, such that the random variables $m(X_i)$ have zero mean.

There are different statistical models for the predictors Z_i. Speckman (1986) assumed that the entries of Z are functionally connected with the regressors X_i. In the present paper we deal with a simpler case by assuming that the matrix Z does not depend on the regressors X_i and the noise ξ_i and that the empirical covariance matrix V

$$V_{kl} = \frac{1}{n}\sum_{i=1}^{n} Z_{ik}Z_{il}.$$

is nonsingular for all $n \geq n_0$.

Since the "noise" $m(X_i) + \xi_i$ has zero mean one could be tempted to use, for instance, the "naive" least squares estimator

$$\widehat{\theta}_n = \arg\min_{\theta \in R^d} \left\{ \frac{1}{n}\sum_{i=1}^{n} \left(Y_i - \theta^T Z_i\right)^2 \right\}$$

to estimate the unknown parameter θ. Evidently

$$\mathbf{E}(\widehat{\theta}_n - \theta)(\widehat{\theta}_n - \theta)^T = \frac{V^{-1}}{n}\{\sigma^2 + \mathbf{E}m^2(X_i)\},$$

where $(\cdot)^T$ denotes transposition. The estimator $\widehat{\theta}_n$ does not use a priori information about the nuisance function $m(\cdot)$. If this function is sufficiently smooth, then the performance of $\widehat{\theta}_n$ can be substantially improved. In Robinson (1987) it was shown that there exists an estimator $\widehat{\theta}_{ef}$ such that

$$\mathbf{E}(\widehat{\theta}_{ef} - \theta)(\widehat{\theta}_{ef} - \theta)^T = \{1 + o(1)\}\frac{\sigma^2 V^{-1}}{n}, \quad n \to \infty. \tag{2}$$

If ξ_i are Gaussian then the estimator $\widehat{\theta}_{ef}$ is often called asymptotically efficient or adaptive, Bickel, Klaassen, Ritov and Wellner (1992). Asymptotically efficient estimates of θ are traditionally constructed in partial linear models in two ways: by using kernel estimators as in Speckman (1988) or by penalization of the log-likelihood. For example, the penalized least squares estimator

$$\widehat{\theta} = \arg\min_{\theta \in R^d} \min_{m} \left\{ \frac{1}{n}\sum_{i=1}^{n} \left\{Y_i - \theta^T Z_i - m(X_j)\right\}^2 + \mu_n^2 \int_0^1 \left\{m^{(\beta)}(t)\right\}^2 dt \right\}, \tag{3}$$

where $m^{(\beta)}(\cdot)$ denotes the derivative of the order β and the smoothing parameter μ_n is of the order of $n^{-\beta/(2\beta+1)}$, is the efficient estimator in the case of Gaussian noise. More precisely for $\beta = 2$, it was proved in Rice (1986) that if

$$\int_0^1 \{m^{(\beta)}(t)\}^2 dt \leq L$$

then

$$\mathbf{E}(\widehat{\theta} - \theta)(\widehat{\theta} - \theta)^T = \frac{\sigma^2 V^{-1}}{n}\left\{1 + O(n^{-2\beta/(2\beta+1)})\right\}, \quad n \to \infty. \tag{4}$$

The goal of the present paper is to make precise the remainder term in (4) and thereby to provide an asymptotic minimax estimator for the parameter θ. The existing first order theory does not help us to perform this program. The available results in the second order theory Carroll and Härdle (1989), Chen (1988), Heckman (1986), Mammen and van de Geer (1997) Speckman (1988), Rice (1986) specify only the order of the second order term in the expansion of the risk. It is known that reasonable candidate estimators have a second order term of the order $n^{-(4\beta+1)/(2\beta+1)}$.

To shed some light on the optimal estimator in partial linear models we do the next natural step in investigation of the second order risk. We calculate it exactly up to the constant. Our considerations show, among other things, that the spline estimator in (3) is not the second order minimax. All asymptotically efficient estimators for partial linear models are based on some choice a smoothing parameter. Our next goal is therefore to propose a data based choice of the smoothing parameter, which gives the second order minimax estimator. This makes the adaptive estimator practically feasible.

The approach we use here is essentially based on the method proposed by Pinsker (1980). To simplify technical details we assume that the regressors X_i are i.i.d. random variables uniformly distributed on $[0,1]$ and $m(x)$ integrates to zero and belongs to a Sobolev ball. In other words $m \in W_2^\beta$, where

$$W_2^\beta = \left\{ m : \int_0^1 \{m^{(\beta)}(t)\}^2 dt \le L, \quad \int_0^1 m(t)dt = 0 \right\}.$$

It is assumed in the sequel that β is integer.

Our results can be extended in different directions. For instance the second order minimax estimator for the case of nonuniform X variables can be obtained. The error variables may be heteroskedastic i.e. $\mathbf{var}\,\xi_i$ may be a function of (X_i, Z_i) in particular of $\theta^T Z_i + m(X_i)$. This case is important in generalized partially linear models, where the variance is a function of the mean. Generalized linear models has been investigated by Severini and Staniswalis (1994) and recently applied to migration by Härdle, Mammen and Müller (1997). But we intentionally choose the simplest partial linear model to demonstrate why the second order theory is essential in semiparametric estimation. We will make comments on some possible extensions of our theory later in the text.

The outline of the paper is as follows. We first calculate a lower bound for the minimax risk. We then study in Section 3 an upper bound and turn finally in Section 4 to a practical method of adaptation.

2 A lower bound

In the derivation of a lower bound we assume only that the random variables ξ_i have a density $p(x)$, $x \in \mathbf{R}^1$ with finite Fisher information

$$I_\xi = \int_{-\infty}^{\infty} \frac{p'^2(x)}{p(x)} dx < \infty.$$

Our approach is based on the well-known idea of parametrization of the functional class W_2^β. We do this by constructing the following orthonormal system in $L_2[0,1]$, which approximates the ellipsoid W_2^β in the minimax sense

$$\{\psi_k\}_1^N = \arg\min_{\varphi_k} \sup_{m\in W_2^\beta} \min_{m_k} \int_0^1 \left\{ m(t) - \sum_{k=1}^N m_k \varphi_k(t) \right\}^2 \mathrm{d}t.$$

It is not difficult to see that $\{\psi_k(t)\}_1^\infty$ is the system with double orthogonality

$$\int_0^1 \psi_k(t)\psi_j(t)\,\mathrm{d}t = \delta_{kj}$$
$$\int_0^1 \psi_k^{(\beta)}(t)\psi_j^{(\beta)}(t)\,\mathrm{d}t = \lambda_k^{-1}\delta_{kj},$$

where

$$\lambda_k^{-1} = \min_{\varphi_s} \sup_{m\in W_2^\beta} \min_{m_s} \int_0^1 \left\{ m(t) - \sum_{s=1}^N m_s \varphi_s(t) \right\}^2 \mathrm{d}t.$$

Note that the first $\beta-1$ functions $\psi_k(t)$ are the standard orthonormal polynomial of the orders $1,\ldots,\beta-1$. Integration by parts easily reveals that λ_s and $\psi_s(t)$ satisfy the following boundary problem

$$\begin{aligned} \lambda_s\psi_s^{(2\beta)}(t) &= (-1)^\beta \psi_s(t), \\ \psi_s^{(k)}(1) &= \psi_s^{(k)}(0) = 0, \quad k = \beta,\ldots,2\beta-1. \end{aligned} \tag{5}$$

Moreover it is well-known that $\{1, \psi_k(t),\ k=1,\ldots\}$ is complete orthonormal system in $L_2(0,1)$. Thus any function $m(t)$ from W_2^β can be represented as

$$m(t) = \sum_{k=1}^\infty \nu_k\psi_k(t), \quad \nu_k = \int_0^1 m(t)\psi_k(t)dt, \tag{6}$$

where the Fourier coefficients ν_k are such that

$$\sum_{k=1}^\infty \nu_k^2\lambda_k^{-1} \le L, \quad \lambda_k^{-1} = 0,\ k=1,\ldots,\beta-1. \tag{7}$$

The asymptotic behavior of λ_k plays a very important role in approximation theory since they define Kolmogorov's diameter of W_2^β, Tikhomirov (1976). From (5) one can show by a simple algebra that uniformly in $s \ge \beta$

$$\lambda_s^{-1} = (\pi s)^{2\beta}\left\{1 + O(s^{-1})\right\}. \tag{8}$$

For more details we refer to Duistermaat (1995) or Härdle and Nussbaum (1994).

Let $B_r(\theta_0)$ be the ball in $\mathbf{R}^d$ of the radius $r > 0$ and with the center at θ_0. The following theorem provides a lower bound for the second order term in the minimax risk expansion.

Theorem 1 *For any estimator* $\widehat{\theta}$

$$\sup_{m\in W_2^\beta}\sup_{\theta\in B_r(\theta_0)}\mathbf{E}(\widehat{\theta}-\theta)(\widehat{\theta}-\theta)^T \geq \frac{V^{-1}}{nI_\xi}\left(1+\frac{1+o(1)}{n}\sum_{s=1}^{\infty}h_s\right), \tag{9}$$

where

$$h_s = \left[1-\mu\lambda_s^{-1/2}\right]_+ \tag{10}$$

and μ *is a root of the equation*

$$\frac{1}{nI_\xi}\sum_{s=1}^{\infty}\lambda_s^{-1}\left[\mu^{-1}\lambda_s^{1/2}-1\right]_+ = L. \tag{11}$$

Thus we see that the second order risk is controlled by the quantity

$$\Delta_n = \frac{1}{nI_\xi}\sum_{s=1}^{\infty}h_s.$$

The statistical interpretation of this value is well-known. The theorem due to Pinsker (1980) states that Δ_n is the asymptotic minimax risk in the following filtering problem. Suppose that we wish to estimate the infinite dimensional vector $(\nu_1,\nu_2,\ldots)^T$ based on observations

$$s_i = \nu_i + n^{-1/2}\varepsilon_i, \quad i=1,2,\ldots$$

where ε_i are i.i.d. $\mathcal{N}(0,I_\xi^{-1})$ and the parameters of interest ν_i obey condition (7). Then as $n\to\infty$

$$\inf_{\widehat{\nu}}\sup_{\nu}\sum_{k=0}^{\infty}\mathbf{E}(\widehat{\nu}_k-\nu_k)^2 = \{1+o(1)\}\Delta_n,$$

where the *inf* is taken over all possible estimators. The value of Δ_n can be calculated as follows. From (8) one concludes with μ solving (11) that

$$\Delta_n = \{1+o(1)\}n^{-1}C(\beta)(LnI_\xi)^{1/(2\beta+1)},$$

where $C(\beta)$ is the Pinsker constant

$$C(\beta) = \pi^{-2\beta/(2\beta+1)}(2\beta+1)^{1/(2\beta+1)}\left\{\beta/(\beta+1)\right\}^{2\beta/(2\beta+1)}.$$

Remark 1. If the regressors X_i have nonuniform density $p(x)$, $x\in[0,1]$ the corresponding basis $\{\psi_k\}_1^\infty$ is obtained as a solution of the following boundary problem

$$\begin{aligned}\lambda_s\psi_s^{(2\beta)}(t) &= (-1)^\beta p(x)\psi_s(t),\\ \psi_s^{(k)}(1) &= \psi_s^{(k)}(0)=0, \quad k=\beta,\ldots,2\beta-1.\end{aligned}$$

In this case the asymptotic behavior of λ_k is given by

$$\lambda_k^{-1} = \{1+o(1)\}(\pi k)^{2\beta}\left(\int_0^1 p^{1/2\beta}(x)\,\mathrm{d}x\right)^{-2\beta}, \quad k\to\infty.$$

For more details we refer to Utreras (1980) and Speckman (1985).

3 An upper bound

In this section we consider penalized least squares estimators. Recall the main heuristic idea of the penalized likelihood. Let the noise ξ_i be Gaussian. Assume that the Fourier coefficients ν_k in (6) are i.i.d. $\mathcal{N}(0, \sigma_k^2)$ and the parameters of interest θ_k are i.i.d. $\mathcal{N}(0, n^2)$. Then is it well known that the estimator

$$\widehat{\theta} = \arg \min_{\theta \in R^d} \min_{\nu_k} \left\{ \frac{1}{\sigma^2} \sum_{i=1}^{n} \left(Y_i - \theta^T Z_i - \sum_{\sigma_k^2 > 0} \nu_k \psi_k(X_i) \right)^2 + \sum_{\sigma_k^2 > 0} \frac{\nu_k^2}{\sigma_k^2} + \frac{\|\theta\|^2}{n^2} \right\} \quad (12)$$

is the Bayesian estimator. Although in the minimax setting the above assumptions are not fulfilled, nevertheless we use $\hat{\theta}$ in this situation. The problem is to calculate its minimax risk and to choose the regularization sequence σ_k^2 to minimize the risk. The following theorem shows how this approach works. Denote by W_n the number of strictly positive σ_k^2.

Theorem 2 *Let* $\mathbf{E}\xi_k = 0$, $\mathbf{E}\xi_k^2 = \sigma^2$, $\mathbf{E}\,|\xi_k|^{2(1+\delta)} < \infty$ *for some* $\delta > 0$, *and*

$$\lim_{n \to \infty} \frac{W_n^2 \log n}{n} = 0.$$

Then for any A, *as* $n \to \infty$, *uniformly in* $m \in W_2^\beta$

$$\sup_{\|\theta\| \le A} \mathbf{E}(\widehat{\theta} - \theta)(\widehat{\theta} - \theta)^T = \frac{V^{-1}\sigma^2}{n} + \frac{V^{-1}\{1 + o(1)\}}{n} \sum_{k=1}^{\infty} \left(\nu_k^2 (1 - h_k)^2 + \frac{\sigma^2}{n} h_k^2 \right), \quad (13)$$

where

$$h_k = \left(1 + \frac{\sigma^2}{n\sigma_k^2} \right)^{-1}, \quad \nu_k = \int_0^1 m(x)\psi_k(x)dx. \quad (14)$$

It follows from the above theorem and Theorem 1 that if $\sigma_k^2 = \sigma^2 h_k \{n(1 - h_k)\}^{-1}$, where h_k are defined by (10) and (11), then $\widehat{\theta}$ is the second order minimax estimator in the case of Gaussian noise. To verify this fact note that if $m \in W_2^\beta$, then from (7) we have

$$\sup_{m \in W_2^\beta} \sum_{k=1}^{n} \nu_k^2 (1 - h_k)^2 \le L \max_k \lambda_k (1 - h_k)^2 = L\mu^2$$

and by (11)

$$I_\xi L\mu^2 + \frac{1}{n} \sum_{k=1}^{n} h_k^2 = \frac{\mu}{n} \sum_{s=1}^{\infty} \lambda_s^{-1/2} \left[1 - \mu \lambda_s^{-1/2} \right]_+ + \frac{1}{n} \sum_{s=1}^{\infty} \left[1 - \mu \lambda_s^{-1/2} \right]_+^2 = \frac{1}{n} \sum_{s=1}^{\infty} h_s$$

thus proving the required result.

Remark 2. We added the additional term $\|\theta\|^2 / n^2$ into the definition of the penalized least squares estimator (12) only to simplify the proofs of Theorem 2 and Theorem 3 below.

Remark 3. In the case when the distribution of the noise is non Gaussian but known the penalized maximum likelihood estimator

$$\widehat{\theta}_p = \arg \max_{\theta \in R^d} \max_{\nu_k} \left\{ \sum_{i=1}^{n} \log p \left(Y_i - \theta^T Z_i - \sum_{\sigma_k^2 > 0} \nu_k \psi_k(X_i) \right) - \frac{1}{2} \sum_{\sigma_k^2 > 0} \frac{\nu_k^2}{\sigma_k^2} \right\}.$$

might be used. Under additional assumptions on the density $p(\cdot)$ one can show that the asymptotic behavior of the risk of this estimator is given by (13) with $\sigma^2 = I_\xi^{-1}$.

The optimal regularization sequence σ_k^2 strongly depends on the parameter L, which defines the functional class W_2^β. Since in practice this parameter is never known we can not make effective use of this estimator $\widehat{\theta}$. Therefore our next step is to construct a practically feasible second order efficient estimator which does not depend L.

4 An adaptive estimator

In this section we consider an adaptive version of the estimator (12). The goal of adaptation is to choose the regularization parameters σ_k^2 based on the observations. Theorem 2 plays an essential role in such a choice since it states that the second order risk is completely defined by

$$IMSE[h] = \sum_{k=1}^{\infty} \nu_k^2 (1 - h_k)^2 + \frac{\sigma^2}{n} \sum_{k=1}^{\infty} h_k^2$$

with $h = (h_1, h_2, \ldots)^T$ and ν_k from (14). In order to minimize the second order term in the risk expansion we have to minimize $IMSE[h]$ with respect to h. Unfortunately this functional depends on the Fourier coefficients ν_k, which we do not know. The main idea to overcome this difficulty is well-known and based on "cross-validation". It is commonly used in adaptive non-parametric estimation (see e.g. Akaike (1973), Mallows (1973), Efroimovich and Pinsker (1984), Golubev and Nussbaum (1992), Birge and Massart (1997)). Consider the functional

$$L[h] = \sum_{k=1}^{\infty} \nu_k^2 (h_k^2 - 2h_k) + \frac{\sigma^2}{n} \sum_{k=1}^{\infty} h_k^2, \tag{15}$$

which coincides with $IMSE[h]$ up to the term $\sum_{k=1}^{\infty} \nu_k^2$. We can estimate this functional by replacing the unknown ν_k^2 by unbiased estimators. Then we minimize the obtained risk predictor and find the optimal h_k or, equivalently, the optimal regularization σ_k^2.

The implementation of this general idea for partial linear models has specific features. In order to obtain an unbiased estimator for ν_k^2 we must use a subsample, namely the first T_n observations $Y_1, \ldots, Y_{T_n}$. The number $T_n \ll n$ will be specified

later on. Based on $Y_1, \ldots, Y_{T_n}$ we calculate the least squares estimates of ν_k as follows

$$\widehat{\nu}_k = \arg\min_{\nu_k} \min_{\theta} \left\{ \sum_{i=1}^{T_n} \left(Y_j - \theta^T Z_i - \sum_{k=1}^{N_n} \nu_k \psi_k(X_i) \right)^2 \right\}, \tag{16}$$

where the number N_n will be specified later on. The unbiased estimators for ν_k^2 are $\widehat{\nu}_k^2 - \sigma^2/T_n$ and the unbiased risk estimator for $L[h]$ is given by

$$L_n[h] = \sum_{k=1}^{N_n} \left(\widehat{\nu}_k^2 - \frac{\sigma^2}{T_n} \right) (h_k^2 - 2h_k) + \frac{\sigma^2}{n} \sum_{k=1}^{N_n} h_k^2. \tag{17}$$

Next the following adaptive regularization is used

$$\sigma_k^{*2} = \frac{\sigma^2 h_k^*}{n(1-h_k^*)}, \qquad \text{with} \quad h^* = \arg\min_{h \in \mathcal{H}_n} L_n[h], \tag{18}$$

where $\mathcal{H}_n$ is the set of admissible filters, see (10)

$$\mathcal{H}_n = \left\{ h_k : h_k = [1 - \mu \lambda_k^{-1/2}]_+, \ h_k = 0, \ k > N_n, \ \mu \in [0, \infty) \right\}. \tag{19}$$

Then we finally define the adaptive estimator as follows

$$\theta^* = \arg\min_{\theta \in R^d} \min_{\nu_k} \left\{ \frac{1}{\sigma^2} \sum_{i=1}^{n} \left(Y_i - \theta^T Z_i - \sum_{k=1}^{N_n} \nu_k \psi_k(X_i) \right)^2 + \sum_{k=1}^{N_n} \frac{\nu_k^2}{\sigma_k^{*2}} + \frac{\|\theta\|^2}{n^2} \right\}. \tag{20}$$

Theorem 3 *Let ξ_j be i.i.d. Gaussian $\mathcal{N}(0, \sigma^2)$ and uniformly in $k, l \in [1, d]$, as $n \to \infty$*

$$\sum_{i=1}^{n} Z_{ik} Z_{il} = \{1 + o(1)\} \sum_{j=T_n}^{n} Z_{ik} Z_{il}. \tag{21}$$

If for some $\delta > 0$

$$N_n = \sqrt{n} \log^{-2-\delta} n, \tag{22}$$

$$T_n = n \log^{-2-\delta} n, \tag{23}$$

then θ^* defined by (18)–(20) is the second order minimax estimator, that is

$$\sup_{m \in W_2^\beta} \sup_{\|\theta\| \le A} \mathbf{E}(\theta^* - \theta)(\theta^* - \theta)^T \le \frac{V^{-1}\sigma^2}{n} \left(1 + \frac{1 + o(1)}{n} \sum_{s=1}^{\infty} h_s \right),$$

where $h_s = \left[1 - \mu \lambda_s^{-1/2} \right]_+$ with μ defined as a root of the equation

$$\frac{\sigma^2}{n} \sum_{s=1}^{\infty} \lambda_s^{-1} \left[\mu^{-1} \lambda_s^{1/2} - 1 \right]_+ = L. \tag{24}$$

Remark 4. The main difficulty in the proving of this theorem lies in the fact that the empirical risk $n(\theta^* - \theta)(\theta^* - \theta)^T$ is not degenerate. More precisely, for any "good" estimator $\hat{\theta}$

$$\sqrt{n}(\hat{\theta} - \theta) = \sigma V^{-1/2}\xi_0 + \{1 + o(1)\}r_2^{1/2}(\hat{\theta}, m)V^{-1/2}\xi_1,$$

where ξ_0, ξ_1 are i.i.d. $\mathcal{N}(0, E)$ and $r_2(\hat{\theta}, m)$ is the empirical L_2-risk of the recovering the nuisance function $m(\cdot)$. We can predict the second order term $r_2(\hat{\theta}, m)$, since this random variable is degenerate. In order to make the dependence between the second order term and the first order weaker we used the subsample $Y_1, \ldots, Y_{T_n}$ with $T_n = n \log^{-1-\delta} n,\ \delta > 0$. This is of course the trick, to avoid technical difficulties. In practice the whole sample must be used for the risk prediction.

Remark 5. We assumed so far that the variance of the noise σ^2 is known. If this is not the case we may use the estimator

$$\sigma_n^2 = \min_{\nu_k} \min_{\theta} \left\{ \frac{1}{T_n} \sum_{i=1}^{T_n} \left(Y_j - \theta^T Z_i - \sum_{k=1}^{N_n} \nu_k \psi_k(X_i) \right)^2 \right\}$$

instead of σ^2 in the construction of our adaptive estimator.

Remark 6. We take into account only N_n nuisance parameters ν_k. The remaining $\nu_k,\ k = N_n, \ldots$ do not effect the second order risk since by (7) and (8)

$$\sum_{k=N_n+1}^{\infty} \nu_k^2 \le L(N_n + 1)^{-2\beta} \le Ln^{-1} \log^{2\beta(2+\delta)} n \ll n^{-2\beta/(2\beta+1)}.$$

5 Proofs of the theorems

5.1 Proof Theorem 1

We begin with a lower bound for the Bayesian risk. Assume that the nuisance function $m(\cdot)$ has the form

$$m(x) = \sum_{k=1}^{n} \nu_k \phi_k(x),$$

where $\phi_k(x)$ is a certain orthonormal system in $L_2(0, 1)$ such that

$$\int_0^1 \phi_k(x)dx = 0.$$

To induce a prior distribution on unknown parameters we assume that ν_k are independent $\mathcal{N}(0, \sigma_k^2)$. It is also assumed that $\theta_k \in (\theta_{0_k} - rd^{-1/2}, \theta_{0_k} + rd^{-1/2})$ are i.i.d. random variables with finite Fisher information I_θ. Let $R(\widehat{\theta}) = \mathbf{E}\left(\widehat{\theta} - \theta\right)\left(\widehat{\theta} - \theta\right)^T$ be the Bayesian risk.

Lemma 1 *If uniformly in* $k \geq 1$

$$\int_0^1 \phi_k^4(x)dx \leq C,$$

then for any estimator $\widehat{\theta}$

$$R(\widehat{\theta}) \geq \frac{V^{-1}}{nI_\xi}\left[1 + \frac{1}{n}\sum_{k=1}^{n} h_k \left\{1 - \frac{1}{n}\sum_{s=1}^{n} h_s - O\left(\frac{1}{n}\right)\right\}\right],$$

where $h_k = n\sigma_k^2 I_\xi/(1 + n\sigma_k^2 I_\xi)$.

Proof. Let $\mu = (\theta^T, \nu^T)^T$. Then from the Van Trees (1968) inequality it follows that for any estimator $\widehat{\mu}$

$$\mathbf{E}\{(\widehat{\mu} - \mu)(\widehat{\mu} - \mu)^T | X_1, \ldots, X_n\} \geq (I + I^\mu)^{-1}, \tag{25}$$

where I^μ is the Fisher information matrix of the prior distribution. This matrix is diagonal with entries

$$I^\mu_{kk} = \left\{ \begin{array}{ll} I_\theta, & k \leq d, \\ \sigma_k^{-2} & k > d. \end{array} \right.$$

The matrix I in (25) is the ordinary information matrix. In the considered model it is defined as

$$\begin{aligned} I_{kl} &= \mathbf{E}_X \frac{\partial}{\partial \mu_k} \sum_{j=1}^{n} \log p\left(Y_j - \theta^T Z_j - \sum_{k=1}^{n} \nu_k \phi_k(X_j)\right) \\ &\quad \times \frac{\partial}{\partial \mu_l} \sum_{j=1}^{n} \log p\left(Y_j - \theta^T Z_j - \sum_{k=1}^{n} \nu_k \phi_k(X_j)\right). \end{aligned} \tag{26}$$

It is easy to see from (26) that I admits the following representation

$$I = nI_\xi \left(\begin{array}{cc} V & \Phi Z \\ Z\Phi & \Phi\Phi \end{array} \right), \tag{27}$$

where the matrixes ΦZ, $Z\Phi$, $\Phi\Phi$ are

$$\Phi\Phi_{kl} = \frac{1}{n}\sum_{j=1}^{n} \phi_k(X_j)\phi_l(X_j), \quad \Phi Z_{kl} = \frac{1}{n}\sum_{j=1}^{n} \phi_k(X_j) Z_{jl}, \quad Z\Phi_{kl} = \frac{1}{n}\sum_{j=1}^{n} Z_{jk}\phi_l(X_j).$$

Let Σ be the diagonal matrix with entrees $\Sigma_{kk} = \sigma_k^{-2}$. Then from (27) one obtains

$$\begin{aligned} (I + I^\mu) &= nI_\xi \left(\begin{array}{cc} V & \Phi Z \\ Z\Phi & \Phi\Phi \end{array} \right) + \left(\begin{array}{cc} I_\theta E & 0 \\ 0 & \Sigma \end{array} \right) \\ &= nI_\xi \left(\begin{array}{cc} V_n & 0 \\ 0 & E + (nI_\xi)^{-1}\Sigma \end{array} \right) + \frac{nI_\xi}{\sqrt{n}} \left(\begin{array}{cc} 0 & \sqrt{n}\Phi Z \\ \sqrt{n} Z\Phi & \sqrt{n}(\Phi\Phi - E) \end{array} \right), \end{aligned} \tag{28}$$

where E is identity matrix and $V_n = V + (nI_\xi)^{-1} I_\theta E$. Denote

$$A = \begin{pmatrix} V_n & 0 \\ 0 & E + (nI_\xi)^{-1}\Sigma \end{pmatrix}, \quad B = \begin{pmatrix} 0 & \sqrt{n}\Phi Z \\ \sqrt{n} Z\Phi & \sqrt{n}(\Phi\Phi - E) \end{pmatrix}. \tag{29}$$

According to (25) we have to evaluate $(A + n^{-1/2}B)^{-1}$ from below. We get

$$\begin{aligned} (A + n^{-1/2}B)^{-1} &= (E + n^{-1/2}A^{-1}B)^{-1}A^{-1} \qquad (30) \\ &= (E - n^{-1/2}A^{-1}B + n^{-1}(A^{-1}B)^2 - n^{-3/2}(A^{-1}B)^3 + \ldots)A^{-1} \\ &\geq (E - n^{-1/2}A^{-1}B + n^{-1}(A^{-1}B)^2 - n^{-3/2}(A^{-1}B)^3)A^{-1}. \end{aligned}$$

Representing the matrix A^{-1} as

$$A^{-1} = \begin{pmatrix} V_n^{-1} & 0 \\ 0 & H \end{pmatrix},$$

where H is the diagonal matrix with entries $H_{kk} = n\sigma_k^2 I_\xi/(1 + n\sigma_k^2 I_\xi)$ and applying a simple algebra, we arrive at

$$\begin{aligned} A^{-1}B &= \begin{pmatrix} 0 & \sqrt{n}V_n^{-1}\cdot\Phi Z \\ \sqrt{n}H\cdot Z\Phi & \sqrt{n}H\cdot(\Phi\Phi - E) \end{pmatrix}, \qquad (31) \\ \left(A^{-1}B\right)^2 &= \begin{pmatrix} nV_n^{-1}\cdot\Phi X\cdot H\cdot Z\Phi & nV_n^{-1}\cdot\Phi Z\cdot H\cdot(\Phi\Phi - E) \\ nH\cdot(\Phi\Phi - E)\cdot H\cdot Z\Phi & * \end{pmatrix}, \\ \left(A^{-1}B\right)^3 &= \begin{pmatrix} n^{3/2}V_n^{-1}\cdot\Phi Z\cdot H\cdot(\Phi\Phi - E)\cdot H\cdot Z\Phi & * \\ * & * \end{pmatrix}. \end{aligned}$$

Here and later in the text $*$ denotes some matrix that is not needed in further calculations. Thus from the above equations and (25), (28), (30) we get

$$R(\widehat{\theta}) \geq \frac{V_n^{-1}}{nI_\xi}\left(E + V_n^{-1}\mathbf{E}\Phi Z\cdot H\cdot Z\Phi - V_n^{-1}\mathbf{E}\Phi Z\cdot H\cdot(\Phi\Phi - E)\cdot H\cdot Z\Phi\right) \tag{32}$$

Note also that

$$\begin{aligned} \mathbf{E}\Phi Z\cdot H\cdot Z\Phi_{kl} &= \mathbf{E}\frac{1}{n^2}\sum_{i,j=1}^n\sum_{m=1}^n Z_{jk}\phi_m(X_j)h_m\phi_m(X_i)Z_{il} \\ &= \mathbf{E}\frac{1}{n^2}\sum_{j=1}^n\sum_{m=1}^n Z_{jk}\phi_m^2(X_j)h_m Z_{jl} = V_{kl}\frac{1}{n}\sum_{m=1}^n h_m \end{aligned}$$

and

$$\begin{aligned} &\mathbf{E}(\Phi Z\cdot H\cdot(\Phi\Phi - E)\cdot H\cdot Z\Phi)_{kl} \\ &= \mathbf{E}\frac{1}{n^3}\sum_{p,i,j=1}^n\sum_{q,m=1}^n Z_{jk}\phi_m(X_j)h_m(\phi_m(X_p)\phi_q(X_p) - \delta(m,q))h_q\phi_q(X_i)Z_{il} \\ &= \mathbf{E}\frac{1}{n^3}\sum_{j=1}^n\sum_{q,m=1}^n Z_{jk}\phi_m(X_j)h_m(\phi_m(X_j)\phi_q(X_j) - \delta(m,q))h_q\phi_q(X_j)Z_{jl} \\ &= V_{kl}\frac{1}{n^2}\sum_{q,m=1}^n \mathbf{E}\phi_m(X_j)h_m(\phi_m(X_j)\phi_q(X_j) - \delta(m,q))h_q\phi_q(X_j) \\ &\leq V_{kl}\frac{1}{n^2}\left(\sum_{m=1}^n h_m\right)^2 + V_{kl}\frac{1}{n^2}\sum_{m=1}^n h_m^2. \end{aligned}$$

These equations together with (32) complete the proof of the lemma.

Proof of Theorem 1 follows now from Pinsker's theorem (1980) and Lemma 1.

5.2 Proof of Theorem 2

We start with some simple properties of the matrixes $\Psi\Psi$, ΨZ, $Z\Psi$.

Lemma 2 *Uniformly in* $x \le W_n^2 n^{1/2}$

$$\mathbf{P}\left\{\left\|\sqrt{n}Z\Psi\right\| > x\right\} \le W_n^2 \exp(-C(x/W_n)^2), \tag{33}$$

$$\mathbf{P}\left\{\left\|\sqrt{n}\Psi Z\right\| > x\right\} \le W_n^2 \exp(-C(x/W_n)^2), \tag{34}$$

$$\mathbf{P}\left\{\left\|\sqrt{n}(\Psi\Psi - E)\right\| > x\right\} \le W_n^2 \exp(-C(x/W_n)^2), \tag{35}$$

where $\|\cdot\|$ *is the ordinary matrix norm* $\|B\| = \max_{\|\mathbf{x}\|\le 1} \mathbf{x}^T B\mathbf{x}$.

Proof. By the Markov inequality one obtains

$$\begin{aligned}\mathbf{P}\left\{\left\|\sqrt{n}Z\Psi\right\| > x\right\} &\le \mathbf{P}\left\{\max_{k,l}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \psi_k(X_i)Z_{il}\right| > \frac{x}{W_n}\right\} \\ &\le 2\exp(-\lambda x/W_n)W_n^2 \max_{k,l}\mathbf{E}\exp\left(\frac{\lambda}{\sqrt{n}}\sum_{i=1}^n \psi_k(X_i)Z_{il}\right).\end{aligned} \tag{36}$$

Since X_i are i.i.d. random variables and $\mathbf{E}\psi_k(X_i) = 0$ we have by Taylor expansion

$$\mathbf{E}\exp\left(\frac{\lambda}{\sqrt{n}}\sum_{i=1}^n \psi_k(X_i)Z_{il}\right) \le \exp(C\lambda^2),$$

uniformly in $\lambda^2/n \le C$. Substituting the above inequality in (36) and minimizing with respect to λ, we arrive at (33). Inequalities (34)–(35) are proved by the same way.

Let $\mathcal{B}$ be a measurable set in $\mathbf{R}^n$. Denote for brevity by $\mathbf{E}_\mathcal{B}$ the conditional expectation

$$\mathbf{E}_\mathcal{B}\{\cdot\} = \mathbf{E}\{\cdot|(X_1,\dots,X_n)^T \in \mathcal{B}\}.$$

Lemma 3 *Let*

$$\widehat{\theta} = \arg\max_{\theta\in R^d}\max_{\nu_k}\left\{-\frac{1}{\sigma^2}\sum_{i=1}^n\left(Y_i - \theta^T Z_i - \sum_k \nu_k\psi_k(X_i)\right)^2 - \sum_k \frac{\nu_k^2}{\sigma_k^2} - \frac{\|\theta\|^2}{n^2}\right\}.$$

be an estimator of θ *with a penalization sequence* σ_k^2 *possibly depending on the observations* Y_i. *Assume that* $\mathbf{E}\,|\xi_k|^{2(1+\delta)} < \infty$ *for some* $\delta > 0$, *and* $\mathcal{B}$ *is such that for any* p

$$\mathbf{P}\{(X_1,\dots,X_n)^T \notin \mathcal{B}\} \le C(p)n^{-p},$$

where $C(p)$ *is some constant, which does not depend on* n, *then*

$$\mathbf{E}(\widehat{\theta}-\theta)(\widehat{\theta}-\theta)^T = \mathbf{E}_\mathcal{B}(\widehat{\theta}-\theta)(\widehat{\theta}-\theta)^T + O\left(n^{-2}\right).$$

Proof. Noting that $\left\|\widehat{\theta}\right\|^2 \le n^2 \sum_{i=1}^n Y_i^2$ and using Hölder's inequality, one obtains

$$\begin{aligned}
\mathbf{E}(\widehat{\theta}-\theta)(\widehat{\theta}-\theta)^T &= \mathbf{E}(\widehat{\theta}-\theta)(\widehat{\theta}-\theta)^T \mathbf{1}\left\{(X_1,\ldots,X_n)^T \in \mathcal{B}\right\} \\
&\quad + \mathbf{E}(\widehat{\theta}-\theta)(\widehat{\theta}-\theta)^T \mathbf{1}\left\{(X_1,\ldots,X_n)^T \notin \mathcal{B}\right\} \\
&\le \mathbf{P}\left\{(X_1,\ldots,X_n)^T \in \mathcal{B}\right\} \mathbf{E}_{\mathcal{B}}(\widehat{\theta}-\theta)(\widehat{\theta}-\theta)^T \\
&\quad + n^2 \left[\mathbf{P}\left\{(X_1,\ldots,X_n)^T \notin \mathcal{B}\right\}\right]^{\delta/(1+\delta)} \left\{\mathbf{E}\left(\sum_{i=1}^n Y_i^2\right)^{1+\delta}\right\}^{1/(1+\delta)}.
\end{aligned}$$

Thus we arrive at the assertion of the lemma.

Proof of Theorem 2. Let $\mathcal{B} = \left\{X_1,\ldots,X_n : \|B\| \le CW_n\sqrt{\log n}\right\}$, then by Lemma 2 for any p

$$\mathbf{P}\left\{(X_1,\ldots,X_n)^T \in \mathcal{B}\right\} = 1 + o(n^{-p}), \tag{37}$$

where C is some sufficiently large constant and the matrix B is defined by (29).

Let

$$\widehat{\mu} = \arg\max_{\mu=(\theta,\nu)} \left\{-\frac{1}{\sigma^2}\sum_{j=1}^n \left(Y_j - \theta^T Z_j - \sum_{\sigma_k^{*2}>0} \nu_k \psi_k(X_j)\right)^2 - \sum_{\sigma_k^2>0} \frac{\nu_k^2}{\sigma_k^2} - \frac{\|\theta\|^2}{n^2}\right\} \tag{38}$$

be the penalized mean square estimator of the parameter of interest θ and the nuisance parameter ν. Differentiating (38) one easily obtains that $\widehat{\mu}$ satisfies the following linear equations see (27)

$$\begin{pmatrix} V + n^{-2}E & \Psi Z \\ Z\Psi & \Psi\Psi + \Sigma \end{pmatrix} (\mu - \widehat{\mu}) = \begin{pmatrix} \xi Z \\ \xi\Psi \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & \Sigma \end{pmatrix} \mu, \tag{39}$$

where

$$\xi Z_l = \frac{1}{n}\sum_{i=1}^n \xi_i Z_{il}, \quad \xi\Psi_k = \frac{1}{n}\sum_{i=1}^n \xi_i \psi_k(X_i)$$

and Σ is the diagonal matrix with entries $\Sigma_{kk} = \sigma^2/(n\sigma_k^2)$. From (39) we see that

$$\begin{aligned}
&[\mu - \mathbf{E}\{\widehat{\mu}|X_1,\ldots,X_n\}][\mu - \mathbf{E}\{\widehat{\mu}|X_1,\ldots,X_n\}]^T \qquad (40)\\
&= \begin{pmatrix} V_n & \Psi Z \\ Z\Psi & \Psi\Psi + \Sigma \end{pmatrix}^{-1} \begin{pmatrix} 0 & 0 \\ 0 & \Sigma\nu\nu^T\Sigma \end{pmatrix} \begin{pmatrix} V + n^{-2}E & \Psi Z \\ Z\Psi & \Psi\Psi + \Sigma \end{pmatrix}^{-1},
\end{aligned}$$

where $V_n = V + n^{-2}E$. To evaluate the right-hand side of the above equation use the representation

$$\begin{pmatrix} V_n & \Psi Z \\ Z\Psi & \Psi\Psi + \Sigma \end{pmatrix} = A + \frac{1}{\sqrt{n}} B,$$

where

$$A = \begin{pmatrix} V_n & 0 \\ 0 & H^{-1} \end{pmatrix}, \quad B = \begin{pmatrix} 0 & \sqrt{n}\Psi Z \\ \sqrt{n}Z\Psi & \sqrt{n}(\Psi\Psi - E) \end{pmatrix},$$

here the diagonal matrix H has entries $H_{kk} = 1 + \sigma^2/(n\sigma_k^2)$. Thus by Taylor expansion and (40) we obtain with (30)

$$
\begin{aligned}
(\mu - \mathbf{E}_{\mathcal{B}}\widehat{\mu})(\mu - \mathbf{E}_{\mathcal{B}}\widehat{\mu})^T &= \begin{pmatrix} 0 & 0 \\ 0 & H\Sigma\nu\nu^T\Sigma H \end{pmatrix} \\
&\quad + \frac{1}{n}\mathbf{E}_{\mathcal{B}} A^{-1}B \begin{pmatrix} 0 & 0 \\ 0 & H\Sigma\nu\nu^T\Sigma H \end{pmatrix} A^{-1}B \\
&\quad + \frac{1}{n}\mathbf{E}_{\mathcal{B}} (A^{-1}B)^2 \begin{pmatrix} 0 & 0 \\ 0 & H\Sigma\nu\nu^T\Sigma H \end{pmatrix} + O\left(\frac{1}{n^{3/2}}\right).
\end{aligned} \tag{41}
$$

Therefore with (31)

$$
(\theta - \mathbf{E}_{\mathcal{B}}\widehat{\theta})(\theta - \mathbf{E}_{\mathcal{B}}\widehat{\theta})^T = V_n^{-1}\mathbf{E}_{\mathcal{B}} Z\Psi \cdot H\Sigma\nu\nu^T\Sigma H \cdot \Psi Z \cdot V_n^{-1} + O\left(\frac{1}{n^{3/2}}\right). \tag{42}
$$

Since $\psi_k(\cdot)$ is orthonormal system,

$$
\begin{aligned}
\mathbf{E}_{\mathcal{B}} Z\Psi \cdot H\Sigma\nu\nu^T\Sigma H \cdot \Psi Z \cdot V^{-1} &= V V_n^{-1} \frac{1}{n}\sum_{k=1}^{n} \nu_k^2 \Sigma_{kk}^2 H_{kk}^2 + O\left(\frac{1}{n^{3/2}}\right) \\
&= E\frac{1+o(1)}{n}\sum_{k=1}^{n} \nu_k^2 (1-h_k)^2 + O\left(\frac{1}{n^{3/2}}\right).
\end{aligned} \tag{43}
$$

Next consider the covariance matrix of the estimator $\widehat{\mu}$. From (39) we have

$$
\begin{aligned}
\mathbf{Cov}_{\mathcal{B}}\widehat{\mu} &= \frac{\sigma^2}{n}\mathbf{E}_{\mathcal{B}}\left(A + \frac{1}{\sqrt{n}}B\right)^{-1} \begin{pmatrix} V & \Psi Z \\ Z\Psi & \Psi\Psi \end{pmatrix} \left(A + \frac{1}{\sqrt{n}}B\right)^{-1} \\
&= \frac{\sigma^2}{n}\mathbf{E}_{\mathcal{B}}\left(A + \frac{1}{\sqrt{n}}B\right)^{-1} + O\left(\frac{\sigma^2}{n^3}\right) \\
&\quad - \frac{\sigma^2}{n}\mathbf{E}_{\mathcal{B}}\left(A + \frac{1}{\sqrt{n}}B\right)^{-1} \begin{pmatrix} 0 & 0 \\ 0 & \Sigma \end{pmatrix} \left(A + \frac{1}{\sqrt{n}}B\right)^{-1}.
\end{aligned} \tag{44}
$$

The first term in the right hand-side of the above equation was already evaluated in proving Theorem 1. It was proved that if $\mathrm{tr}H = o(n)$ then

$$
\mathbf{E}_{\mathcal{B}}\left(A + \frac{1}{\sqrt{n}}B\right)^{-1} = \begin{pmatrix} V^{-1}[1 + \{1+o(1)\}\mathrm{tr}H/n] & * \\ * & * \end{pmatrix}. \tag{45}
$$

The last term is easily evaluated by comparison with (40). Applying the same arguments we have from (41)–(43)

$$
\begin{aligned}
&\mathbf{E}_{\mathcal{B}}\left(A + \frac{1}{\sqrt{n}}B\right)^{-1} \begin{pmatrix} 0 & 0 \\ 0 & \Sigma \end{pmatrix} \left(A + \frac{1}{\sqrt{n}}B\right)^{-1} \\
&\quad = \frac{1+o(1)}{n} \begin{pmatrix} V^{-1}\sum_{k=1}^{n} H_{kk}^2 \Sigma_{kk} & * \\ * & * \end{pmatrix} + O\left(\frac{1}{n^{3/2}}\right).
\end{aligned}
$$

According to the definition of the matrixes H and Σ we have

$$\sum_{k=1}^{n} H_{kk}^2 \Sigma_{kk} = \sum_{k=1}^{n} h_k^2 \frac{\sigma^2}{n\sigma_k^2} = \sum_{k=1}^{n} h_k^2 (h_k^{-1} - 1).$$

From (44), (45) we see that

$$\mathbf{cov}\{\widehat{\theta}|\mathcal{B}\} = V^{-1}\frac{\sigma^2}{n}\left(1 + \frac{1+o(1)}{n}\sum_{k=1}^{n} h_k^2\right).$$

This together with (42), (43) gives

$$\mathbf{E}_{\mathcal{B}}(\widehat{\theta}-\theta)(\widehat{\theta}-\theta)^T = \frac{V^{-1}\sigma^2}{n} + \frac{V^{-1}\{1+o(1)\}}{n}\sum_{k=1}^{\infty}\left(\nu_k^2(1-h_k)^2 + \frac{\sigma^2}{n}h_k^2\right).$$

The above equation together with (37) and Lemma 3 proves the theorem.

5.3 Proof of Theorem 3

We start the proof with some auxiliary results.

Lemma 4 *Let ξ_k be i.i.d. $\mathcal{N}(0,1)$ and $h_k \in \mathcal{H}_n$ be a sequence possibly depending on ξ_k. Assume that ν_k is a sequence from $l_2(1,\infty)$, which does not depend on ξ_k. Then*

$$\mathbf{E}\left\{\sum_{k=1}^{\infty}\xi_k\nu_k(1-h_k)^2\right\}^2 \le \frac{C\,\|\nu\|^2}{n} + C\log n\,\mathbf{E}\sum_{k=1}^{\infty}\nu_k^2(1-h_k)^2. \tag{46}$$

Proof. Let $\mathcal{H}_n^{\varepsilon}$ be the minimal ε-net in $\mathcal{H}_n$. Choose $\varepsilon = (n\log n)^{-1}$. It is clear that the cardinality of $\mathcal{H}_n^{\varepsilon}$ is less than ε^{-2}. Let h_k^{ε} be a point in $\mathcal{H}_n^{\varepsilon}$ such that

$$\|h - h^{\varepsilon}\| \le \varepsilon. \tag{47}$$

Then we have

$$\begin{aligned}\mathbf{E}\left\{\sum_{k=1}^{\infty}\xi_k\nu_k(1-h_k)^2\right\}^2 \;\le\; & 2\,\mathbf{E}\left\{\sum_{k=1}^{\infty}\xi_k\nu_k(1-h_k^{\varepsilon})^2\right\}^2 \\ & + 2\mathbf{E}\left[\sum_{k=1}^{\infty}\xi_k\nu_k\left\{(1-h_k)^2 - (1-h_k^{\varepsilon})^2\right\}\right]^2.\end{aligned} \tag{48}$$

The last term in the right-hand side of the above equation can be estimated from above by the Cauchy-Schwartz inequality

$$\mathbf{E}\left[\sum_{k=1}^{\infty}\xi_k\nu_k\left\{(1-h_k)^2 - (1-h_k^{\varepsilon})^2\right\}\right]^2 \le 4\varepsilon^2\,\|\nu\|^2. \tag{49}$$

To estimate the first term in the right-hand side (48) denote for brevity

$$S(h) = \sum_{k=1}^{\infty}\xi_k\nu_k(1-h_k)\left\{\sum_{k=1}^{\infty}\nu_k^2(1-h_k)^2\right\}^{-1/2}.$$

Thus we have

$$\mathbf{E}\left\{\sum_{k=1}^{\infty}\xi_k\nu_k(1-h_k^\varepsilon)^2\right\}^2 \le \mathbf{E}\left\{\sum_{k=1}^{\infty}\nu_k^2(1-h_k^\varepsilon)^2\right\}\max_{h^*\in\mathcal{H}_n^\varepsilon}|S(h^*)|^2. \tag{50}$$

Next according to Markov's inequality one obtains that for arbitrary $\omega > 0$

$$\begin{aligned}\mathbf{P}\left\{\max_{h^*\in\mathcal{H}_n^\varepsilon}|S(h^*)|^2 > x\right\} &\le \exp(-\omega\sqrt{x})\mathbf{E}\left[\sum_{h^*\in\mathcal{H}_n^\varepsilon}\exp\{\omega S(h^*)\}+\exp\{-\omega S(h^*)\}\right]\\ &= \exp(-\omega\sqrt{x}+\omega^2/2)\mathrm{card}(\mathcal{H}_n^\varepsilon).\end{aligned}$$

Choosing $\omega = \sqrt{x}$ we arrive at

$$\mathbf{P}\left\{\max_{h^*\in\mathcal{H}_n^\varepsilon}|S(h^*)|^2 > x\right\} \le Cn^3\exp(-x/2). \tag{51}$$

Finally noting that

$$\left|\sum_{k=1}^{\infty}\nu_k^2(1-h_k^\varepsilon)^2-\sum_{k=1}^{\infty}\nu_k^2(1-h_k)^2\right| \le 2\,\|h-h^\varepsilon\|\,\|\nu\|^2$$

and using (48)–(51) one completes the proof of the lemma.

We will use in the sequel the following result analogous to Lemma 4.

Lemma 5 *Let ξ_k be i.i.d. $\mathcal{N}(0,1)$ and $h_k \in \mathcal{H}_n$ be some sequence possibly depending on ξ_k. Then*

$$\mathbf{E}\left\{\sum_{k=1}^{\infty}(\xi_k^2-1)(h_k^2-2h_k)\right\}^2 \le C + C\log^2 n\,\mathbf{E}\sum_{k=1}^{\infty}h_k^2. \tag{52}$$

Proof. We use here almost the same arguments as in proving Lemma 4. Choose the minimal ε-net $\mathcal{H}_n^\varepsilon$ in $\mathcal{H}_n$ with $\varepsilon = n^{-1/2}\log^{-2} n$. Denote by h_k^ε a point in $\mathcal{H}_n^\varepsilon$ such that $\|h-h^\varepsilon\| \le \varepsilon$. Then we can write

$$\begin{aligned}\mathbf{E}\left\{\sum_{k=1}^{\infty}(\xi_k^2-1)(h_k^2-2h_k)\right\}^2 &\le 2\mathbf{E}\left[\sum_{k=1}^{\infty}(\xi_k^2-1)\{(h_k^\varepsilon)^2-2h_k^\varepsilon\}\right]^2 \\ &\quad + 2\mathbf{E}\left[\sum_{k=1}^{n}(\xi_k^2-1)\left\{(1-h_k)^2-(1-h_k^\varepsilon)^2\right\}\right]^2.\end{aligned} \tag{53}$$

Applying the Cauchy-Schwartz inequality to estimate the last term in the right-hand side of the above equation we get

$$\mathbf{E}\left[\sum_{k=1}^{n}(\xi_k^2-1)\left\{(1-h_k)^2-(1-h_k^\varepsilon)^2\right\}\right]^2 \le C\varepsilon^2 n. \tag{54}$$

To estimate the first term in (53) denote

$$S(h) = \sum_{k=1}^{\infty}(\xi_k^2-1)\{(h_k^\varepsilon)^2-2h_k^\varepsilon\}\left[\sum_{k=1}^{\infty}\{(h_k^\varepsilon)^2-2h_k^\varepsilon\}^2\right]^{-1/2},$$

Then we have

$$\mathbf{E}\left[\sum_{k=1}^{\infty}(\xi_k^2-1)\{(h_k^\varepsilon)^2-2h_k^\varepsilon\}\right]^2 \le \mathbf{E}\sum_{k=1}^{\infty}\{(h_k^\varepsilon)^2-2h_k^\varepsilon\}^2 \max_{h^*\in\mathcal{H}_n^\varepsilon}|S(h^*)|^2 . \qquad (55)$$

Noting that $\mathbf{E}\exp(\pm S(h)/4) \le \exp(1/4)$ and using Markov's inequality, one obtains

$$\mathbf{P}\left\{\max_{h^*\in\mathcal{H}_n^\varepsilon}|S(h^*)|^2 > x\right\} \le \exp(-\sqrt{x}/4)$$
$$\times\,\mathbf{E}\left[\sum_{h^*\in\mathcal{H}_n^\varepsilon}\exp\{S(h^*)/4\}+\exp\{-S(h^*)/4\}\right] = 2\mathrm{card}(\mathcal{H}_n^\varepsilon)\exp(-\sqrt{x}/4).$$

Thus

$$\left|\sum_{k=1}^{n}\{(h_k^\varepsilon)^2-2h_k^\varepsilon\}^2-\sum_{k=1}^{n}(h_k^2-2h_k)^2\right| \le 2\,\|h-h^\varepsilon\|\,n^{1/2}$$

and using (53)–(55) one completes the proof.

The proofs of the next two lemmas are quite analogous to the proofs of Lemmas 4, 5 and therefore omitted.

Lemma 6 *Let H be the diagonal matrix with entries $H_{kk}=h_k$, where $h\in\mathcal{H}_n$. Then*

$$\mathbf{E}\left\|V^{-1/2}\Psi Z_T(E-H)\nu\right\|^4 \le \frac{C\log^4 n}{T_n^2}\mathbf{E}\left\{\sum_{k=1}^{\infty}(1-h_k)^2\nu_k^2\right\}^2 + O\left(\frac{1}{n^4}\right).$$

Lemma 7 *Let H be the matrix from Lemma 6*

$$\mathbf{E}\left\|V^{-1/2}\Psi Z_T H\right\|^4 + \mathbf{E}\left\|V^{-1/2}H\xi Z_T\right\|^4 \le \frac{C\log^4 n}{T_n^2}\mathbf{E}\left(\sum_{k=1}^{\infty}h_k^2\right)^2,$$
$$\mathbf{E}\left\|V^{-1/2}\Psi Z_T H^{1/2}\right\|^4 + \mathbf{E}\left\|V^{-1/2}H^{1/2}\xi Z_T\right\|^4 \le \frac{C\log^4 n}{T_n^2}\mathbf{E}\left(\sum_{k=1}^{\infty}h_k\right)^2.$$

Consider two matrices

$$B=\begin{pmatrix}0 & \sqrt{n}\Psi Z\\ \sqrt{n}Z\Psi & \sqrt{n}(\Psi\Psi-E)\end{pmatrix},\quad B_T=\begin{pmatrix}0 & \sqrt{T_n}\Psi Z_T\\ \sqrt{T_n}Z\Psi_T & \sqrt{T_n}(\Psi\Psi_T-E)\end{pmatrix}$$

where the index T indicates the use of only the first T_n observations. Define the set

$$\mathcal{B}=\left\{X_1,\dots,X_n : \|B\|\le CW_n\sqrt{\log n},\ \|B_T\|\le CW_n\sqrt{\log n}\right\}, \qquad (56)$$

where C is a sufficiently large constant. From Lemma 2 it follows that for any p

$$\mathbf{P}\left\{(X_1,\dots,X_n)^T\in\mathcal{B}\right\} = 1+o(n^{-p}). \qquad (57)$$

Our first step is to estimate $\mathbf{E}_{\mathcal{B}} L[h^*]$ (see (15)). From (16) we see that the least squares estimator $\hat{\nu}$ can be represented as

$$\begin{pmatrix} * \\ \nu - \widehat{\nu} \end{pmatrix} = \begin{pmatrix} V_T & \Psi Z_T \\ Z\Psi_T & \Psi\Psi_T \end{pmatrix}^{-1} \begin{pmatrix} \xi Z_T \\ \xi \Psi_T \end{pmatrix}.$$

Evidently the above equation can be rewritten in the following equivalent form

$$\begin{pmatrix} * \\ \nu - \widehat{\nu} \end{pmatrix} = \frac{\sigma}{\sqrt{T_n}} \begin{pmatrix} V_T & \Psi Z_T \\ Z\Psi_T & \Psi\Psi_T \end{pmatrix}^{-1/2} \begin{pmatrix} \zeta \\ \eta \end{pmatrix}, \tag{58}$$

where ζ, η are independent white Gaussian noises.

Lemma 8 *For any $h^0 \in \mathcal{H}_n$*

$$\mathbf{E}_{\mathcal{B}} L[h^*] \leq \{1 + o(1)\} L[h^0] + O\left(n^{-1} \log^{6+3\delta} n\right).$$

Proof. Define the random vector $\widehat{\nu}' \in \mathbf{R}^{N_n}$ as

$$\widehat{\nu}' = \nu + \sigma T_n^{-1/2} \eta. \tag{59}$$

From (58) and Taylor expansion we see that the Gaussian random variables $\widehat{\nu}'_k - \widehat{\nu}_k$ have variance of order T_n^{-2}. Hence for any p

$$\mathbf{P}_{\mathcal{B}} \left\{ \max_k |\widehat{\nu}'_k - \widehat{\nu}_k| > C\sigma T_n^{-1} \log^{1/2} n \right\} \leq C(p) n^{-p},$$

where C a sufficiently large constant depending on p. Using this inequality we get

$$\begin{aligned} & \mathbf{E}_{\mathcal{B}} \sum_{k=1}^{N_n} (h_k^{*2} - 2h_k^*) \{ (\widehat{\nu}'_k)^2 - (\widehat{\nu}_k)^2 \} \qquad (60) \\ & \leq \mathbf{E}_{\mathcal{B}} \sum_{k=1}^{N_n} (h_k^{*2} - 2h_k^*) \left(2\nu_k + \sigma T_n^{-1/2} \eta_k + \widehat{\nu}'_k - \widehat{\nu}_k \right) (\widehat{\nu}'_k - \widehat{\nu}_k) \\ & \leq \frac{C \log^{1/2} n}{T_n} + \frac{C \log n}{T_n^{3/2}} \mathbf{E}_{\mathcal{B}} \sum_{k=1}^{N_n} h_k^* + \frac{C \log n}{T_n^2} \mathbf{E}_{\mathcal{B}} \sum_{k=1}^{N_n} h_k^*. \end{aligned}$$

From definition of the set $\mathcal{H}_n$, see (19) it follows that

$$\sum_{k=1}^{N_n} h_k^* \leq C \sum_{k=1}^{N_n} h_k^{*2}, \tag{61}$$

and for any $h^0 \in \mathcal{H}_n$

$$L_n[h^*] \leq L_n[h^0]. \tag{62}$$

Thus from (59)–(61) and Lemmas 4, 5 we have

$$\begin{aligned}
\mathbf{E}_{\mathcal{B}} L_n[h^*] + \sum_{k=1}^{\infty} \nu_k^2 &\le \{1+o(1)\}\mathbf{E}_{\mathcal{B}} L[h^*] + \frac{\sigma^2}{T_n}\mathbf{E}_{\mathcal{B}} \sum_{k=1}^{N_n} (h_k^{*2} - 2h_k^*)\left(\eta_k^2 - 1\right) \\
&\quad + \frac{2\sigma}{\sqrt{T_n}}\mathbf{E}_{\mathcal{B}} \sum_{k=1}^{N_n} (h_k^{*2} - 2h_k^*)\eta_k\nu_k + O\left(T_n^{-1}\log^{1/2} n\right) \\
&= \{1+o(1)\}\mathbf{E}_{\mathcal{B}} L[h^*] + O\left\{T_n^{-1} n^{1/2}\log n\ (\mathbf{E}L[h^*])^{1/2}\right\} \\
&\quad + O\left(T_n^{-1}\log^{1/2} n\right) + O\left\{T_n^{-1/2}\log n\ (\mathbf{E}L[h^*])^{1/2}\right\} \\
&\le \{1+o(1)\}\mathbf{E}_{\mathcal{B}} L[h^*] + O\left\{(nT_n^{-2} + T_n^{-1})\log^{2+\delta} n\right\}.
\end{aligned}$$

This equation together with (62) yields the assertion of the lemma.

Proof of Theorem 3. Assume that the regressors $X_1, \ldots, X_n$ belong to the set $\mathcal{B}$, see (56). Then using Taylor expansion we have based on (31) and (22)

$$\begin{aligned}
\theta - \widehat{\theta} &= V^{-1}\cdot\xi Z - V^{-1}\cdot\Psi Z\cdot H^*\cdot\xi\Psi + V^{-1}\cdot\Psi Z\cdot(E - H^*)\nu \\
&\quad + \{1+o(1)\}V^{-1}\cdot\Psi Z\cdot H^*\cdot Z\Psi\cdot V^{-1}\cdot\xi Z \\
&\quad + \{1+o(1)\}V^{-1}\cdot\Psi Z\cdot H^*\cdot(\Psi\Psi - E)\cdot(E - H^*)\nu \\
&\quad + \{1+o(1)\}V^{-1}\cdot\Psi Z\cdot H^*\cdot(\Psi\Psi - E)\cdot H^*\cdot\xi\Psi.
\end{aligned} \tag{63}$$

Note that the matrix H^* depends only on observations $Y_i,\ i = 1, \ldots, T_n$. Therefore it is convenient to represent the matrixes $\Psi Z,\ \xi\Psi,\ \xi Z,\ \Psi\Psi - E$ in the right-hand side of (63) in the form

$$\Psi Z = \frac{T_n}{n}\Psi Z_T + \Psi Z', \quad \text{where} \quad \Psi Z'_{kl} = \frac{1}{n}\sum_{i=T_n+1}^{n} \psi_k(X_i) Z_{il}.$$

Note that

$$\begin{aligned}
\Psi Z\cdot H^*\cdot\xi\Psi &= \Psi Z'\cdot H^*\cdot\xi\Psi' + T_n n^{-1}\Psi Z'\cdot H^*\cdot\xi\Psi_T \\
&\quad + T_n n^{-1}\Psi Z_T\cdot H^*\cdot\xi\Psi' + T_n^2 n^{-2}\Psi Z_T\cdot H^*\cdot\xi\Psi_T
\end{aligned} \tag{64}$$

and by (23), (21)

$$\mathbf{E}_{\mathcal{B}}\Psi Z'\cdot H^*\cdot\xi\Psi'\cdot(\Psi Z'\cdot H^*\cdot\xi\Psi')^T = \{1+o(1)\}V\sigma^2 n^{-2}\,\mathbf{E}_{\mathcal{B}}\sum_{k=1}^{\infty} h_k^{*2}.$$

Using Lemma 7 one obtains

$$\begin{aligned}
\mathbf{E}_{\mathcal{B}}\Psi Z_T\cdot H^*\cdot\xi\Psi'\cdot(\Psi Z_T\cdot H^*\cdot\xi\Psi')^T &= V^2 O\left(n^{-1}T_n^1\log^2 n\,\mathbf{E}_{\mathcal{B}}\sum_{k=1}^{\infty} h_k^{*2}\right), \\
\mathbf{E}_{\mathcal{B}}\Psi Z'\cdot H^*\cdot\xi\Psi_T\cdot(\Psi Z'\cdot H^*\cdot\xi\Psi_T)^T &= V^2 O\left(n^{-1}T_n^{-1}\log^2 n\,\mathbf{E}_{\mathcal{B}}\sum_{k=1}^{\infty} h_k^{*2}\right), \\
\mathbf{E}_{\mathcal{B}}\Psi Z_T\cdot H^*\cdot\xi\Psi_T\cdot(\Psi Z_T\cdot H^*\cdot\xi\Psi_T)^T &= V^2 O\left\{T_n^{-2}\log^4 n\,\mathbf{E}_{\mathcal{B}}\left(\sum_{k=1}^{\infty} h_k^{*2}\right)^2\right\}.
\end{aligned}$$

Hence

$$\mathbf{E}_{\mathcal{B}}\Psi Z\cdot H^*\cdot\xi\Psi\cdot(\Psi Z\cdot H^*\cdot\xi\Psi)^T=\{1+o(1)\}V\sigma^2n^{-2}\,\mathbf{E}_{\mathcal{B}}\sum_{k=1}^{\infty}h_k^{*2}. \tag{65}$$

By the same arguments and Lemma 6

$$\mathbf{E}_{\mathcal{B}}\Psi Z\cdot(E-H^*)\nu\cdot(\Psi Z\cdot(E-H^*)\nu)^T=\{1+o(1)\}Vn^{-1}\,\mathbf{E}_{\mathcal{B}}\sum_{k=1}^{\infty}(1-h_k^*)^2\nu_k^2. \tag{66}$$

Now let us look at the three last terms in (63). It is not very difficult to see that they are sufficiently small. Indeed, from Lemma 7 we have

$$\begin{aligned}&\mathbf{E}_{\mathcal{B}}\Psi Z\cdot H^*\cdot Z\Psi\cdot V^{-1}\cdot\xi Z\cdot(\Psi Z\cdot H^*\cdot Z\Psi\cdot V^{-1}\cdot\xi Z)^T \\ &\quad=\ Vn^{-2}O\left\{n^{-1}\log^4 n\,\mathbf{E}_{\mathcal{B}}\left(\sum_{k=1}^{\infty}h_k^*\right)^2\right\}=Vo\left(n^{-1}\,\mathbf{E}_{\mathcal{B}}L[h^*]\right).\end{aligned} \tag{67}$$

From Lemmas 6, 7 we get

$$\begin{aligned}&\mathbf{E}_{\mathcal{B}}\Psi Z\cdot H^*\cdot(\Psi\Psi-E)(E-H^*)\nu\{\Psi Z\cdot H^*\cdot(\Psi\Psi-E)(E-H^*)\nu\}^T \\ &=\ VO\left(\frac{\log^5 n}{n}\,\mathbf{E}_{\mathcal{B}}\,\|\Psi Z\cdot H^*\|^4+\frac{n}{\log^5 n}\,\mathbf{E}_{\mathcal{B}}\,\|(\Psi\Psi-E)\cdot(E-H^*)\nu\|^4\right) \\ &=\ VO\left\{\frac{\log^9 n}{n^3}\,\mathbf{E}_{\mathcal{B}}\left(\sum_{k=1}^{\infty}h_k^{*2}\right)^2+\frac{1}{n\log n}\,\mathbf{E}_{\mathcal{B}}\left(\sum_{k=1}^{\infty}(1-h_k)^2\nu_k^{*2}\right)^2\right\} \\ &=Vo\left(n^{-1}\,\mathbf{E}_{\mathcal{B}}L[h^*]\right),\end{aligned} \tag{68}$$

and

$$\begin{aligned}&\mathbf{E}_{\mathcal{B}}\Psi Z\cdot H^*\cdot(\Psi\Psi-E)\cdot H^*\cdot\xi\Psi\cdot(\Psi Z\cdot H^*\cdot(\Psi\Psi-E)\cdot H^*\cdot\xi\Psi)^T \\ &\quad=\ VO\left(n^{-1}\,\mathbf{E}_{\mathcal{B}}\,\|\Psi Z\cdot H^*\|^4+n^{-1}\,\mathbf{E}_{\mathcal{B}}\,\|H^*\cdot\xi\Psi\|^4\right) \\ &\quad\le\ VO\left\{n^{-4}\log^4 n\,\mathbf{E}_{\mathcal{B}}\left(\sum_{k=1}^{\infty}h_k^{*2}\right)^2\right\}=Vo\left(n^{-1}\mathbf{E}_{\mathcal{B}}L[h^*]\right).\end{aligned} \tag{69}$$

Consider now the interference terms. We begin with $\mathbf{E}_{\mathcal{B}}\Psi Z\cdot H^*\cdot\xi\Psi(V^{-1}\cdot\xi Z)^T$. From (64), Cauchy-Schwartz's inequality and Lemma 7 we have

$$\begin{aligned}&\mathbf{E}_{\mathcal{B}}\Psi Z\cdot H^*\cdot\xi\Psi\cdot(V^{-1}\cdot\xi Z)^T=\mathbf{E}_{\mathcal{B}}\Psi Z'\cdot H^*\cdot\xi\Psi'\cdot(V^{-1}\cdot\xi Z')^T \\ &\quad+\ T_nn^{-1}\,\mathbf{E}_{\mathcal{B}}\Psi Z_T\cdot H^*\cdot\xi\Psi_T\cdot V^{-1}\cdot\xi Z_T^T=\{1+o(1)\}En^{-2}\sigma^2\mathbf{E}_{\mathcal{B}}\sum_{k=1}^{\infty}h_k^*.\end{aligned} \tag{70}$$

The next interference term is also easy to estimate from above using Lemma 6 and the Cauchy-Schwartz inequality

$$\left\|\mathbf{E}_{\mathcal{B}}\Psi Z\cdot(E-H^*)\nu\cdot(V^{-1}\xi Z)^T\right\| \tag{71}$$

$$
\begin{aligned}
&= T_n^2 n^{-2} \left\| \mathbf{E}_{\mathcal{B}} \Psi Z_T \cdot (E - H^*)\nu \cdot (V^{-1}\xi Z_T)^T \right\| \\
&\le C T_n^2 \left(n^{-5} \mathbf{E}_{\mathcal{B}} \left\| \Psi Z_T \cdot (E - H^*)\nu \right\|^2 \right)^{1/2} \\
&\le C \log n \left(T_n^3 n^{-5} \mathbf{E}_{\mathcal{B}} \sum_{k=1}^{\infty} (1 - h_k^*)^2 \nu_k^2 \right)^{1/2} \\
&\le o(n^{-1}) \, \mathbf{E}_{\mathcal{B}} \sum_{k=1}^{\infty} (1 - h_k^*)^2 \nu_k^2 + O(n^{-2}) = o\left(n^{-1} \, \mathbf{E}_{\mathcal{B}} L[h^*] \right).
\end{aligned}
$$

It remains to consider one more interference term, which has in fact the main order. Namely from Lemma 7 one obtains

$$
\begin{aligned}
&\mathbf{E}_{\mathcal{B}} \Psi Z \cdot H^* \cdot Z\Psi \cdot V^{-1}\xi Z \cdot (V^{-1}\xi Z)^T \qquad (72) \\
&\quad = \mathbf{E}_{\mathcal{B}} \Psi Z' \cdot H^* \cdot Z\Psi' \cdot V^{-1}\xi Z' \cdot (V^{-1}\xi Z')^T \\
&\qquad + T_n^2 n^{-2} \, \mathbf{E}_{\mathcal{B}} \Psi Z_T \cdot H^* \cdot Z\Psi_T \cdot V^{-1}\xi Z' \cdot (V^{-1}\xi Z')^T \\
&\qquad + T_n^2 n^{-2} \, \mathbf{E}_{\mathcal{B}} \Psi Z' \cdot H^* \cdot Z\Psi' \cdot V^{-1}\xi Z_T \cdot (V^{-1}\xi Z_T)^T \\
&\qquad + T_n^4 n^{-4} \, \mathbf{E}_{\mathcal{B}} \Psi Z_T \cdot H^* \cdot Z\Psi_T \cdot V^{-1}\xi Z_T \cdot (V^{-1}\xi Z_T)^T \\
&\quad = E\left\{ 1 + o(1) + O\left(n^{-1} T_n \log^2 n \right) \right\} \sigma^2 n^{-2} \, \mathbf{E}_{\mathcal{B}} \sum_{k=1}^{\infty} h_k^*.
\end{aligned}
$$

The remaining interference terms are of the order $o(\mathbf{E}_{\mathcal{B}} L[h^*])$. This follows easily from the Cauchy-Schwartz inequality and (65)–(69). Therefore by noting that

$$
\mathbf{E}_{\mathcal{B}} \xi Z \cdot \xi Z^T = \{1 + O(n^{-1})\} \frac{V\sigma^2}{n}
$$

we have from (63) and (65)–(72)

$$
\mathbf{E}_{\mathcal{B}} (\theta - \widehat{\theta})(\theta - \widehat{\theta})^T \le V^{-1} \left\{ \frac{\sigma^2}{n} + \frac{1 + o(1)}{n^2} \mathbf{E}_{\mathcal{B}} L[h^*] + O\left(\frac{1}{n^2} \right) \right\}.
$$

This inequality together with Lemma 8 yields that for any $h^0 \in \mathcal{H}_n$

$$
\mathbf{E}_{\mathcal{B}} (\theta - \widehat{\theta})(\theta - \widehat{\theta})^T \le V^{-1} \left\{ \frac{\sigma^2}{n} + \frac{1 + o(1)}{n^2} L[h^0] + O\left(\frac{\log^{6+3\delta} n}{n^2} \right) \right\}.
$$

Choosing $h_k^0 = \left[1 - \mu \lambda_s^{-1/2} \right]_+$ with μ defined in (24) applying Lemma 3 and (57) completes the proof of the theorem.

References

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

[2] Bickel, P.J., Klaassen, C.A.J., Rytov, Y. and Wellner, J.A. (1992). *Efficient and Adaptive Estimation in Semiparametric Models.* John Hopkins Univ. Press.

[3] Birgé, L. Massart, P. (1997). From model selection to adaptive estimation. In *Festschrift for Lucien Le Cam.* Pollard, D., Torgesen, E., Yang, G. eds. Springer, 55–87.

[4] Carroll, R.J., Härdle, W. (1989). A note on second-order effects in a semiparametric context. *Statistics.* **20** 179–186.

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

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

[7] Duistermaat J. (1995). The Sturm-Liouville problem for the operator $(-\mathrm{d}^2/\mathrm{d}\,x^2)^m$, with Neumann or Dirichlet boundary conditions. Technical report 899, Department of Mathematics, University Untrecht.

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

[9] Härdle, W. and Nussbaum, M. (1994). Kernel estimation: the spline-smoothing method. *Pub. Inst. Stat. Univ. Paris* XXXVIII, fasc. 3 61–86.

[10] Härdle, W., Mammen, E. and Müller, M. (1996). Testing parametric versus semiparametric in generalized linear models. Preprint 28 Sonderforschungsbereich 373. Humboldt-Universität zu Berlin. Available via http://sfb.wiwi.huberlin.de.

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

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

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

[14] Pinsker, M.S. (1980). Optimal filtering of square integrable signals in Gaussian white noise. *Problems Inform. Transm.* **16** 120–133.

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

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

[17] Severini, T.A. and Staniswalis, J.G. (1994). Quasi-likelihood estimation in semiparametric models. *J. Amer. Statist. Assoc.* **89** 501–511.

[18] Speckman, P. (1985). Spline smoothing and optimal rates of convergence in nonparametric regression models. *Ann. Statist.* **13** 970–983.

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

[20] Tikhomirov, V.M. (1976). *Some questions of approximation theory.* Moscow State University Press. (in Russian).

[21] Utreras, F. (1980) Sur le choix des parametre d'adjustements dans le lissage par functions spline. *Numer. Math.* **34** 15–28.

[22] Van Trees H.L. (1968). *Detection, Estimation and Modulation Theory.* Vol. 1, Wiley, New York.