The Annals of Statistics
2001, Vol. 29, No. 6, 1601–1619

ADAPTIVE PREDICTION AND ESTIMATION IN LINEAR REGRESSION WITH INFINITELY MANY PARAMETERS[1]

BY A. GOLDENSHLUGER AND A. TSYBAKOV

University of Haifa and Université Paris VI

The problem of adaptive prediction and estimation in the stochastic linear regression model with infinitely many parameters is considered. We suggest a prediction method that is sharp asymptotically minimax adaptive over ellipsoids in ℓ_2. The method consists in an application of blockwise Stein's rule with "weakly" geometrically increasing blocks to the penalized least squares fits of the first N coefficients. To prove the results we develop oracle inequalities for a sequence model with correlated data.

1. Introduction. Consider the regression model

$$y = \sum_{k=1}^{\infty} \theta_k x_k + \varepsilon, \tag{1}$$

where $\{x_k\}_{k=1,2,\ldots}$ is a sequence of explanatory variables, y is the corresponding response, ε is the error, and $\theta = (\theta_1, \theta_2, \ldots) \in \ell_2$ is an unknown regression sequence. Assume that $\{x_k\}$ and ε are random variables, and $\mathbb{E}\varepsilon = 0$ and $\mathbb{E}\varepsilon^2 = \sigma^2$. The stochastic series in (1) and later are assumed to converge in the mean squared sense. Suppose we are given n independent realizations of y and $\{x_k\}$,

$$\mathcal{U}_n = \{y(t); x_1(t), x_2(t), \ldots; t = 1, \ldots, n\}$$

coming from the model (1), that is,

$$y(t) = \sum_{k=1}^{\infty} \theta_k x_k(t) + \varepsilon(t), \qquad t = 1, \ldots, n.$$

Given $\mathcal{X}_{n+1} = \{x_1(n+1), x_2(n+1), \ldots\}$, the objective is to predict the corresponding response $y(n+1)$. A *predictor* (or *prediction method*) is a random variable $\hat{y} = \hat{y}(n+1)$ measurable with respect to $(\mathcal{U}_n, \mathcal{X}_{n+1})$.

The problem of prediction in the model (1) has been considered by Shibata (1981), Breiman and Freedman (1983), Goldenshluger and Tsybakov (1999). In particular, Shibata (1981) and Breiman and Freedman (1983) study the least squares predictor of the form

$$\hat{y}(n+1) = \sum_{k=1}^{d} \hat{\theta}_k^{\mathrm{OLS}} x_k(n+1), \tag{2}$$

Received June 2000; revised June 2001.

[1]Supported in part by an Arc-en-ciel/Keshet grant.

AMS 2000 *subject classifications*. 62G05, 62G20.

Key words and phrases. Linear regression with infinitely many parameters, adaptive prediction, exact asymptotics of minimax risk, blockwise Stein's rule, oracle inequalities.

where $(\hat{\theta}_1^{\mathrm{OLS}}, \ldots, \hat{\theta}_d^{\mathrm{OLS}})$ is the ordinary least squares (OLS) estimator of $(\theta_1, \ldots, \theta_d)$ based on the reduced data $\{y(t), x_1(t), \ldots, x_d(t),\ t = 1, \ldots, n\}$. They discuss data-driven choices of d. Goldenshluger and Tsybakov (1999) suggest the predictor

$$\hat{y}(n+1) = \sum_{k=1}^{d} \lambda_k \hat{\theta}_k^{\mathrm{P}} x_k(n+1), \tag{3}$$

where $(\hat{\theta}_1^{\mathrm{P}}, \ldots, \hat{\theta}_d^{\mathrm{P}})$ is a penalized least squares estimator of $(\theta_1, \ldots, \theta_d)$, and $\{\lambda_k\}$ are some weights. They show that the predictor (3) is asymptotically sharp minimax on the classes of ellipsoids in the space of coefficients θ, provided $\{\lambda_k\}$ are chosen in a proper way. In particular, the predictor (3) outperforms the OLS predictor (2) in the minimax sense. The weights $\{\lambda_k\}$ depend on the parameters of the ellipsoid, and the method (3) is not adaptive to these parameters.

In this paper we suggest an adaptive prediction method which is asymptotically as good as (3) (i.e., is asymptotically sharp minimax) on any ellipsoid within a wide scale. The method does not depend on the parameters of an ellipsoid and is not related to prior assumptions on an ellipsoidal structure. The idea is to apply the blockwise Stein rule to penalized least squares fits $\hat{\theta}_k^{\mathrm{P}}$ of the first N coefficients θ_k. The blockwise Stein rule has been used recently to get adaptive estimators in different statistical problems [Donoho and Johnstone (1995), Johnstone (1998), Cai (1999), Cavalier and Tsybakov (2000)]. In these papers asymptotically minimax adaptivity is proved by means of oracle inequalities in sequence space. Our reasoning is similar, but the difficulty of the present setting is that the sequence space representation is non-Gaussian, correlated and biased. We get oracle inequalities that work in this situation. We also prove that, under general conditions, blockwise linear estimators are almost as good as linear monotone oracles. These results are of independent interest, and can be used in other contexts as well. The sharp minimax adaptivity is proved as a consequence of these results for a special construction of "weakly" geometrically increasing blocks. This differs from the polynomially increasing blocks as in Efromovich and Pinsker (1984, 1996), Efromovich (1999), or dyadic blocks as in the wavelet context [Donoho and Johnstone (1995), Johnstone (1998, 1999)], but is closely related to Nemirovski (2000) and Cavalier and Tsybakov (2000), where other statistical models have been studied.

The paper is organized as follows. In Section 2 we define our adaptive prediction method, analyze its properties and state our main results. Section 3 considers an equivalent sequence space model and includes basic oracle inequalities underlying our proofs. In Section 4 we prove the main results. The Appendix contains auxiliary lemmas related to properties of the Stein estimator for correlated data.

2. Adaptive prediction method. We define the predictor $\hat{y}^*(n+1)$ of $y(n+1)$ as follows. Fix a positive integer $N = N_n$, and denote $\mathscr{I} = \{1, \ldots, N_n\}$.

Let $\phi_N(t) = (x_1(t), \dots, x_N(t))'$, $t = 1, \dots, n$, and consider a penalized least squares estimator

$$\tilde{y}_{\mathscr{I}} = \hat{\theta}^{\mathrm{P}}_{\mathscr{I}} = Q^{-1}_{\mathscr{I}}\left(\frac{1}{n}\sum_{t=1}^{n}\phi_N(t)y(t)\right), \tag{4}$$

where

$$Q_{\mathscr{I}} = \frac{1}{n}\sum_{t=1}^{n}\phi_N(t)\phi'_N(t) + n^{-1}I$$

and I is the $N_n \times N_n$ identity matrix. Here and later v_B denotes the vector $\{v_k, k \in B\}$, where B is a set of integers.

For a monotone increasing sequence of integers $\{\kappa_j\}$ such that $\kappa_1 = 1$, we define the partition of the set $\{1, \dots, N\}$ into blocks B_j as follows

$$B_j = \{\kappa_j, \kappa_j + 1, \dots, \kappa_{j+1} - 1\}, \qquad j = 1, \dots, J, \tag{5}$$

$$\kappa_{J+1} - 1 = N. \tag{6}$$

[We assume w.l.o.g. that the sequence $\{\kappa_j\}$ and N are such that (6) holds.] Denote $n_j = \kappa_j - \kappa_{j-1}$, $\theta_{(j)} = \theta_{B_j}$, and $\tilde{y}_{(j)} = \tilde{y}_{B_j}$, $j = 1, \dots, J$.

Let

$$\hat{\theta}^* = (\hat{\theta}^*_{(1)}, \hat{\theta}^*_{(2)}, \dots, \hat{\theta}^*_{(J)}, 0, 0, \dots), \tag{7}$$

where $\hat{\theta}^*_{(j)}$ is the Stein estimator for the block B_j,

$$\hat{\theta}^*_{(j)} = \left(1 - \frac{\sigma^2 n_j}{n\|\tilde{y}_{(j)}\|^2}\right)\tilde{y}_{(j)}, \qquad j = 1, \dots, J. \tag{8}$$

Here and later $\|\cdot\|$ is the Euclidean norm when applied to a finite dimensional vector.

Define the prediction method

$$\hat{y}^*(n+1) = \sum_{k=1}^{N_n}\hat{\theta}^*_k x_k(n+1). \tag{9}$$

We show that this method is asymptotically sharp adaptive in a minimax sense on the ellipsoids in the space of sequences θ. Consider the ellipsoids

$$\Theta(a, L) = \left\{\theta \in \ell_2 : \sum_{k=1}^{\infty} a_k^2\theta_k^2 \le L^2\right\},$$

where $L > 0$ and $a = \{a_k\}$ is a monotone nondecreasing positive sequence such that $a_k \to \infty$ as $k \to \infty$.

The prediction error of any predictor $\hat{y}$ is defined as $\mathbb{E}[\hat{y}(n+1) - y(n+1)]^2$. Note that this error cannot be arbitrarily small; it is at least σ^2 for large n, because of the nonvanishing innovation component $\varepsilon(n+1)$ independent of

$(\mathscr{U}_n, \mathscr{X}_{n+1})$. We therefore consider the difference $\mathbb{E}[\hat{y}(n+1) - y(n+1)]^2 - \sigma^2$, and define the maximal risk over $\Theta(a, L)$ in the form

$$\mathscr{R}[\hat{y}; \Theta(a, L)] = \sup_{\theta \in \Theta(a,L)} \mathbb{E}[\hat{y}(n+1) - y(n+1)]^2 - \sigma^2.$$

We show that $\hat{y}^* = \hat{y}^*(n+1)$ is an asymptotically minimax predictor; that is, it minimizes the maximal prediction error,

$$\mathscr{R}[\hat{y}^*; \Theta(a, L)] = \inf_{\hat{y}} \mathscr{R}[\hat{y}; \Theta(a, L)](1 + o(1)), \qquad n \to \infty,$$

where inf is taken over all possible prediction methods based on the observations $(\mathscr{U}_n, \mathscr{X}_{n+1})$.

The following assumptions will be used.

ASSUMPTION 1. The random variables $\{x_k\}_{k=1,2,\ldots}$ are independent, $\mathbb{E}x_k = 0$, $\mathbb{E}x_k^2 = 1$, and there exist constants $H > 0$ and $c_* > 0$ such that

$$\mathbb{E} \exp\{\lambda x_k^2\} \le c_* < \infty, \qquad |\lambda| < H, \quad k = 1, 2, \ldots . \tag{10}$$

The assumption that $\{x_k\}$ are uncorrelated zero mean random variables with variance 1 is quite natural in the prediction context, since typically $\{x_k\}$ are considered as "principal components" of some original random covariates [see the discussion in Breiman and Freedman (1983) and Goldenshluger and Tsybakov (1999)]. In particular, Breiman and Freedman (1983) work with i.i.d. standard Gaussian $\{x_k\}$.

Note also that under Assumption 1 there is a simple and natural relationship between prediction risk and the ℓ_2-risk in estimation of the regression coefficients. In this case the prediction risk $\mathbb{E}[\hat{y}(n+1) - y(n+1)]^2 - \sigma^2$ of any predictor of the type $\hat{y}(n+1) = \sum_{k=1}^{\infty} \hat{\theta}_k x_k(n+1)$ with $\hat{\theta}_k$'s based on $\mathscr{U}_n$ only, coincides with $\mathbb{E}\|\hat{\theta} - \theta\|^2$. Thus, the prediction problem is equivalent to estimating coefficients of the corresponding regression model with explanatory variables satisfying Assumption 1.

ASSUMPTION 2. The random variable ε is Gaussian $\mathscr{N}(0, \sigma^2)$, and ε is independent of $\{x_k\}_{k=1,2,\ldots}$.

ASSUMPTION 3. The partition (5) and (6) satisfies

$$\kappa_1 = 1, \ \kappa_2 = \nu_n + 1, \ \kappa_j = \kappa_{j-1} + \lfloor (1 + \rho_n)^{j-1+\nu_n} \rfloor, \ j = 3, \ldots, J,$$

where $\rho_n = (\ln \ln n)^{-1}$, $\nu_n = \lfloor \rho_n^{-1} \ln \rho_n^{-1} \rfloor$, and J is such that $N_n \asymp \sqrt{n}(\ln n)^{-1}$.

Let c_n denote the solution to the equation [cf. Pinsker (1980)]

$$\sigma^2 n^{-1} \sum_{k=1}^{\infty} a_k (1 - c_n a_k)_+ = c_n L^2$$

(note that the solution is unique since $a_k \to \infty$), and let

$$r_n = r_n(\Theta(a, L)) = \sigma^2 n^{-1} \sum_{k=1}^{\infty} (1 - c_n a_k)_+.$$

For the i.i.d. sequence space model, the value r_n is exactly the minimax linear risk on the ellipsoid $\Theta(a, L)$, and asymptotically r_n equals to the minimax risk among all estimators on $\Theta(a, L)$ [Pinsker (1980); see also Belitser and Levit (1995)]. As shown in Goldenshluger and Tsybakov (1999), the value r_n gives also a lower bound for the minimax risk in our problem, and this bound cannot be improved among all prediction methods. The next theorem is a corollary of the lower bound in Goldenshluger and Tsybakov (1999).

THEOREM 1. *Let Assumptions* 1 *and* 2 *hold. Assume that the ellipsoid* $\Theta(a, L)$ *is such that the sequence* $\{a_k\}$ *is monotone nondecreasing and there exist* $\beta > 1/2$, *and the positive constants* $a_{\min}$, $a_{\max}$ *such that*

$$a_{\min} k^{\beta} \le a_k \le a_{\max} k^{\beta}, \qquad k = 1, 2, \ldots .$$

Then for every prediction method $\hat{y} = \hat{y}(n + 1)$ *one has*

$$\mathscr{R}[\hat{y}; \Theta(a, L)] \ge r_n(1 + o(1)), \qquad n \to \infty. \tag{11}$$

Now we state the main results of the paper. First, we claim that the lower bound (11) is attained by the predictor $\hat{y}^*$; that is, this predictor is asymptotically minimax sharp adaptive on the scale of ellipsoids satisfying the assumptions of Theorem 1.

THEOREM 2. *Let Assumptions* 1–3 *hold. Let* $\{a_k\}$ *satisfy the assumptions of Theorem* 1. *Then the prediction method defined in* (9) *satisfies*

$$\mathscr{R}[\hat{y}^*; \Theta(a, L)] \le r_n(1 + o(1)), \qquad n \to \infty. \tag{12}$$

Next, the result of Theorem 2 can be extended to a larger class of ellipsoids, and the $o(1)$ term in (12) is uniformly small over this class as stated in the following theorem.

THEOREM 3. *Let Assumptions* 1–3 *hold. Let* $A(\beta_0, \beta_1)$ *be the set of monotone nondecreasing sequences* $a = \{a_k\}$ *such that*

$$a_{\min} k^{\beta_0} \le a_k \le a_{\max} k^{\beta_1}, \qquad k = 1, 2, \ldots,$$

where $2\beta_1/(2\beta_1+1) < \beta_0 \le \beta_1 < \infty$. *For given numbers* $0 < L_{\min} \le L_{\max} < \infty$, *let* $\mathscr{A}$ *denote the collection of pairs* (a, L) *such that* $a \in A(\beta_0, \beta_1)$ *and* $L \in [L_{\min}, L_{\max}]$. *Then the prediction method defined in* (9) *satisfies*

$$\sup_{(a, L) \in \mathscr{A}} \{\mathscr{R}[\hat{y}^*; \Theta(a, L)]/r_n\} = 1 + o(1), \qquad n \to \infty.$$

Proofs of Theorems 2 and 3 are given in Section 4.

REMARK. Theorems 2 and 3 can be extended in a standard way to the case where the variance σ^2 is not known. It suffices to replace σ^2 in the definition of $\hat{\theta}^*$ by a consistent estimator $\hat{\sigma}^2$. This can be a standard estimator of variance based on the sum of squares of the OLS residuals.

3. The main tools.

3.1. *Equivalent sequence model.* It follows from (4) that

$$\tilde{y}_{\mathscr{I}} = \theta_{\mathscr{I}} + \delta_{\mathscr{I}} + \frac{\sigma}{\sqrt{n}}\xi_{\mathscr{I}}, \qquad \mathscr{I} = \{1, \dots, N_n\}, \tag{13}$$

where

$$\begin{aligned} \delta_{\mathscr{I}} &= Q_{\mathscr{I}}^{-1}\left(-\frac{\theta_{\mathscr{I}}}{n} + \frac{1}{n}\sum_{t=1}^{n}\phi_N(t)\sum_{k=N_n+1}^{\infty}\theta_k x_k(t)\right), \\ \xi_{\mathscr{I}} &= Q_{\mathscr{I}}^{-1}\left(\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\phi_N(t)\tilde{\varepsilon}(t)\right), \end{aligned} \tag{14}$$

and $\tilde{\varepsilon} = \sigma^{-1}\varepsilon \sim \mathscr{N}(0,1)$.

Considering $\tilde{y}_{\mathscr{I}}$ as "new observations," we note that the model (13) is nongaussian with correlated nonzero mean errors. We note, however, that conditionally on the σ-algebra $\mathscr{F}_x^n = \sigma(\{x_k(t)\}_{k=1,2,\dots}, t = 1, \dots, n)$ the random vector $\tilde{y}_{\mathscr{I}}$ is Gaussian. Below we state a lemma showing that on a set of "large" probability (13) can be regarded as a Gaussian model with small correlations.

Note that

$$\mathbb{E}(\xi_{\mathscr{I}}|\mathscr{F}_x^n) = 0, \qquad S_{\mathscr{I}} = \mathbb{E}(\xi_{\mathscr{I}}\xi'_{\mathscr{I}}|\mathscr{F}_x^n) = Q_{\mathscr{I}}^{-1} - n^{-1}Q_{\mathscr{I}}^{-2}. \tag{15}$$

Let $v_{ij} = n^{-1}\sum_{t=1}^{n} x_i(t)x_j(t) - \delta_{ij}$, where $i, j = 1, \dots, N_n$, and δ_{ij} is the Kronecker delta.

LEMMA 1. *Let Assumptions* 1 *and* 2 *hold. Let* $\alpha \in (0,1)$. *Then there exists a constant* $q = q(c_*, H) > 0$ *such that*

$$\mathbb{P}(\Omega_\alpha) \geq 1 - \alpha, \tag{16}$$

where

$$\Omega_\alpha = \left\{\omega \in \Omega : \max_{i,\, j=1,\dots,N_n} |v_{ij}| \leq \sqrt{\frac{q}{n}\ln\frac{2N_n^2}{\alpha}}\right\}. \tag{17}$$

Furthermore, let

$$N_n\mu_n(\alpha) < \tfrac{1}{2}, \tag{18}$$

where

$$\mu_n(\alpha) = \sqrt{\frac{q}{n}\ln\frac{2N_n^2}{\alpha}} + \frac{1}{n}.$$

Then on the event Ω_α *we have* $S_{\mathscr{I}} = I - A_{\mathscr{I}}$, *where* $A_{\mathscr{I}}$ *is an* $N_n \times N_n$ *matrix satisfying*

$$\max_{i,\, j=1,\ldots,N_n} |[A_{\mathscr{I}}]_{ij}| \le 3\mu_n(\alpha), \tag{19}$$

and $[\cdot]_{ij}$ *stands for the* (i, j)*th entry of a matrix.*

Proof of the lemma is given in Section 4.

In view of Lemma 1 conditionally on $\mathscr{F}_x^n$ on the event Ω_α the model (13) can be regarded as a Gaussian model with small correlations. Our proof of Theorems 2 and 3 is based on combining several oracle inequalities. The first inequality shows that the risk of the blockwise Stein estimator is almost as small as the risk of the blockwise linear oracle in the Gaussian model with small correlations. The second inequality uses the first one and guarantees that the effect of the bias $\delta_{\mathscr{I}}$ is asymptotically negligible. Finally, the third inequality allows linking the risk of the blockwise linear oracle to that of the monotone linear one.

3.2. *An oracle inequality for a sequence model with correlated errors.* Denote $\varepsilon = \sigma n^{-1/2}$. Consider the model

$$y_{\mathscr{I}} = \theta_{\mathscr{I}} + \varepsilon\xi_{\mathscr{I}}, \tag{20}$$

where $\xi_{\mathscr{I}}$ is the Gaussian vector with zero mean, and the covariance matrix $Q_{\mathscr{I}} = I - A_{\mathscr{I}}$ such that $\max_{i,\, j} |[A_{\mathscr{I}}]_{ij}| \le \mu = \mu_\varepsilon < 1$.

Denote $y_{(j)} = y_{B_j}$, $j = 1, \ldots, J$, and introduce the Stein estimators,

$$\hat{\theta}_{(j)} = \left(1 - \frac{\varepsilon^2 n_j}{\|y_{(j)}\|^2}\right) y_{(j)}, \qquad j = 1, \ldots, J.$$

The estimate $\hat{\theta}_{\mathscr{I}}$ of the sequence $\theta_{\mathscr{I}}$ is given by

$$\hat{\theta}_{\mathscr{I}} = (\hat{\theta}_{(1)}, \hat{\theta}_{(2)}, \ldots, \hat{\theta}_{(J)}). \tag{21}$$

Consider the ideal blockwise linear risk

$$r_n^{\mathrm{BL}} = \sum_{j=1}^{J} \frac{\|\theta_{(j)}\|^2 \varepsilon^2 n_j}{\|\theta_{(j)}\|^2 + \varepsilon^2 n_j}.$$

It is well known [see, e.g., Efroimovich and Pinsker (1984), Johnstone (1998)] that in the sequence space model with independent errors, r_n^{BL} represents the minimal risk of linear estimators whose weights are constant on the blocks $\{B_j\}$. This minimum is attained on a "pseudo-estimator" that depends on θ (called a *blockwise linear oracle*).

LEMMA 2. *Let* $\{B_j\}$ *be the partition* (5) *and* (6). *Let* $n_j > 4$ *and* $\mu < 1/6 - 2/(3n_j)$ *for all* $j = 1, \ldots, J$. *Then for the estimate* (21) *in the model* (20) *we*

have

$$\mathbb{E}\|\hat{\theta}_{\mathscr{I}} - \theta_{\mathscr{I}}\|^2 \le r_n^{\mathrm{BL}} + 7\varepsilon^2\mu N + 4\varepsilon^2 J. \tag{22}$$

Proof of the lemma is given in the Appendix.

3.3. *Passage to infinite sequences and the* i.i.d. *approximation.* Denote by $\bar{\theta}(h)$ a linear estimator of $\theta \in \ell_2$ with nonrandom weights $h = \{h_k\}$, that is,

$$\bar{\theta}(h) = (h_1 y_1, h_2 y_2, \ldots).$$

Let $\mathbb{E}_*$ be the expectation with respect to $\{y_k\}$ satisfying

$$y_k = \theta_k + \varepsilon\xi_k, \qquad k = 1, 2, \ldots, \ \xi_k \text{ i.i.d. } \mathscr{N}(0,1). \tag{23}$$

The risk of the estimator $\bar{\theta}(h)$ in this model is

$$\mathscr{R}_\varepsilon(h,\theta) = \mathbb{E}_*\|\bar{\theta}(h) - \theta\|^2 = \sum_k (1-h_k)^2\theta_k^2 + \varepsilon^2\sum_k h_k^2,$$

where $\|\cdot\|$ is the ℓ_2-norm.

Let $\mathscr{H}$ be a set of sequences $\{h_k\}$ piecewise constant on the blocks B_j:

$$\mathscr{H} = \{\{h_k\}\colon h_k = h_{\kappa_j},\ \forall\, k \in B_j,\ \ j = 1, \ldots, J, \text{ and } h_k = 0, \forall\, k > N\}.$$

It is easy to see that

$$\inf_{h\in\mathscr{H}} \mathscr{R}_\varepsilon(h,\theta) = \sum_{j=1}^{J} \frac{\|\theta_{(j)}\|^2\varepsilon^2 n_j}{\|\theta_{(j)}\|^2 + \varepsilon^2 n_j} + \sum_{k=N+1}^{\infty}\theta_k^2 = r_n^{\mathrm{BL}} + \sum_{k=N+1}^{\infty}\theta_k^2.$$

LEMMA 3. *Let Assumptions* 1 *and* 2 *hold. Assume that the partition* (5) *and* (6) *satisfies the following conditions*:

$$n_j > 4, \ \ j = 1, \ldots, J_n, \ \ N_n \asymp \sqrt{n}(\ln n)^{-1}. \tag{24}$$

Then for $\hat{\theta}^*$ *defined in* (7) *and* (8) *and for every* $\theta \in \ell_2$ *such that* $\|\theta\| \le L$ *we have*

$$\mathbb{E}\|\hat{\theta}^* - \theta\|^2 \le \left(\inf_{h\in\mathscr{H}} \mathscr{R}_{\sigma/\sqrt{n}}(h,\theta) + \frac{4\sigma^2 J_n}{n}\right)(1 + o(1)), \qquad n \to \infty, \tag{25}$$

here $o(1)$ *is uniform over* $\|\theta\| \le L$.

The proof is given in Section 4.

Lemma 3 shows that, up to the factor $1 + o(1)$, the behavior of our estimator $\hat{\theta}^*$ is asymptotically at least as good as that of the blockwise Stein estimator for the i.i.d. model (23). In fact, in view of (22), we have for the model (23) that the risk of the blockwise Stein estimator $(\hat{\theta}_{(1)}, \ldots, \hat{\theta}_{(J)}, 0, 0, \ldots)$ is bounded by $\inf_{h\in\mathscr{H}} \mathscr{R}_\varepsilon(h,\theta) + 4\varepsilon^2 J$.

3.4. *The blockwise linear oracle is almost as good as the monotone linear one.* The oracle inequality (25) shows that our estimator is asymptotically almost as good as the best blockwise linear oracle for the i.i.d. Gaussian sequence space model. Here we will see that a similar result is valid for the linear monotone oracle in place of the blockwise linear oracle. This follows from Lemma 3 and the next lemma.

Denote $\mathscr{H}_{\mathrm{mon}}$ the set of monotone nonincreasing sequences $h = \{h_k\} \in \ell_2$,

$$\mathscr{H}_{\mathrm{mon}} = \{h = \{h_k\} \in \ell_2 \colon 1 \geq h_1 \geq h_2 \geq \cdots \geq 0\}.$$

LEMMA 4. *For any $\theta \in \ell_2$ and any partition* (5) *and* (6) *such that*

$$\frac{\mathrm{card}(B_{j+1})}{\mathrm{card}(B_j)} \leq 1 + \eta, \tag{26}$$

where $\eta = \eta_\varepsilon > 0$ we have

$$\inf_{h \in \mathscr{H}} \mathscr{R}_\varepsilon(h, \theta) \leq (1+\eta) \inf_{h \in \mathscr{H}_{\mathrm{mon}}} \mathscr{R}_\varepsilon(h, \theta) + \varepsilon^2 \,\mathrm{card}(B_1) + \sum_{k=N+1}^{\infty} \theta_k^2. \tag{27}$$

PROOF. It suffices to prove that for any $h \in \mathscr{H}_{\mathrm{mon}}$ there exists $\bar{h} \in \mathscr{H}$ such that

$$\mathscr{R}_\varepsilon(\bar{h}, \theta) \leq (1+\eta)\mathscr{R}_\varepsilon(h, \theta) + \varepsilon^2 \,\mathrm{card}(B_1) + \sum_{k=N+1}^{\infty} \theta_k^2. \tag{28}$$

Given $h \in \mathscr{H}_{\mathrm{mon}}$, define $\bar{h} \in \mathscr{H}$ by

$$\bar{h}_k = \begin{cases} 1, & k \in B_1, \\ h_{\kappa_j}, & k \in B_j,\ j = 2, \ldots, J, \\ 0, & k > N. \end{cases}$$

Since $0 \leq h_k \leq \bar{h}_k \leq 1$ for $k \leq N$ we have

$$\begin{aligned} \mathscr{R}_\varepsilon(\bar{h}, \theta) &= \sum_{k=1}^{\infty} (1 - \bar{h}_k)^2 \theta_k^2 + \varepsilon^2 \sum_{k=1}^{\infty} \bar{h}_k^2 \\ &\leq \sum_{k=1}^{N} (1 - h_k)^2 \theta_k^2 + \varepsilon^2 \sum_{k=1}^{N} \bar{h}_k^2 + \sum_{k=N+1}^{\infty} \theta_k^2. \end{aligned}$$

Hence to show (28) it suffices to prove that

$$\varepsilon^2 \sum_{k=1}^{N} \bar{h}_k^2 \leq (1+\eta)\varepsilon^2 \sum_{k=1}^{N} h_k^2 + \varepsilon^2 \,\mathrm{card}(B_1).$$

Note that $\sum_{k=1}^{N} \bar{h}_k^2 = \operatorname{card}(B_1) + \sum_{j=2}^{J} h_{\kappa_j}^2 \operatorname{card}(B_j)$. By (26) and monotonicity of $\{h_k\}$,

$$\sum_{j=2}^{J} h_{\kappa_j}^2 \operatorname{card}(B_j) \le (1+\eta) \sum_{j=1}^{J-1} h_{\kappa_{j+1}}^2 \operatorname{card}(B_j)$$

$$\le (1+\eta) \sum_{j=1}^{J-1} \sum_{k \in B_j} h_k^2 \le (1+\eta) \sum_{k=1}^{N} h_k^2. \qquad \square$$

REMARK. Inequality (27) can be used to prove minimax adaptivity of block rules for different statistical models after reduction to the i.i.d. Gaussian sequence model. In fact, minimaxity of block rules for a class of sequences Θ follows from Lemma 4 under the condition that the minimax estimator on Θ is linear and has the form $\bar{\theta}(h^*)$ with $h^* \in \mathscr{H}_{\mathrm{mon}}$.

4. Proofs.

PROOF OF LEMMA 1. First we will prove (16). For any $s > 0$ we have

$$\mathbb{P}\left\{ \max_{i,\, j=1,\ldots,N_n} |v_{ij}| \ge s \right\} \le \mathbb{P}\left\{ \max_{i=1,\ldots,N_n} \left| \frac{1}{n} \sum_{t=1}^{n} x_i^2(t) - 1 \right| \ge s \right\}$$

$$+ \mathbb{P}\left\{ \max_{i,\, j=1,\ldots,N_n,\, i \ne j} \left| \frac{1}{n} \sum_{t=1}^{n} x_i(t) x_j(t) \right| \ge s \right\}$$

$$\equiv P_1 + P_2.$$

It follows from (10) that there exist positive constants H_1, q_1 and q_2 such that for $k, j = 1, \ldots, N_n$, $k \ne j$,

$$\mathbb{E} \exp\{\lambda(x_k^2 - 1)\} \le \exp\{q_1 \lambda^2\},$$

$$\mathbb{E} \exp\{\lambda x_k x_j\} \le \exp\{q_2 \lambda^2\} \quad \text{for } |\lambda| < H_1.$$

Therefore by Theorem 2.6 from Petrov (1995) we have for $s < 2H_1 q_1$ and any $k = 1, \ldots, N_n$,

$$\mathbb{P}\left\{ \left| \frac{1}{n} \sum_{t=1}^{n} (x_k^2(t) - 1) \right| \ge s \right\} \le 2 \exp\left\{ -\frac{ns^2}{4q_1} \right\} \quad \text{and} \quad P_1 \le 2N_n \exp\left\{ -\frac{ns^2}{4q_1} \right\}.$$

The term P_2 is bounded similarly. Finally, we conclude that for $s \le 2H_1 \times \min(q_1, q_2)$

$$\mathbb{P}\left\{ \max_{i,\, j=1,\ldots,N_n} |v_{ij}| \ge s \right\} \le 2N_n^2 \exp\left\{ -\frac{ns^2}{4q_3} \right\},$$

where $q_3 = q_1 \wedge q_2$. This completes the proof of (16).

Let $C_{\mathscr{I}} = I - Q_{\mathscr{I}}$; then on the set Ω_α,

$$|[C_{\mathscr{I}}]_{ij}| \le |v_{ij}| + \delta_{ij} n^{-1} \le \mu_n(\alpha).$$

By (18), $\|C_{\mathscr{I}}\| \le N_n\mu_n(\alpha) < 1/2$, where $\|\cdot\|$ denotes the standard spectral matrix norm. Hence on the set Ω_α,

$$Q_{\mathscr{I}}^{-1} = (I - C_{\mathscr{I}})^{-1} = I + D_{\mathscr{I}}, \qquad D_{\mathscr{I}} = \sum_{k=1}^{\infty} C_{\mathscr{I}}^k. \tag{29}$$

We note also that

$$\begin{aligned}\max_{i,\,j=1,\ldots,N} |[C_{\mathscr{I}}^k]_{ij}| &\le N_n\mu_n(\alpha) \max_{i,\,j=1,\ldots,N_n} |[C_{\mathscr{I}}^{k-1}]_{ij}| \\ &\le N_n^{k-1}\mu_n^k(\alpha), \qquad k = 1, 2, \ldots,\end{aligned}$$

and therefore by (29) and (18),

$$\max_{i,\,j=1,\ldots,N_n} |[D_{\mathscr{I}}]_{ij}| \le \sum_{k=1}^{\infty} N_n^{k-1}\mu_n^k(\alpha) \le 2\mu_n(\alpha). \tag{30}$$

Now define $A_{\mathscr{I}} = -D_{\mathscr{I}}(1 - 2n^{-1}) + D_{\mathscr{I}}^2 n^{-1} + n^{-1}I$; then by definition $S_{\mathscr{I}} = I - A_{\mathscr{I}}$. It follows from (30) and (18) that

$$\begin{aligned}\max_{i,\,j=1,\ldots,N_n} |[A_{\mathscr{I}}]_{ij}| &\le 2\mu_n(\alpha)\left(1 - \frac{2}{n}\right) + \frac{4}{n}N_n\mu_n^2(\alpha) + \frac{1}{n} \\ &\le 2\mu_n(\alpha) - \frac{2\mu_n(\alpha)}{n} + \frac{1}{n} \le 3\mu_n(\alpha).\end{aligned}$$

This completes the proof. □

The following auxiliary result will be used in the proof of Lemma 3.

LEMMA 5. *Let Assumption* 1 *hold. Denote* $U_n(\alpha) = 1 + 2N_n\mu_n(\alpha)$. *Then for the subvectors* $\delta_{(j)} = \delta_{B_j}$, $j = 1, \ldots, J$ *we have*

(i)

$$\mathbb{E}\big[\|\delta_{(j)}\|^2 \mathbf{1}(\Omega_\alpha)\big] \le 2U_n^2(\alpha)\left[\frac{\|\theta_{(j)}\|^2}{n^2} + \frac{\sqrt{2q}\,n_j}{n}\sum_{k=N_n+1}^{\infty}\theta_k^2\right]; \tag{31}$$

(ii)

$$\|\mathbb{E}[\delta_{(j)}\mathbf{1}(\Omega_\alpha)]\| \le U_n(\alpha)\|\theta_{(j)}\|n^{-1}. \tag{32}$$

PROOF. (i) By (29) and (30) on the set Ω_α we have $\|Q_{\mathscr{I}}^{-1}\| \le U_n(\alpha)$. Then (31) is easily obtained as in the proof of Lemma 2 in Goldenshluger and Tsybakov (1999).

(ii) By (29) and (30) on the set Ω_α we have $Q_{\mathscr{I}}^{-1} = I + D_{\mathscr{I}}$. Therefore, conditioning on $\{x_k(t);\, t = 1, \ldots, n;\, k = 1, \ldots, N\}$ and using independence of

$x_k(t)$, $k = 1, 2, \ldots$ and the fact that, by Assumption 1, $\mathbb{E}(x_k(t)) = 0$, we obtain

$$\begin{aligned}\mathbb{E}[\delta_{\mathscr{I}}\mathbf{1}(\Omega_\alpha)] &= -\frac{\theta_{\mathscr{I}}}{n} + \mathbb{E}\biggl[\frac{1}{n}\sum_{t=1}^{n}\phi_N(t)\sum_{k=N_n+1}^{\infty}\theta_k x_k(t)\mathbf{1}(\Omega_\alpha)\biggr] \\ &\quad + \mathbb{E}\biggl[D_{\mathscr{I}}\biggl(-\frac{\theta_{\mathscr{I}}}{n} + \frac{1}{n}\sum_{t=1}^{n}\phi_N(t)\sum_{k=N_n+1}^{\infty}\theta_k x_k(t)\biggr)\mathbf{1}(\Omega_\alpha)\biggr] \\ &= -\frac{\theta_{\mathscr{I}}}{n}\bigl(I + \mathbb{E}[D_{\mathscr{I}}\mathbf{1}(\Omega_\alpha)]\bigr).\end{aligned}$$

Then (32) follows from (30). □

PROOF OF LEMMA 3. We proceed in steps.

Step 1′. Let Ω_α be defined by (17), and choose $\alpha = \alpha_* = 2N_n^2 n^{-8}$. Denote $\Omega_* = \Omega_{\alpha_*}$ and write

$$\mathbb{E}\|\hat{\theta}^* - \theta\|^2 = \mathbb{E}\bigl[\|\hat{\theta}^*_{\mathscr{I}} - \theta_{\mathscr{I}}\|^2\mathbf{1}(\Omega_*)\bigr] + \mathbb{E}\bigl[\|\hat{\theta}^*_{\mathscr{I}} - \theta_{\mathscr{I}}\|^2\mathbf{1}(\overline{\Omega}_*)\bigr] + \sum_{k=N_n+1}^{\infty}\theta_k^2. \tag{33}$$

Now we establish upper bounds on the first two terms in the r.h.s. of (33).

Step 2′. First, note that Ω_* is $\mathscr{F}_x^n$ measurable and

$$\mathbb{E}\bigl[\|\hat{\theta}^*_{\mathscr{I}} - \theta_{\mathscr{I}}\|\mathbf{1}(\Omega_*)\bigr] = \mathbb{E}\bigl\{\mathbb{E}\bigl[\|\hat{\theta}^*_{\mathscr{I}} - \theta_{\mathscr{I}}\|^2 \mid \mathscr{F}_x^n\bigr]\mathbf{1}(\Omega_*)\bigr\}.$$

Now consider the sequence model (13), and observe that conditionally on $\mathscr{F}_x^n$ the vector $\tilde{y}_{\mathscr{I}}$ is Gaussian. Note that with our choice $\alpha = \alpha_*$ we have $\mu_n(\alpha_*) = \mu_n^* = 2\sqrt{2qn^{-1}\ln n} + n^{-1}$. Thus, (18) is fulfilled for n large enough. If the event Ω_* holds, then by Lemma 1 the conditional covariance matrix of $\xi_{\mathscr{I}}$ equals $I - A_{\mathscr{I}}$, and inequality (19) is valid. Therefore, due to (24), we can apply Lemma 2 with $\mu_\varepsilon = 3\mu_n^*$ conditionally on $\mathscr{F}_x^n$ to estimate the vector $\theta_{\mathscr{I}} + \delta_{\mathscr{I}}$.

Thus, on the event Ω_* one has

$$\begin{aligned}\mathbb{E}\bigl[\|\hat{\theta}^*_{\mathscr{I}} - \theta_{\mathscr{I}} - \delta_{\mathscr{I}}\|^2 \mid \mathscr{F}_x^n\bigr] &\le \sum_{j=1}^{J_n}\frac{\|\theta_{(j)} + \delta_{(j)}\|^2(\sigma^2/n)n_j}{\|\theta_{(j)} + \delta_{(j)}\|^2 + (\sigma^2/n)n_j} \\ &\quad + \frac{\sigma^2}{n}(21\mu_n^* N_n + 4J_n).\end{aligned}$$

Taking expectation and using the Jensen inequality, we obtain

$$\begin{aligned}\mathbb{E}\bigl[\|\hat{\theta}^*_{\mathscr{I}} - \theta_{\mathscr{I}} - \delta_{\mathscr{I}}\|^2\mathbf{1}(\Omega_*)\bigr] &\le \sum_{j=1}^{J_n}\frac{\mathbb{E}\bigl[\|\theta_{(j)} + \delta_{(j)}\|^2\mathbf{1}(\Omega_*)\bigr](\sigma^2/n)n_j}{\mathbb{E}\bigl[\|\theta_{(j)} + \delta_{(j)}\|^2\mathbf{1}(\Omega_*)\bigr] + (\sigma^2/n)n_j} \\ &\quad + \frac{\sigma^2}{n}(21\mu_n^* N_n + 4J_n).\end{aligned}$$

Let $V_{j,n}(\alpha)$ denote the expression on the r.h.s. of (31). By Lemma 5,

$$\begin{aligned}\mathbb{E}\big[\|\theta_{(j)}+\delta_{(j)}\|^2\mathbf{1}(\Omega_*)\big] &\le \|\theta_{(j)}\|^2\mathbb{P}(\Omega_*)+\mathbb{E}\big[\|\delta_{(j)}\|^2\mathbf{1}(\Omega_*)\big]+2U_n^*n^{-1}\|\theta_{(j)}\|^2\\ &\le \|\theta_{(j)}\|^2(1+2U_n^*n^{-1})+V_{j,n}^*\end{aligned}$$

[here we use the notation $U_n^*=U_n(\alpha_*)$ and $V_{j,n}^*=V_{j,n}(\alpha_*)$]. On the other hand,

$$\mathbb{E}\big[\|\theta_{(j)}+\delta_{(j)}\|^2\mathbf{1}(\Omega_*)\big]\ge\|\theta_{(j)}\|^2(1-\alpha_*-2U_n^*n^{-1}).$$

Thus,

$$\begin{aligned}&\mathbb{E}\big[\|\hat\theta_{\mathscr{I}}^*-\theta_{\mathscr{I}}-\delta_{\mathscr{I}}\|^2\mathbf{1}(\Omega_*)\big]\\ &\quad\le\sum_{j=1}^{J_n}\frac{\{\|\theta_{(j)}\|^2(1+2U_n^*n^{-1})+V_{j,n}^*\}(\sigma^2/n)n_j}{\|\theta_{(j)}\|^2(1-\alpha_*-2U_n^*n^{-1})+(\sigma^2/n)n_j}+\frac{\sigma^2}{n}(21\mu_n^*N_n+4J_n)\\ &\quad\le\sum_{j=1}^{J_n}\bigg(\frac{\|\theta_{(j)}\|^2(\sigma^2/n)n_j}{\|\theta_{(j)}\|^2(1-\alpha_*-2U_n^*n^{-1})+(\sigma^2/n)n_j}+4U_n^*\frac{\sigma^2n_j}{n^2}+V_{j,n}^*\bigg)\\ &\qquad+\frac{\sigma^2}{n}(21\mu_n^*N_n+4J_n)\qquad(34)\\ &\quad\le\sum_{j=1}^{J_n}\frac{\|\theta_{(j)}\|^2(\sigma^2/n)n_j}{\|\theta_{(j)}\|^2+(\sigma^2/n)n_j}\bigg(1+\frac{\alpha_*+2U_n^*n^{-1}}{1-\alpha_*-2U_n^*n^{-1}}\bigg)\\ &\qquad+2U_n^*\frac{\sigma^2N_n}{n^2}+\sum_{j=1}^{J_n}V_{j,n}^*+\frac{\sigma^2}{n}(21\mu_n^*N_n+4J_n)\\ &\quad=\bigg(r_n^{\mathrm{BL}}+\frac{4\sigma^2J_n}{n}\bigg)(1+o(1))+o\bigg(\sum_{k=N_n+1}^{\infty}\theta_k^2\bigg),\qquad n\to\infty.\end{aligned}$$

Here we used that $2U_n^*N_nn^{-2}=o(n^{-1})$, $\mu_n^*=O(\sqrt{\ln n/n})$, $N_n=O(\sqrt{n}/\ln n)$, and

$$\begin{aligned}\sum_{j=1}^{J_n}V_{j,n}^*&=2(U_n^*)^2\bigg[\frac{\|\theta_{\mathscr{I}}\|^2}{n^2}+\frac{N_n}{n}\sum_{k\ge N_n+1}\theta_k^2\bigg]\\ &=O\bigg(\frac{1}{\sqrt{n}\ln n}\bigg)\sum_{k=N_n+1}^{\infty}\theta_k^2+O\bigg(\frac{1}{n^2}\bigg).\end{aligned}\qquad(35)$$

On the other hand, we have

$$\begin{aligned}&\mathbb{E}\big[\|\hat\theta_{\mathscr{I}}^*-\theta_{\mathscr{I}}\|^2\mathbf{1}(\Omega_*)\big]\\ &\quad\le\mathbb{E}\big[\|\hat\theta_{\mathscr{I}}^*-\theta_{\mathscr{I}}-\delta_{\mathscr{I}}\|^2\mathbf{1}(\Omega_*)\big]+\mathbb{E}[\|\delta_{\mathscr{I}}\|^2\mathbf{1}(\Omega_*)]\\ &\qquad+2\big(\mathbb{E}\big[\|\hat\theta_{\mathscr{I}}^*-\theta_{\mathscr{I}}-\delta_{\mathscr{I}}\|^2\mathbf{1}(\Omega_*)\big]\big)^{1/2}\big(\mathbb{E}\big[\|\delta_{\mathscr{I}}\|^2\mathbf{1}(\Omega_*)\big]\big)^{1/2}.\end{aligned}\qquad(36)$$

It follows from (31) and (35) that

$$\mathbb{E}\big[\|\delta_{\mathscr{I}}\|^2\mathbf{1}(\Omega_*)\big] \le \sum_{j=1}^{J_n} V^*_{j,n} = O\Big(\frac{1}{\sqrt{n}}\Big)\Big[\Big(r_n^{\mathrm{BL}} + \frac{4\sigma^2 J_n}{n}\Big) + o\Big(\sum_{k=N_n+1}^{\infty}\theta_k^2\Big)\Big].$$

Combining this with (34) and (36) we conclude that

$$\begin{aligned}\mathbb{E}\big[\|\hat\theta^*_{\mathscr{I}} - \theta_{\mathscr{I}}\|^2\mathbf{1}(\Omega_*)\big] &\le \Big(r_n^{\mathrm{BL}} + \frac{4\sigma^2 J_n}{n}\Big)(1+o(1)) \\ &\quad + o\Big(\sum_{k=N_n+1}^{\infty}\theta_k^2\Big), \qquad n\to\infty.\end{aligned} \tag{37}$$

Step 3′. Now we establish an upper bound on the second term in the r.h.s. of (33). We have

$$\mathbb{E}\big[\|\hat\theta^*_{\mathscr{I}} - \theta_{\mathscr{I}}\|^2\mathbf{1}(\overline{\Omega}_*)\big] = \sum_{j=1}^{J_n}\mathbb{E}\big[\|\hat\theta_{(j)} - \theta_{(j)}\|^2\mathbf{1}(\overline{\Omega}_*)\big] \le K_1 + K_2, \tag{38}$$

where

$$K_1 = 2\sum_{j=1}^{J_n}\mathbb{E}\big[\|\tilde y_{(j)} - \theta_{(j)}\|^2\mathbf{1}(\overline{\Omega}_*)\big],$$

$$K_2 = \frac{2\sigma^4}{n^2}\sum_{j=1}^{J_n}\Big(n_j^2\mathbb{E}\big[\|\tilde y_{(j)}\|^{-2}\mathbf{1}(\overline{\Omega}_*)\big]\Big).$$

It follows immediately from (14) and definition of the set Ω_* that

$$\begin{aligned}&\mathbb{E}\big[\|\delta_{\mathscr{I}}\|^2\mathbf{1}(\overline{\Omega_*})\big] \\ &\quad\le 2n^2\Big(n^{-1}\|\theta_{\mathscr{I}}\|^2\mathbb{P}(\overline{\Omega}_*) + \mathbb{E}\Big[\Big\|\frac{1}{n}\sum_{t=1}^{n}\phi_N(t)\sum_{k=N_n+1}^{\infty}\theta_k x_k(t)\Big\|^2\mathbf{1}(\overline{\Omega}_*)\Big]\Big) \\ &\quad\le 2n\|\theta_{\mathscr{I}}\|\alpha_* + 2n^2\Big[\mathbb{E}\Big\|\frac{1}{n}\sum_{t=1}^{n}\phi_N(t)\sum_{k=N_n+1}^{\infty}\theta_k x_k(t)\Big\|^4\Big]^{1/2}\sqrt{\alpha_*}.\end{aligned} \tag{39}$$

For every $k = 1,\dots,N$, due to Assumption 1,

$$\begin{aligned}&\mathbb{E}\Big(\frac{1}{n}\sum_{t=1}^{n}x_k(t)\sum_{j=N+1}^{\infty}\theta_j x_j(t)\Big)^4 \\ &\quad= \frac{1}{n^4}\mathbb{E}\Big[\sum_{t,s=1}^{n}x_k^2(t)x_k^2(s)\Big(\sum_{j=N+1}^{\infty}\theta_j x_j(t)\Big)^2\Big(\sum_{j=N+1}^{\infty}\theta_j x_j(s)\Big)^2\Big] \\ &\quad\le \frac{1}{n^2}\Big(\sum_{j=N+1}^{\infty}\theta_j^2\Big)^2 + \frac{C_1}{n^4}\sum_{t=1}^{n}\mathbb{E}\Big(\sum_{j=N+1}^{\infty}\theta_j x_j(t)\Big)^4 \\ &\quad\le \frac{L^4}{n^2} + \frac{C_1}{n^3}\sum_{j_1,\dots,j_4=N+1}^{\infty}\theta_{j_1}\theta_{j_2}\theta_{j_3}\theta_{j_4}\mathbb{E}\big[x_{j_1}(t)x_{j_2}(t)x_{j_3}(t)x_{j_4}(t)\big] \le C_2 n^{-2}\end{aligned}$$

(here $C_1 > 0$, $C_2 > 0$ are constants depending on L, c_* and H only). Using this bound and (39) we conclude that

$$\mathbb{E}\big[\|\delta_{\mathscr{I}}\|^2 \mathbf{1}(\overline{\Omega}_*)\big] \le 2n\|\theta_{\mathscr{I}}\|\alpha_* + 2C_2 N^2\sqrt{\alpha_*} = o(n^{-2}), \qquad n \to \infty.$$

Further, acting as in Lemma 3 in Goldenshluger and Tsybakov (1999) we get $\mathbb{E}(\|\xi_{\mathscr{I}}\|^4)^{1/2} \le C_3 N_n$, where $C_3 > 0$ depends on c_* and H only. Thus,

$$\begin{aligned} K_1 &= \mathbb{E}\big[\|\tilde{y}_{\mathscr{I}} - \theta_{\mathscr{I}}\|^2 \mathbf{1}(\overline{\Omega}_*)\big] \\ &\le 2\mathbb{E}\big[\|\delta_{\mathscr{I}}\|^2 \mathbf{1}(\overline{\Omega}_*)\big] + \frac{2\sigma^2}{n}\mathbb{E}\big[\|\xi_{\mathscr{I}}\|^4\big]^{1/2}\sqrt{\alpha_*} \\ &= o(n^{-2}), \qquad n \to \infty. \end{aligned} \tag{40}$$

It remains to bound K_2. We note again that conditionally on $\mathscr{F}_x^n$, $\tilde{y}_{(j)}$ is a Gaussian vector with mean $\theta_{(j)} + \delta_{(j)}$ and covariance matrix $\sigma^2 n^{-1} S_{(j)}$, where $S_{(j)}$ denotes the corresponding principle submatrix of the matrix $S_{\mathscr{I}}$ [see (15)]. Further, write $\tilde{y}_{(j)} = \sigma n^{-1/2} S_{(j)}^{1/2} \tilde{z}_{(j)}$, where $\tilde{z}_{(j)} \mid \mathscr{F}_x^n \sim \mathscr{N}(\sigma^{-1} n^{1/2} S_{(j)}^{-1/2}[\theta_{(j)} + \delta_{(j)}], I)$. Thus, we have

$$\begin{aligned} \mathbb{E}\big[\|\tilde{y}_{(j)}\|^{-2} | \mathscr{F}_x^n\big] &\le \frac{n}{\sigma^2 \lambda_{\min}[S_{(j)}]}\mathbb{E}\big[\|\tilde{z}_{(j)}\|^{-2} | \mathscr{F}_x^n\big] \\ &\le \frac{n}{\sigma^2 \lambda_{\min}[S_{(j)}](n_j - 2)}. \end{aligned} \tag{41}$$

The last inequality is a consequence of the following. Note that $\mathbb{E}[\|\tilde{z}_{(j)}\|^{-2} | \mathscr{F}_x^n] \le \mathbb{E}[\|\bar{z}_{(j)}\|^{-2} | \mathscr{F}_x^n]$ [cf. Lehmann and Casella (1998), page 355], where $\bar{z}_{(j)} | \mathscr{F}_x^n \sim \mathscr{N}(0, I)$. Then (41) follows from the fact that $\|\bar{z}_{(j)}\|^2 | \mathscr{F}_x^n \sim \chi^2_{(n_j)}$ and $n_j > 4$. Thus,

$$\begin{aligned} \mathbb{E}\big[\|\tilde{y}_{(j)}\|^{-2} \mathbf{1}(\overline{\Omega}_*)\big] &= \mathbb{E}\big\{\mathbb{E}\big[\|\tilde{y}_{(j)}\|^{-2} | \mathscr{F}_x^n\big]\mathbf{1}(\overline{\Omega}_*)\big\} \\ &\le \frac{n}{\sigma^2(n_j - 2)}\mathbb{E}\big\{\lambda_{\min}^{-1}[S_{(j)}]\mathbf{1}(\overline{\Omega}_*)\big\} \\ &\le \frac{n}{\sigma^2(n_j - 2)}\big\{\mathbb{E}\lambda_{\min}^{-2}[S_{(j)}]\big\}^{1/2}\sqrt{\alpha_*}. \end{aligned} \tag{42}$$

Since $S_{(j)}$ is a principle submatrix of $S_{\mathscr{I}}$, we have $\lambda_{\min}[S_{(j)}] \ge \lambda_{\min}[S_{\mathscr{I}}]$ [see, e.g., Marcus and Minc (1992), Chapter III, Section 3.6.5]. By definition of $Q_{\mathscr{I}}$ and independently of the event Ω_*, we have $\lambda_{\min}[Q_{\mathscr{I}}^{-1}] = \lambda_{\max}^{-1}[Q_{\mathscr{I}}] \ge (\sqrt{2q}N_n + n^{-1})^{-1}$. Therefore, if n is large enough, then $\lambda_{\min}[S_{\mathscr{I}}] \ge C_4 N_n^{-1}$, where $C_4 > 0$ depends only on q. Combining these inequalities with (42), we obtain

$$\mathbb{E}\big[\|\tilde{y}_{(j)}\|^{-2} \mathbf{1}(\overline{\Omega}_*)\big] \le \frac{C_5 n N_n^2}{\sigma^2(n_j - 2)n^4},$$

where $C_5 > 0$. This entails $K_2 = o(n^{-2})$, $n \to \infty$. Combining this with (33), (37), (38) and (40) we complete the proof. □

PROOF OF THEOREMS 2 AND 3. We prove only Theorem 2 because the proof of Theorem 3 is quite similar. We indicate how the statement of Theorem 3 follows from the proof of Theorem 2.

First note that Assumption 3 on the blocks guarantees that (24) is satisfied for n large enough. Note also that under this assumption,

$$J_n = O(\rho_n^{-1} \ln N_n) \quad \text{and} \quad \frac{\operatorname{card}(B_{j+1})}{\operatorname{card}(B_j)} \le 1 + O(\rho_n).$$

Therefore, using Lemmas 3 and 4 we get for $\theta \in \Theta(a, L)$ and for $\hat{\theta}^*$ defined in (7) and (8),

$$\begin{aligned}
&\mathbb{E}\|\hat{\theta}^* - \theta\|^2 \\
&\quad \le \left(\inf_{h \in \mathscr{H}} \mathscr{R}_{\sigma/\sqrt{n}}(h, \theta) + \frac{4\sigma^2 J_n}{n} \right)(1 + o(1)) \\
&\quad \le \left([1 + O(\rho_n)] \inf_{h \in \mathscr{H}_{\mathrm{mon}}} \mathscr{R}_{\sigma/\sqrt{n}}(h, \theta) + \frac{\sigma^2}{n}(4J_n + \nu_n) + \sum_{k=N+1}^{\infty} \theta_k^2 \right)(1 + o(1)),
\end{aligned}$$

where the $o(1)$ term is uniform in $\|\theta\| \le L$. Observe that

$$\mathbb{E}[\hat{y}^*(n+1) - y(n+1)]^2 - \sigma^2 = \mathbb{E}\|\hat{\theta}^* - \theta\|^2.$$

Hence

$$\begin{aligned}
&\mathbb{E}[\hat{y}^*(n+1) - y(n+1)]^2 - \sigma^2 \\
&\quad \le \left(\inf_{h \in \mathscr{H}_{\mathrm{mon}}} \mathscr{R}_{\sigma/\sqrt{n}}(h, \theta) + \frac{\sigma^2}{n}(4\rho_n^{-1} \ln N_n + \nu_n) + \frac{L^2}{a_{\min}^2} N_n^{-2\beta} \right)(1 + o(1)) \qquad (43) \\
&\quad \le (1 + o(1)) \inf_{h \in \mathscr{H}_{\mathrm{mon}}} \mathscr{R}_{\sigma/\sqrt{n}}(h, \theta) + O\left(\frac{\ln n}{n\rho_n} + (\ln n)^{2\beta} n^{-\beta} \right),
\end{aligned}$$

where we used that $\sum_{k=N_n+1}^{\infty} \theta_k^2 \le L^2/a_{N_n}^2 \le L_{\max}^2 N_n^{-2\beta}/a_{\min}^2$. Note that both $o(1)$ and $O(\cdot)$ in (43) are uniform over $\theta \in \cup_{(a,L) \in \mathscr{A}} \Theta(a, L)$. Taking the supremum of both sides of (43) over $\theta \in \Theta(a, L)$ and using the monotonicity of $\{a_k\}$ we get

$$\begin{aligned}
\mathscr{R}[\hat{y}^*; \Theta(a, L)] &\le \inf_{\bar{\theta}} \sup_{\theta \in \Theta(a, L)} \mathbb{E}_* \|\bar{\theta} - \theta\|^2 [1 + o(1)] \\
&\quad + O\left(\frac{\ln n}{n\rho_n} + (\ln n)^{2\beta} n^{-\beta} \right). \qquad (44)
\end{aligned}$$

By Pinsker's (1980) theorem [see also Belitser and Levit (1995)], we get

$$\inf_{\bar{\theta}} \sup_{\theta \in \Theta(a, L)} \mathbb{E}_* \|\bar{\theta} - \theta\|^2 = r_n(1 + o(1)) \ge c n^{-2\beta/(2\beta+1)} \qquad (45)$$

and $c > 0$ depends only on β, a and L. Recalling that $\beta > 1/2$ and combining the last two inequalities we get the result of Theorem 2.

The proof of Theorem 3 follows the same lines as the above proof. We need to modify only the last step. Since $\Theta(a, L) \supseteq \Theta(a^1, L_{\min}), \forall (a, L) \in \mathscr{A}$, where $a^1 = \{a_{\max} k^{\beta_1}\}$, instead of (45) we may write

$$\inf_{\bar\theta} \sup_{\theta \in \Theta(a, L)} \mathbb{E}_* \|\bar\theta - \theta\|^2 \geq \inf_{\bar\theta} \sup_{\theta \in \Theta(a^1, L_{\min})} \mathbb{E}_* \|\bar\theta - \theta\|^2 \geq c' n^{-2\beta_1/(2\beta_1+1)}, \tag{46}$$

where $c' > 0$ depends only on β_1, $a_{\max}$ and $L_{\min}$. On the other hand, $O(\frac{\ln n}{n\rho_n} + (\ln n)^{2\beta} n^{-\beta}) = O((\ln n)^{2\beta_1} n^{-\beta_0})$, since $\beta_0 \leq \beta \leq \beta_1$. This and (46) show that under the condition $2\beta_1/(2\beta_1 + 1) < \beta_0 \leq \beta_1$ the first term in the r.h.s. of (44) asymptotically dominates the second one uniformly over all pairs $(a, L) \in \mathscr{A}$. $\square$

APPENDIX

PROOF OF LEMMA 2. The proof is obtained as an immediate consequence of Lemma 7 given below. First we present an auxiliary lemma concerning the estimation of a d-dimensional Gaussian mean.

Let $y_k = \theta_k + \varepsilon\xi_k$, $k = 1, \ldots, d$, where $y = (y_1, \ldots, y_d)'$ is the observation vector, and $\theta = (\theta_1, \ldots, \theta_d)'$ is the parameter to be estimated. We assume that $\xi = (\xi_1, \ldots, \xi_d)' \sim \mathscr{N}_d(0, Q)$, where Q is a positive definite $d \times d$ matrix.

LEMMA 6. *Let* $g\colon \mathbb{R}^d \to \mathbb{R}^d$ *and*

$$\mathbb{E}\left(|y_k||g_k(y)| + \left| \frac{\partial g_k(y)}{\partial y_k} \right| \right) < \infty, \qquad k = 1, \ldots, d.$$

Then

$$\mathbb{E}\|y + g(y) - \theta\|^2 = \varepsilon^2 \operatorname{tr}(Q) + 2\varepsilon^2 \mathbb{E}\operatorname{tr}[QD_g(y)] + \mathbb{E}\|g(y)\|^2,$$

where $D_g = D_g(y) = \{\partial g_i/\partial y_j\}_{i,\, j=1,\ldots,d}$.

The proof is an easy modification of Stein's (1981) theorem [e.g., Johnstone (1998)].

Assume that $Q = I - A$, where A is a symmetric $d \times d$ matrix such that $\max_{i,\, j=1,\ldots,d} |[A]_{ij}| \leq \mu < 1$. Consider the Stein estimator

$$\tilde\theta = \left(1 - \frac{\varepsilon^2 d}{\|y\|^2}\right) y.$$

LEMMA 7. *Assume that* $d > 4$, *and* $\mu < 1/6 - 2/(3d)$. *Then*

$$\mathbb{E}\|\tilde\theta - \theta\|^2 \leq \frac{\|\theta\|^2 \varepsilon^2 d}{\|\theta\|^2 + \varepsilon^2 d} + 7\varepsilon^2 \mu d + 4\varepsilon^2. \tag{47}$$

For $\mu = 0$ this inequality is given in Donoho and Johnstone (1995).

PROOF. We apply Lemma 6 with $g(y) = -\varepsilon^2 d\|y\|^{-2} y$. First we note that

$$\frac{\partial g_i(y)}{\partial y_j} = \begin{cases} -\varepsilon^2 d\left(\dfrac{1}{\|y\|^2} - \dfrac{2y_i^2}{\|y\|^4}\right), & i = j, \\ 2\varepsilon^2 d\dfrac{y_i y_j}{\|y\|^4}, & i \neq j. \end{cases} \tag{48}$$

Thus, $\mathbb{E}\ \mathrm{tr}[D_g(y)] = -\varepsilon^2 d(d-2)\mathbb{E}\|y\|^{-2}$, and simple algebra leads to

$$\begin{aligned} \mathbb{E}\|\tilde{\theta} - \theta\|^2 &= \varepsilon^2\ \mathrm{tr}(I - A) + 2\varepsilon^2 \mathbb{E}\ \mathrm{tr}[(I - A)D_g(y)] + \mathbb{E}\|g(y)\|^2 \\ &= \varepsilon^2 d - \varepsilon^4 d(d-4)\mathbb{E}\|y\|^{-2} - \varepsilon^2\ \mathrm{tr}(A) - 2\varepsilon^2 \mathbb{E}\ \mathrm{tr}(AD_g). \end{aligned} \tag{49}$$

Further,

$$\mathbb{E}\ \mathrm{tr}(AD_g) = \ \mathrm{tr}(A\mathbb{E}(D_g)) \le \mu\mathbb{E}\left(\sum_{i,\,j=1}^{d} |[D_g]_{ij}|\right). \tag{50}$$

In view of (48), $\mathbb{E}|[D_g]_{ii}| \le \varepsilon^2 d\,\mathbb{E}(\|y\|^{-2} + 2y_i^2\|y\|^{-4})$, and $\mathbb{E}|[D_g]_{ij}| \le 2\varepsilon^2 d\mathbb{E}(|y_i y_j|\|y\|^{-4})$ for $i \neq j$. Therefore,

$$\begin{aligned} \mathbb{E}\left(\sum_{i,\,j=1}^{d} |[D_g]_{ij}|\right) &\le \varepsilon^2 d\sum_{i=1}^{d}\left[\mathbb{E}\|y\|^{-2} + 2\mathbb{E}(y_i^2\|y\|^{-4})\right] + 2\varepsilon^2 d \sum_{i,\,j=1,\,i\neq j}^{d} \mathbb{E}\left(|y_i y_j|\|y\|^{-4}\right) \\ &= \varepsilon^2(d^2+2d)\mathbb{E}\|y\|^{-2} + 2\varepsilon^2 d\mathbb{E}\left[\|y\|^{-4}\left(\left(\sum_{i=1}^{d}|y_i|\right)^2 - \sum_{i=1}^{d}|y_i|^2\right)\right] \\ &\le \varepsilon^2(d^2+2d)\mathbb{E}\|y\|^{-2} + 2\varepsilon^2 d(d-1)\mathbb{E}\left(\|y\|^{-4}\sum_{i=1}^{d}|y_i|^2\right) \\ &= 3\varepsilon^2 d^2 \mathbb{E}\|y\|^{-2}. \end{aligned}$$

Taking into account (50) and (49) we come to

$$\begin{aligned} \mathbb{E}\|\tilde{\theta} - \theta\|^2 &\le \varepsilon^2 d - \varepsilon^4 d(d-4)\mathbb{E}\|y\|^{-2} + \varepsilon^2\mu d + 6\varepsilon^4\mu d^2\mathbb{E}\|y\|^{-2} \\ &= \varepsilon^2 d - \varepsilon^4 d(d-4)\mathbb{E}\|y\|^{-2}\left(1 - \frac{6\mu d}{d-4}\right) + \varepsilon^2\mu d. \end{aligned} \tag{51}$$

By Jensen's inequality $\mathbb{E}\|y\|^{-2} \ge (\mathbb{E}\|y\|^2)^{-1} = (\|\theta\|^2 + \varepsilon^2 d)^{-1}$. This along with (51) and $6\mu d < d - 4$, leads to (47). □

REFERENCES

BELITSER, E. N. and LEVIT, B. YA. (1995). On minimax filtering on ellipsoids. *Math. Method Statist.* **4** 259–273.

BREIMAN, L. and FREEDMAN, D. (1983). How many variables should be entered in a regression equation? *J. Amer. Statist. Assoc.* **78** 131–136.

CAI, T. (1999). Adaptive wavelet estimation: a block thresholding and oracle inequality approach. *Ann. Statist.* **27** 898–924.

CAVALIER, L. and TSYBAKOV, A. (2000). Sharp adaptation for inverse problems with random noise. *Probab. Theory Related Fields*. To appear. Available at www.proba.jussieu.fr.

DONOHO, D. and JOHNSTONE, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.* **90** 1200–1224.

EFROMOVICH, S. (1999). *Nonparametric Curve Estimation*. Springer, New York.

EFROIMOVICH, S. YU. and PINSKER, M. S. (1984). Learning algorithm for nonparametric filtering. *Automat. Remote Control*, **11** 1434–1440.

EFROMOVICH, S. and PINSKER, M. S. (1996). Sharp-optimal and adaptive estimation for heteroscedastic nonparametric regression. *Statist. Sinica* **6** 925–942.

GOLDENSHLUGER, A. and TSYBAKOV, A. (1999). Optimal prediction for linear regression with infinitely many parameters. Available at www.proba.jussieu.fr.

JOHNSTONE, I. M. (1998). Function estimation in Gaussian noise: sequence models. Available at www-stat.stanford.edu/.

JOHNSTONE, I. M. (1999). Wavelet shrinkage for correlated data and inverse problems: adaptivity results. *Statist. Sinica* **9** 51–83.

LEHMANN, E. and CASELLA, G. (1998). *Theory of Point Estimation.* Springer, New York.

MARCUS, M. and MINC, H. (1992). *A Survey of Matrix Theory and Matrix Inequalities.* Dover, New York.

NEMIROVSKI, A. (2000). *Topics in Non-Parametric Statistics. Ecole d'eté de Probabilités Saint Flour XXVIII. Lecture Notes in Math.* **1738** 89–277. Springer, Berlin.

PETROV, V. V. (1995). *Limit Theorems of Probability Theory.* Clarendon Press, Oxford.

PINSKER, M. S. (1980). Optimal filtering of square integrable signals in Gaussian white noise. *Problems Inform. Transmission* **16** 120–133.

SHIBATA, R. (1981). An optimal selection of regression variables. *Biometrika* **68** 45–54.

STEIN, CH. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* **9** 1135–1151.

DEPARTMENT OF STATISTICS
UNIVERSITY OF HAIFA
HAIFA 31905
ISRAEL
E-MAIL: goldensh@stat.haifa.ac.il

LABORATOIRE DE PROBABILITÉS
ET MODÈLES ALÉATOIRE
UNIVERSITÉ PARIS VI
4PL. JUSSIEU
PARIS 75252
FRANCE