Electronic Journal of Statistics
Vol. 8 (2014) 432–455
ISSN: 1935-7524
DOI: 10.1214/14-EJS889

Sparse model selection under heterogeneous noise: Exact penalisation and data-driven thresholding

Laurent Cavalier[*,†]

Université Aix-Marseille, LATP, CMI
39 rue Joliot-Curie
F-13453 Marseille cedex 13, France

and

Markus Reiß[*,†]

Institut für Mathematik
Humboldt-Universität zu Berlin
Unter den Linden 6
10099 Berlin, Germany
e-mail: mreiss@math.hu-berlin.de

Abstract: We consider a Gaussian sequence space model $X_\lambda = f_\lambda + \xi_\lambda$, where the noise variables $(\xi_\lambda)_\lambda$ are independent, but with heterogeneous variances $(\sigma_\lambda^2)_\lambda$. Our goal is to estimate the unknown signal vector (f_λ) by a model selection approach. We focus on the situation where the non-zero entries f_λ are sparse. Then the heterogenous case is much more involved than the homogeneous model where $\sigma_\lambda^2 = \sigma^2$ is constant. Indeed, we can no longer profit from symmetry inside the stochastic process that one needs to control. The problem and the penalty do not only depend on the number of coefficients that one selects, but also on their position. This appears also in the minimax bounds where the worst coefficients will go to the larger variances. With a careful and explicit choice of the penalty, however, we are able to select the correct coefficients and get a sharp non-asymptotic control of the risk of our procedure. Some finite sample results from simulations are provided.

AMS 2000 subject classifications: Primary 62G05; secondary 62J05.
Keywords and phrases: Sparse oracle inequality, optimal threshold, statistical inverse problem, risk hull, penalized empirical risk, full subset selection, heteroskedastic noise.

Received December 2013.

*We thank Iain Johnstone, Debashis Paul and Thorsten Dickhaus for interesting discussions. We are grateful for the feedback by the referees, the associate editor and the editor which have lead to significant improvements. Financial support from the Deutsche Forschungsgemeinschaft via research unit FOR 1735 *Structural Inference in Statistics: Adaptation and Efficiency* is gratefully acknowledged.

†Tragically, after submission Laurent Cavalier passed away and the second author dedicates this paper to him.

1. Introduction

1.1. Motivation and main results

We consider the following sequence space model

$$X_\lambda = f_\lambda + \xi_\lambda, \quad \lambda \in \Lambda, \tag{1.1}$$

where (f_λ) are the real-valued coefficients of a signal and the noise variables $(\xi_\lambda) \sim \mathcal{N}(0, \Sigma)$ have a diagonal covariance matrix $\Sigma = \mathrm{diag}(\sigma_\lambda^2)$. Here Λ is a finite, but large index set. This heterogeneous model may appear in several frameworks where the variance is fluctuating, for example in heterogeneous regression, coloured noise, fractional Brownian motion models or especially in statistical inverse problems. For the latter setting the general literature is quite exhaustive, e.g. Johnstone and Silverman (1997); Abramovich and Silverman (1998); Cavalier *et al.* (2002); Cavalier (2004); Cavalier and Raimondo (2007); Cohen *et al.* (2004); Cavalier (2011); Donoho (1995); Hoffmann and Reiß (2008); Johnstone and Paul (2013); Rochet (2013), but mostly focusses on specific questions like universal thresholding, asymptotic minimax rates or level-wise thresholding. The goal here is to estimate the unknown parameter vector (f_λ) from the observations (X_λ) under general and unknown sparsity constraints. To this end a penalised empirical risk criterion, based on the so-called risk hull approach, is proposed for general families of possibly data-driven selection rules. This can be viewed as a (data-dependent) model selection procedure and results in a sparse oracle-type inequality.

Model selection is a core problem in statistics. One of the main reference in the field dates back to the information criterion AIC by Akaike (1973), but there is a huge amount of more recent work on this subject, in particular a precise analysis for high-dimensional and sparse data, see e.g. Birgé and Massart (2001); Golubev (2002); Abramovich *et al.* (2006); Massart (2007); Golubev (2011); Rochet (2013); Wu and Zhou (2013). Model selection is usually linked to the choice of a penalty and its precise choice is the main difficulty in model selection both from a theoretical and a practical perspective. Moreover, there is a close relationship between model selection and the popular thresholding procedures, where coefficients below a certain noise-related level are suppressed, cf. Golubev (2002); Abramovich *et al.* (2006); Massart (2007). The idea is that the search for a "good penalty" in model selection is indeed very much related to the choice of a "good threshold" in wavelet procedures. There exists also a fascinating connection between the false discovery rate control (FDR) and both thresholding and model selection, as studied in Abramovich *et al.* (2006); Benjamini and Hochberg (1995), which will become apparent in the second part of our paper. Our main structural assumption is that the parameter vector (f_λ) of interest is sparse, while we do neither know the position nor the number of non-zero entries. Sparsity is one of the leading paradigms nowadays and signals with a sparse representation in some basis (for example wavelets) or functions with sparse coefficients appear in many scientific fields, compare the discussions

in Abramovich *et al.* (2006); Golubev (2002, 2011); Wu and Zhou (2013) among many others.

In this paper, we consider the sequence space model with heterogeneous errors. Our goal is then to select among a family of models the best possible one, by use of a data-driven selection rule. In particular, one has to deal with the special heterogeneous nature of the observations, which must be reflected by the choice of the penalty. The heterogenous case is much more involved than the direct (homogeneous) model. Indeed, there is no more symmetry inside the stochastic process that one needs to control, since each empirical coefficient has its own variance. The problem and the penalty do not only depend on the number of coefficients that one selects, but also on their position. The penalty is in this sense non-local. We treat the case of general families of data-driven selection rules first and then specify to the full subset selection procedures and the computationally much easier thresholding rules via an FDR-type control. Using our model selection approach, the procedures are almost exact minimax (up to a factor 2). Moreover, the procedure is fully adaptive. Indeed, the sparsity index γ_n is unknown and we obtain an explicit penalty, valid in the mathematical proofs and directly applicable in simulations.

The heterogeneity also appears in the minimax lower bounds where the coefficients in the least favourable model will go to the larger variances. In the case of known sparsity γ_n, we consider also a non-adaptive threshold estimator and derive its minimax upper bound. This estimator exactly attains the lower bound for typical specifications of the noise levels (σ_λ^2) and is then minimax.

The paper is organized as follows. In the following subsection we give examples of problems where our heterogeneous model appears. Section 2 specifies our method and a provides a general oracle-type inequality for general families of selection rules. In Section 3 we consider the sparsity assumptions and obtain concrete results for the full subset selection and thresholding procedures. Section 4 presents the results on minimax lower and upper bounds. In Section 5 we present numerical results that document the finite-sample properties of the methods and we discuss implementation issues. All proofs are deferred to Section 6.

1.2. Examples

Heterogeneous regression

Consider first a model of heterogeneous regression

$$Y_i = f(x_i) + \sigma(x_i)\varepsilon_i, \quad i = 1, \dots, n,$$

where ε_i are i.i.d. standard Gaussian, but their variance are fluctuating depending on the design points x_i and f is some spiky unknown function. In this model $\Lambda = \{1, \dots, n\}$. By spiky function we mean that $f(x_i)$ is close to zero apart from a small subset of all design points x_i. These signals are frequently encountered in

applications (though rarely modeled in theoretical statistics), e.g. when measuring absorption spectra in physical chemistry (i.e. rare well-localised and strong signals) or jumps in log returns of asset prices (i.e. log-price increments which fluctuate at low levels except when larger shocks occur).

Coloured noise

Often in applications coloured noise models are adequate. Let us consider here the problem of estimating an unknown function observed with a noise defined by some fractional Brownian motion,

$$dY(t) = f(t)dt + \varepsilon dW_{-\alpha}(t), \quad t \in [0,1], \tag{1.2}$$

where f is an unknown $1-$periodic function in $L^2(0,1)$, $\int_0^1 f(t)dt=0$, ε is the noise level and $W_{-\alpha}$ is a fractional Brownian motion of index α (e.g., see Sowell (1990)). The fractional Brownian motion appears in econometric applications to model the long-memory phenomena, e.g. in Comte and Renault (1996). We are not interested in the fractional Brownian motion itself, but we want to estimate the unknown function f based on the noisy data $Y(t)$, as in Cavalier (2004); Johnstone (2011); Wang (1996). The model (1.2) is close to the standard Gaussian white noise model, which corresponds to the case $\alpha = 0$. Here, the behaviour of the noise is different. Let us point out the potential use of our approach here.

An important tool is fractional integration. In this framework, if the function f is supposed to be $1-$periodic, then the natural way is to consider the periodic version of fractional integration (given in (1.3)) such that

$$d^{-\alpha}f(x) = \int_{-\infty}^{x} \frac{(x-t)^{\alpha-1}}{\Gamma(\alpha)} f(t)dt, \tag{1.3}$$

and thus (see p.135 in Zygmund (1959)),

$$d^{-\alpha}e^{2\pi ikx} = \frac{e^{2\pi ikx}}{(2\pi ik)^{\alpha}}. \tag{1.4}$$

By integration and projection on the cosine (or sine) basis and using (1.4), one obtains the sequence space model (as in Cavalier (2004)),

$$X_\lambda = f_\lambda + \xi_\lambda, \ \lambda \in \Lambda = \mathbb{N},$$

where $\{\xi_\lambda\}$ are independent with $(\xi_\lambda)_\lambda \sim \mathcal{N}(0,\Sigma)$, where $\Sigma = \mathrm{diag}(\sigma_\lambda^2)$ and $\sigma_\lambda^2 = \varepsilon^2(2\pi\lambda)^{2\alpha}$.

Inverse problems

Consider the following framework of a general inverse problem

$$Y = Af + \varepsilon \dot{W},$$

where A is a known injective compact linear bounded operator, f an unknown d-dimensional function, $\dot{W}$ is a Gaussian white noise and $\varepsilon > 0$ the noise level. We will use here the framework of Singular Values Decomposition (SVD), see e.g. Cavalier (2011). Denote by φ_λ the eigenfunctions of the operator A^*A associated with the strictly positive eigenvalues $b_\lambda^2 > 0$. Remark that any function f may be decomposed in this orthonormal basis as $f = \sum_{\lambda\in\Lambda} f_\lambda \varphi_\lambda$, where $\lambda \in \Lambda$.

Let $\{\psi_\lambda\}_{\lambda\in\Lambda}$ be the normalized image basis $\psi_\lambda = b_\lambda^{-1} A\varphi_\lambda$. By projection and division by the singular values, we may obtain the empirical coefficients

$$b_\lambda^{-1}\langle Y, \psi_\lambda\rangle = b_\lambda^{-1}\langle Af, b_\lambda^{-1}A\varphi_\lambda\rangle + b_\lambda^{-1}\langle \varepsilon\dot{W}, \psi_\lambda\rangle = \langle f, \psi_\lambda\rangle + b_\lambda^{-1}\langle \varepsilon\dot{W}, \psi_\lambda\rangle.$$

We then obtain a model in the sequence space (see Cavalier *et al.* (2002))

$$X_\lambda = f_\lambda + \xi_\lambda,\ \lambda \in \Lambda,$$

with $(\xi_\lambda)_\lambda \sim \mathcal{N}(0, \Sigma)$ and $\Sigma = \mathrm{diag}(\varepsilon^2 b_\lambda^{-2})$.

2. Data-driven-subset selection

We consider the sequence space model (1.1) for coefficients of an unknown L^2-function f with respect to an orthornormal system (ψ_λ). The estimator over an arbitrary large, but finite index set Λ is then defined by

$$\hat{f}(h) = \sum_{\lambda\in\Lambda} \hat{f}_\lambda(h)\psi_\lambda \text{ with } \hat{f}_\lambda(h) := h_\lambda X_\lambda,$$

where $h = (h_\lambda)_\lambda \in \{0,1\}^\Lambda$. The empirical version of f is defined as

$$\tilde{f} = \sum_{\lambda\in\Lambda} X_\lambda \psi_\lambda.$$

We write $|h| = \#\{h_\lambda = 1\}$ and $n = \#\Lambda$ for the cardinality of Λ. By $\|A\|$ we denote the operator norm, i.e. the largest absolute eigenvalue.

The random elements $(X_\lambda)_\lambda$ take values in the sample space $\mathcal{X} = \mathbb{R}^\Lambda$. We now consider an arbitrary family $\mathcal{H} \subseteq \mathcal{H}_0 := \{h : \mathcal{X} \to \{0,1\}^\Lambda\}$ of Borel-measurable data-driven subset selection rules. Define an estimator by minimizing in the family $\mathcal{H}$ the penalized empirical risk:

$$h^\star = \arg\min_{h\in\mathcal{H}} \left\{ \|\hat{f}(h) - \tilde{f}\|^2 + 2Pen(h) \right\}. \tag{2.1}$$

Remark that $h^\star$ is defined in an equivalent way by

$$h^\star = \arg\min_{h\in\mathcal{H}} \bar{R}_{pen}(X, h),$$

where

$$\bar{R}_{pen}(X, h) = -\sum_{\lambda\in\Lambda} h_\lambda X_\lambda^2 + 2Pen(h).$$

Then, define the data-driven estimator

$$f^\star = \sum_{\lambda\in\Lambda} h^\star_\lambda X_\lambda \psi_\lambda. \tag{2.2}$$

In order to find an explicit and adequate penalty, we follow Cavalier and Golubev (2006) and apply the concept of a risk hull $\ell(f,h)$. The function ℓ provides an upper bound for the stochastic error in estimating f uniformly over possibly random selction rules h. Although it may thus still be stochastic, it is much easier to work with since it does no longer depend on (X_λ) directly. The following penalty function turns out to be natural:

$$Pen(h) = 2\sum_{j=1}^{|h|} \sigma^2_{(j)_h}(\log(ne/j) + j^{-1}\log_+(n\|\Sigma\|)), \tag{2.3}$$

where $\sigma^2_{(j)_h}$ denotes the j-th largest value among $\{h_\lambda\sigma^2_\lambda\}$ and $\log_+(z) = \max(\log z, 0)$. In general the second summand is of lower order and mainly provides non-asymptotic control. Note further that in the homogeneous case $\sigma_\lambda = \sigma$ the close approximation $\sum_{j=1}^{|h|}\log(ne/j) \approx |h|\log(ne/|h|)$ shows that the typical $2\sigma^2 k\log(n/k)$-form of the penalty for the sparsity index k is recovered, see Abramovich *et al.* (2006) for a survey of different results in this direction.

The next lemma shows that the penalty indeed provides a risk hull. The proof is based on bounding the tails of the corresponding order statistics and on worst-case permutations of the entries. We are not going to dwell on measurability issues there, taking outer expectations if necessary, which is easily justified by the arguments.

2.1 Lemma. *The function*

$$\ell(f,h) = \sum_{\lambda\in\Lambda}(1-h_\lambda)f^2_\lambda + Pen(h) + \sqrt{2}\min\left(\tfrac{1}{n}, \|\Sigma\|\right), \tag{2.4}$$

with the penalty from (2.3) *is a risk hull, i.e. we have*

$$\mathbf{E}\sup_{h\in\mathcal{H}_0}\left(\|\hat f(h) - f\|^2 - \ell(f,h)\right) \leqslant 0. \tag{2.5}$$

Based on this lemma we are able to derive a general oracle inequality for data-driven subset selection, which form the first fundamental result of the paper.

2.2 Theorem. *Let $h^\star$ be the data-driven rule defined in (2.1). For any $\delta \in (0,1)$, we have*

$$\begin{aligned}&\mathbf{E}_f\,\|\hat f(h^\star) - f\|^2\\ &\leqslant (1+\delta)\,\mathbf{E}_f\left[\inf_{h\in\mathcal{H}}\Big(\sum_{\lambda\in\Lambda}(1-h_\lambda)f^2_\lambda - \sum_{\lambda\in\Lambda}h_\lambda(X^2_\lambda - f^2_\lambda) + 2Pen(h)\Big)\right] + \Omega_\delta,\end{aligned}$$

where

$$\Omega_\delta := 4\sqrt{2}\min\left(\tfrac{1}{n}, \|\Sigma\|\right) + \frac{2}{\delta}\sum_{\lambda\in\Lambda}\min(f^2_\lambda, \sigma^2_\lambda).$$

3. Sparse representations

Let us consider the intuitive version of sparsity by assuming a small proportion of nonzero coefficients, in the spirit of Abramovich *et al.* (2006), i.e. the family

$$\mathcal{F}_0(\gamma_n) := \Big\{ f : \sum_{\lambda\in\Lambda} \mathbf{1}(f_\lambda \neq 0) \leqslant n\gamma_n \Big\}$$

where γ_n denotes the maximal proportion of nonzero coefficients.

Throughout, we assume that this proportion γ_n is such that asymptotically

$$\gamma_n \to 0 \text{ and } n\gamma_n \to \infty.$$

We first derive the results for the interesting cases of full subset selection and adaptive thresholding before discussing their relevance and comparing them with the literature.

Full subset selection

The goal here is to study the accuracy of the full model selection over the whole family of estimators, which offers 2^n possibilities to select a sub-model. Each coefficient may be chosen to be inside or outside the model. Let us consider the case where $\mathcal{H}$ denotes all deterministic subset selections,

$$\mathcal{H} = \Big\{ h : \mathcal{X} \to \{0,1\}^\Lambda \,|\, h(x) = \mathbf{1}_{\Lambda'},\ \Lambda' \subseteq \Lambda \Big\}. \tag{3.1}$$

We strive for an oracle inequality that involves only the noise levels of the truly active coordinates, i.e. involving $h_f := \mathbf{1}(f_\lambda \neq 0)_{\lambda\in\Lambda}$. To this end, write Σ_h for the covariance matrix of the ξ_λ restricted to the indices λ for which $h_\lambda = 1$, i.e.

$$\Sigma_h = \operatorname{diag}(\sigma_\lambda^2)_{\lambda\in\Lambda(h)}$$

with $\Lambda(h) = \{\lambda : h_\lambda = 1\}$.

3.1 Theorem. *Let $h^\star$ be the data-driven rule defined in (2.1) with $\mathcal{H}$ as in (3.1). We have, for $n \to \infty$, uniformly over $f \in \mathcal{F}_0(\gamma_n)$,*

$$\begin{aligned}\mathbf{E}_f \|\hat f(h^\star) - f\|^2 \leqslant (4+o(1))\|\Sigma_{h^f}\| \big(n\gamma_n \log(\gamma_n^{-1}) + \log(n\gamma_n)\log_+(n\|\Sigma\|)\big) \\ + 4\sqrt{2}\min\big(\tfrac{1}{n}, \|\Sigma\|\big).\end{aligned} \tag{3.2}$$

In particular, if $\log_+(\|\Sigma\|) = O(\log n)$ (i.e., any polynomial growth for $\|\Sigma\|$ is admissible) and $\frac{\|\Sigma_{h^f}\|}{\|\Sigma\|}\max(n\|\Sigma\|,1)n\gamma_n\log(\gamma_n^{-1}) \to \infty$, then we obtain

$$\mathbf{E}_f \|\hat f(h^\star) - f\|^2 \leqslant (4+o(1))\|\Sigma_{h^f}\| n\gamma_n \log(\gamma_n^{-1}). \tag{3.3}$$

In Section 5 we shall propose a greedy algorithm to find approximately the full subset selection rule. As the results there indicate, however, the statistical properties of the full subset selection rule are not so impressive that a profound study of algorithmic ways to overcome the complexity bound $O(2^n)$ seems necessary. In fact, adaptive thresholding not only in practice, but also theoretically offers a simple and quite successful way to perform sparse model selection.

Threshold estimators

Consider now a family of adaptive threshold estimators. The problem is to study the data-driven selection of the threshold. Let us consider the case where $\mathcal{H}$ denotes the threshold selection rules with arbitrary threshold values $t > 0$:

$$\mathcal{H} = \Big\{ h((X_\lambda)_\lambda) = \mathbf{1}(\lambda : |X_\lambda| > \sigma_\lambda t) \,|\, t > 0 \Big\}. \tag{3.4}$$

Note that $\mathcal{H}$ consists of $n = \#\Lambda$ different subset selection rules only and can be implemented efficiently using the order statistics of $(|X_\lambda|/\sigma_\lambda)_\lambda$.

3.2 Theorem. *Let $h^\star$ be the data-driven rules defined in (2.1) with $\mathcal{H}$ as in (3.4). If $\|\Sigma_{h_f}\| \log(\gamma_n^{-1}) \to \infty$, then we have, for $n \to \infty$, uniformly over $f \in \mathcal{F}_0(\gamma_n)$*

$$\mathbf{E}_f \, \|\hat{f}(h^\star) - f\|^2 \leqslant \Big(4n\gamma_n (\|\Sigma_{h_f}\| \log(\gamma_n^{-1}) + 8\|\Sigma\| \gamma_n (\log(\gamma_n^{-1}))^{1/2}) \\ + 2\log_+(n\|\Sigma\|)(2\|\Sigma_{h_f}\| \log(n\gamma_n) + 4\|\Sigma\| \log_+(n\gamma_n^2)) \Big)(1 + o(1)). \tag{3.5}$$

Assuming for Σ the growth bounds

$$\|\Sigma\| = O(\|\Sigma_{h_f}\| \gamma_n^{-1}) \text{ and } \|\Sigma\| \log_+(n\|\Sigma\|) = o(\|\Sigma_{h_f}\| n\gamma_n \log(\gamma_n^{-1}) / \log_+(n\gamma_n^2)),$$

with the second condition always verified if $\log_+(n\gamma_n^2) = 0$, this inequality simplifies to

$$\mathbf{E}_f \, \|\hat{f}(h^\star) - f\|^2 \leqslant (4 + o(1)) \|\Sigma_{h^f}\| n\gamma_n \log(\gamma_n^{-1}).$$

Discussion

Heterogeneous case. One may compare the method and its accuracy with other results in related frameworks. For example, Rochet (2013) considers a very close framework of model selection in inverse problems by using the SVD approach. This results in a noise (ξ_λ) which is heterogeneous and diagonal. Johnstone (2011); Johnstone and Paul (2013) study the related topic of inverse problems and Wavelet Vaguelette Decomposition (WVD), built on Birgé and Massart (2001). The framework in Johnstone (2011) is more general than ours. However, this leads to less precise results. In all their results Johnstone and Paul (2013); Rochet (2013) use universal constants which are not really controlled. This is even more important for the constants inside the method, for example in the penalty. Our method contains an explicit penalty. It is used in the mathematical results and also in simulations without additional tuning. A possible extension of our method to the dependent WVD case does not seem straight-forward.

Homogeneous case. Let us compare with other work for the homogeneous setting $\Sigma = \sigma^2 Id$. There exist a lot of results in this framework, see e.g. Abramovich *et al.* (2006); Johnstone (2011); Massart (2007); Wu and Zhou (2013). Again those results contain universal constants, not only in the mathematical results, but

even inside the methods. For example, constants in front of the penalty, but also inside the FDR technique, with an hyper-parameter q_n which has to be tuned.

The perhaps closest paper to our work is Golubev (2011) in the homogeneous case. Our penalty is analogous to "twice the optimal" penalty considered in Golubev (2011). This is due to difficulties in the heterogenous case, where the stochastic process that one needs to control is much more involved in this setting. Indeed, there is no more symmetry inside this stochastic process, since each empirical coefficient has its own variance. The problem and the penalty do not only depend on the number of coefficients that one selects, but also on their position. It is not obvious how the theory of ordered processes, the main tool in the homogeneous case, extends to the heterogeneous setting.

This leads to our main risk term $4\|\Sigma\|n\gamma_n \log(\gamma_n^{-1})$, while the risk $2\sigma^2 n\gamma_n \log(\gamma_n^{-1})$ is obtained in Golubev (2011). The potential loss of the factor 2 in the heterogeneous framework is possibly avoidable in theory, but in simulations the results seem comparably less sensitive to this factor than to other modifications, e.g. to how many data points, among the $n\gamma_n$ non-zero coefficients, are close to the critical threshold level, which defines some kind of effective sparsity of the problem (often much less than $n\gamma_n$). This effect is not treated in the theoretical setup in most of the FDR-related studies, where implicitly a worst case scenario of the coefficients' magnitude is understood.

4. Minimax bounds

Lower bound

First we present a lower bound in the minimax sense over $\mathcal{F}_0(\gamma_n)$. The proof is achieved by exhibiting a least favourable Bayesian prior on the signal coefficients. This reveals the statistical complexity of the model selection problem.

4.1 Theorem. *For any estimator $\hat{f}_n$ based on n observations we have the minimax lower bound*

$$\sup_{f\in\mathcal{F}_0(\gamma_n)} \mathbf{E}_f[\|\hat{f}_n - f\|^2] \geqslant \sup_{\alpha_n\in S_\Lambda(n\gamma_n,c_n)} 2\big(1+o(1)\big)\Big(\sum_{\lambda\in\Lambda} \sigma_\lambda^2 \alpha_{\lambda,n} \log(\alpha_{\lambda,n}^{-1})\Big)$$

for some $c_n \to 0$ where $S_\Lambda(R,c) = \{\alpha \in [0,c]^\Lambda \mid \sum_\lambda \alpha_\lambda \leqslant R(1-c)\}$ denotes the intersection of c-times the n-dimensional unit cube with $R(1-c)$-times the n-simplex and where $o(1) \to 0$ as $n \to \infty$.

The introduction of the sequence (c_n) is used to have a uniform control of the coefficients and to make deviations from expectations sufficiently small. In specific cases the expression remains asymptotically the same when the supremum is taken over all sequences (α_λ) with $\alpha_\lambda \in [0,1]$ and $\sum_\lambda \alpha_\lambda \leqslant n\gamma_n$. A more concrete lower bound is given in the following corollary.

4.2 Corollary. *Distributing mass uniformly over the r_n indices with largest values σ_λ in Theorem 4.1 yields the lower bound, as $n \to \infty$,*

$$\sup_{f \in \mathcal{F}_0(\gamma_n)} \mathbf{E}_f[\|\hat{f}_n - f\|^2] \geqslant 2n\gamma_n \log(\gamma_n^{-1})\big(1 + o(1)\big)\frac{1}{r_n}\sum_{i=1}^{r_n} \sigma_{(i)}^2$$

in terms of the inverse order statistics $\sigma_{(i)}^2$, provided $\log(n/r_n) = o(\log(\gamma_n^{-1}))$ (i.e., r_n must be somewhat larger than $n\gamma_n$).

For polynomial growth $\sigma_{(i)}^2 \sim (n-i)^\beta$, $\beta > 0$, the lower bound simplifies to, as $n \to \infty$,

$$\sup_{f \in \mathcal{F}_0(\gamma_n)} \mathbf{E}_f[\|\hat{f}_n - f\|^2] \geqslant 2\big(1 + o(1)\big)\|\Sigma\| n\gamma_n \log(\gamma_n^{-1}).$$

Remark that the lower bound is a kind of weighted entropy. In contrast to the upper bounds above the minimax (and the Bayes) lower bound does not involve the quantity $\|\Sigma_{h_f}\|$, individual to each unknown f. In the proof for this heterogeneous model, conceptually we need to allow for a high complexity of the class $\mathcal{F}_0(\gamma_n)$, leading to the entropy factor $\log(\gamma_n^{-1})$, and to put more prior probability on coefficients with larger variance, which explains the abstract weighted entropy expression.

Upper bound

Consider now the setting where the sparsity γ_n is known and a correctly tuned threshold estimator is applied in order to identify the unknown positions of the significant non-zero coefficients f_λ. This leads to the following non-adaptive minimax upper bound.

4.3 Theorem. *Consider the threshold estimator defined coordinate-wise by*

$$\hat{f}_\lambda = X_\lambda \mathbf{1}_{\{X_\lambda^2 > 2\sigma_\lambda^2 \log(\alpha_{\lambda,n}^{-1})\}} \text{ with } \alpha_{\lambda,n} := e^{-\beta_n/\sigma_\lambda^2}$$

and $\beta_n > 0$ chosen such that $\sum_{\lambda \in \Lambda} \alpha_{\lambda,n} = n\gamma_n$. Then, as $n \to \infty$,

$$\sup_{f \in \mathcal{F}_0(\gamma_n)} \mathbf{E}_f[\|\hat{f}_n - f\|^2] \leqslant 2n\gamma_n\beta_n(1 + o(1))$$

holds. This implies that, as $n \to \infty$,

$$\sup_{f \in \mathcal{F}_0(\gamma_n)} \mathbf{E}_f[\|\hat{f}_n - f\|^2] \leqslant 2n\gamma_n \log(\gamma_n^{-1})\|\Sigma\|(1 + o(1)),$$

which is minimax optimal for at most polynomial growth in (σ_λ^2) by the lower bound in Corollary 4.2.

Let us mention that for faster growth than polynomial, we might well have $\beta_n = \log(\gamma_n^{-1})o(\|\Sigma\|)$. So, in general the upper bound matches exactly the lower bound with respect to the term $2n\gamma_n \log(\gamma_n^{-1})$, while the influence of the heterogeneous noise depends on the specific case. This procedure, however, is non-adaptive since the threshold relies on the knowledge of the sparsity γ_n.

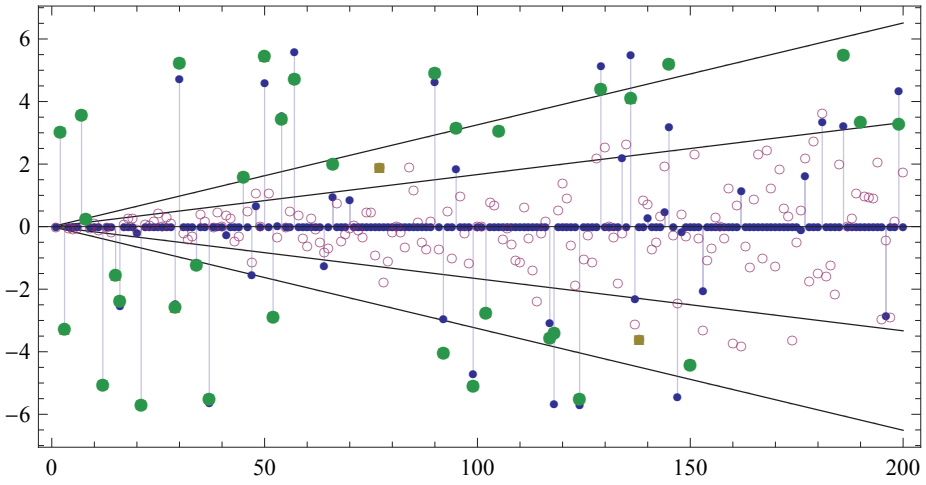

FIG 1. *Coefficients* (f_λ) *(blue), observations* (X_λ) *(green circle in full subset, green circle/yellow box in adaptive threshold, magenta open circle not taken) and universal/minimax thresholds as black lines (parameter values:* $n = 200$, $\gamma_n = 0.25$, $\sigma_\lambda = 0.01\lambda$ *for* $\lambda = 1 \dots n$*).*

5. A numerical example

In Figure 1 a typical realisation of the coefficients f_λ is shown in blue with 50 non-zero coefficients chosen uniformly on $[-6, 6]$ and increasing noise level $\sigma_\lambda = 0.01\lambda$ for every $\lambda = 1, \dots, 200$. The inner black diagonal lines indicate the minimax threshold (with oracle value of γ_n) and the outer diagonal lines the universal threshold. For each λ the non-blue points depict noisy observations X_λ, obtained by adding $N(0, \sigma_\lambda^2)$ noise to the blue point f_λ. Observations included in the adaptive full subset selection estimator are coloured green, while those included for the adaptive threshold estimator are the union of green and yellow points (in fact, for this sample the adaptive thresholding selects all full subset selected points), the discarded observations are in magenta.

We have run 1000 Monte Carlo experiments for the parameters $n = 200$, $\sigma_\lambda = 0.01\lambda$ in the sparse ($\gamma_n = 0.05$) and dense ($\gamma_n = 0.25$) case. In Figure 2 the first 100 relative errors are plotted for the different estimation procedures in the dense case, which offers more detailed sample information than mere boxplots. The errors are taken as a quotient with the sample-wise oracle threshold value applied to the renormalised X_λ/σ_λ. Therefore only the full subset selection can sometimes have relative errors less than one. Table 1 lists the relative Monte Carlo errors for the two cases. The last column reports the relative error of the oracle procedure with $h_\lambda = \mathbf{1}(f_\lambda \neq 0)$ that discards all observations X_λ with $f_\lambda = 0$ (not noticing the model selection complexity).

The simulation results are quite stable for variations of the setup. Altogether the thresholding works globally well. The (approximate) full subset selection procedure (see below for the greedy algorithm used) is slightly worse and exhibits

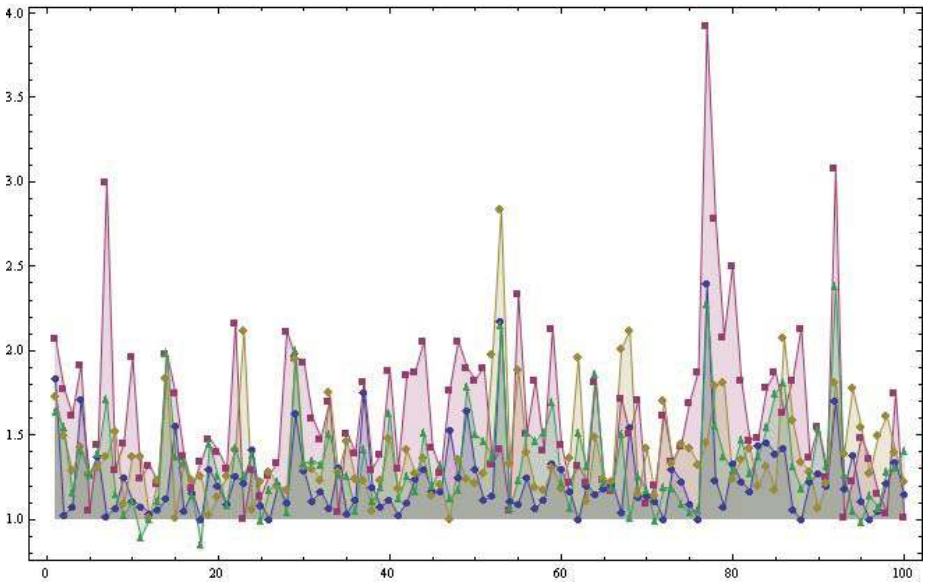

FIG 2. *First 100 Monte Carlo relative errors as in Table 1: adaptive (blue), universal (magenta) and minimax (yellow) thresholding, full subset selection (green).*

TABLE 1
Relative errors from 1000 Monte Carlo simulations as in Figure 1

γ_n	Adaptive Thr.	Universal Thr.	Sparse Thr.	Full Subset	Known Model
0.05	1.81	1.80	2.26	1.86	0.55
0.25	1.22	1.62	1.39	1.33	0.53

a higher variability, but is still pretty good. By construction, in the dense case the oracle minimax threshold works better than the universal threshold, while the universal threshold works better in very sparse situations. The reason why the minimax threshold even with a theoretical oracle choice of γ_n does not work so well is that the entire theoretical analysis is based upon potentially most difficult signal-to-noise ratios, that is coefficients f_λ of the size of the threshold or the noise level. Here, however, the effective sparsity is larger (i.e., effective γ_n is smaller) because the uniformly generated non-zero coefficients can be relatively small especially at indices with high noise level, see also Figure 1.

Let us briefly describe how the adaptive full subset selection procedure has been implemented. The formula (2.3) attributes to each selected coefficient X_λ the individual penalty $p_\lambda^h = 2\sigma_\lambda^2(\log(ne/r_\lambda(h)) + \log_+(n\|\Sigma\|)/r_\lambda(h)$ with the inverse rank $r_\lambda(h)$ of $(h_\lambda\sigma_\lambda^2)_\lambda$ (e.g., $r_\lambda(h) = 1$ if $h_\lambda\sigma_\lambda^2 = \max_{\lambda'} h_{\lambda'}\sigma_{\lambda'}^2$). Due to $p_\lambda^h \leqslant 2\sigma_\lambda^2(\log(ne) + \log_+(n\|\Sigma\|))$ all coefficients with

$$X_\lambda^2/\sigma_\lambda^2 \geqslant 4(\log(ne) + \log_+(n\|\Sigma\|))$$

are included into h_1^* in an initial step. Then, iteratively h_i^* is extended to h_{i+1}^*

by including all coefficients with

$$X_\lambda^2/\sigma_\lambda^2 \geqslant 4(\log(ne/r_\lambda(h_i^*)) + \log_+(n\|\Sigma\|)/r_\lambda(h_i^*)).$$

The iteration stops when no further coefficients can be included. The estimator h_I^* at this stage definitely contains all coefficients also taken by h^*. In a second iteration we now add in a more greedy way coefficients that will decrease the total penalized empirical risk. Including a new coefficient X_{λ_0}, adds to the penalized empirical risk the (positive or negative) value

$$\begin{aligned} &- X_{\lambda_0}^2 + 4\sigma_{\lambda_0}^2(\log(ne/r_{\lambda_0}(h_I^*)) + \log_+(n\|\Sigma\|))/r_{\lambda_0}(h_I^*) \\ &\quad - 4\sum_{\lambda:\sigma_\lambda<\sigma_{\lambda_0}} (h_I^*)_\lambda \sigma_\lambda^2(\log(1+1/r_\lambda(h_I^*)) + \log_+(n\|\Sigma\|)/(r_\lambda(h_I^*)(r_\lambda(h_I^*)+1))). \end{aligned}$$

Here, $r_{\lambda_0}(h_I^*)$ is to be understood as the rank at λ_0 when setting $(h_I^*)_{\lambda_0} = 1$. Consequently, the second iteration extends h_I^* each time by one coefficient X_{λ_0} for which the displayed formula gives a negative value until no further reduction of the total penalized empirical risk is obtainable. This second greedy optimisation does not necessarily yield the optimal full subset selection solution, but most often in practice it yields a coefficient selection h^* with a significantly smaller penalized empirical risk than the adaptive threshold procedure. The numerical complexity of the algorithm is of order $O(n^2)$ due to the second iteration in contrast to the exponential order $O(2^n)$ when scanning all possible subsets. A more refined analysis of our procedure would be interesting, but might have minor statistical impact in view of the good results for the straight-forward adaptive thresholding scheme.

6. Proofs

6.1. Proof of Lemma 2.1

Recall $n = \#\Lambda$ and introduce the stochastic term

$$\eta(h) = \sum_{\lambda\in\Lambda} h_\lambda \xi_\lambda^2. \tag{6.1}$$

Remark that $\|\hat f(h) - f\|^2 = \sum_{\lambda\in\Lambda}(1-h_\lambda)f_\lambda^2 + \eta(h)$ such that

$$\mathbf{E}\sup_{h\in\mathcal{H}_0}\left(\|\hat f(h)-f\|^2 - \sum_{\lambda\in\Lambda}(1-h_\lambda)f_\lambda^2 - Pen(h) - \sqrt{2}\min\left(\tfrac{1}{n}, \|\Sigma\|\right)\right) \leqslant 0 \tag{6.2}$$

follows from

$$\mathbf{E}\sup_{h\in\mathcal{H}_0}(\eta(h) - Pen(h)) \leqslant \sqrt{2}\min\left(\tfrac{1}{n}, \|\Sigma\|\right). \tag{6.3}$$

Let us write $\zeta_\lambda = \sigma_\lambda^{-1}\xi_\lambda \sim \mathcal{N}(0,1)$ and let $r_\lambda(h)$ denote the inverse rank of $h_\lambda\sigma_\lambda^2$ in $(h_{\lambda'}\sigma_{\lambda'}^2)_{\lambda'}$ (e.g., $r_\lambda(h)=1$ if $h_\lambda\sigma_\lambda^2 = \max_{\lambda'} h_{\lambda'}\sigma_{\lambda'}^2$) such that

$$\eta(h) - Pen(h) = \sum_{\lambda\in\Lambda} h_\lambda\sigma_\lambda^2\Big(\zeta_\lambda^2 - 2\Big(\log\Big(\frac{ne}{r_\lambda(h)}\Big) + r_\lambda(h)^{-1}\log_+(n\|\Sigma\|)\Big)\Big).$$

Note that for any enumeration $(\lambda_j)_{j=1,\dots,k}$ of $\{\lambda \,|\, h_\lambda = 1\}$ by monotonicity:

$$\begin{aligned}&\sum_{\lambda\in\Lambda} h_\lambda\sigma_\lambda^2\Big(\log(ne/r_\lambda(h)) + r_\lambda(h)^{-1}\log_+(n\|\Sigma\|)\Big)\\ &\geqslant \sum_{j=1}^{k}\sigma_{\lambda_j}^2\Big(\log(ne/j) + j^{-1}\log_+(n\|\Sigma\|)\Big)\end{aligned}$$

holds. We therefore obtain with the inverse order statistics $(\sigma_{(i)}^2)$ and $(\zeta_{(i)}^2)$ (i.e. $\sigma_{(1)}^2 \geqslant \sigma_{(2)}^2 \geqslant \cdots$ etc.) of $(\sigma_\lambda^2)_{\lambda\in\Lambda}$ and $(\zeta_\lambda^2)_{\lambda\in\Lambda}$, respectively,

$$\begin{aligned}&\mathbf{E}\Big[\sup_{h\in\mathcal{H}_0}(\eta(h) - Pen(h))_+\Big]\\ &\leqslant \mathbf{E}\Big[\sum_{j=1}^{n}\sigma_{(j)}^2\Big(\zeta_{(j)}^2 - 2(\log(ne/j) + j^{-1}\log_+(n\|\Sigma\|))\Big)_+\Big].\end{aligned}$$

It remains to evaluate $\mathbf{E}[(\zeta_{(j)}^2 - 2(\log(ne/j) + j^{-1}\log_+(n\|\Sigma\|)))_+]$. We obtain by independence, $\log(\binom{n}{k}) \leqslant k\log(ne/k)$ and by the Mill ratio inequality $P(\zeta_\lambda > t) \leqslant t^{-1}e^{-t^2/2}$

$$\begin{aligned}P(\zeta_{(j)}^2 > \kappa) &= P(\exists i_1,\dots,i_j \forall l \in \{1,\dots,j\} : \zeta_{i_l}^2 > \kappa)\\ &\leqslant \binom{n}{j}P(\zeta_\lambda^2 > \kappa)^j \leqslant \kappa^{-1/2}\exp(j\log(ne/j) - j\kappa/2).\end{aligned}$$

This implies for any $p > 0$

$$\mathbf{E}[(\zeta_{(j)}^2 - p)_+] = \int_p^\infty P(\zeta_{(j)}^2 > \kappa)\,d\kappa \leqslant 2j^{-1}p^{-j/2}\exp(j\log(ne/j) - jp/2).$$

We conclude

$$\begin{aligned}&\mathbf{E}\Big[\sum_{j=1}^{n}\sigma_{(j)}^2\Big(\zeta_{(j)}^2 - 2(\log(ne/j) + j^{-1}\log_+(n\|\Sigma\|))\Big)_+\Big]\\ &\leqslant 2\|\Sigma\|\sum_{j=1}^{n} j^{-1}(2\log(ne/j))^{-j/2}\exp(-\log_+(n\|\Sigma\|))\\ &\leqslant \min\big(\tfrac{1}{n}, \|\Sigma\|\big)\sup_n 2\sum_{j=1}^{n} j^{-1}(2\log(ne/j))^{-j/2} \leqslant \sqrt{2}\min\big(\tfrac{1}{n}, \|\Sigma\|\big),\end{aligned}$$

where $\sigma_{(j)}^2 \leqslant \|\Sigma\|$ and the supremum is attained at $n=1$ with value $\sqrt{2}$.

6.2. Proof of Theorem 2.2

In view of Lemma 2.1,

$$\ell(f,h)=\sum_{\lambda\in\Lambda}(1-h_\lambda)f_\lambda^2+Pen(h)+\sqrt{2}\min\left(\tfrac{1}{n},\|\Sigma\|\right) \tag{6.4}$$

is a risk hull, and therefore we have

$$\mathbf{E}_f\|\hat f(h^\star)-f\|^2\leqslant\mathbf{E}_f\ell(f,h^\star). \tag{6.5}$$

On the other hand, since $h^\star$ minimizes $\bar R_{pen}(X,h)$ we have

$$\mathbf{E}_f\bar R_{pen}(X,h^\star)=\mathbf{E}_f\Big[\min_{h\in\mathcal{H}}\bar R_{pen}(X,h)\Big]. \tag{6.6}$$

In order to combine the inequalities (6.5) and (6.6), we rewrite $\ell(f,h^\star)$ in terms of $\bar R_{pen}(X,h^\star)$

$$\begin{aligned}\ell(f,h^\star)=\bar R_{pen}(X,h^\star)+\|f\|^2+\sqrt{2}\min\left(\tfrac{1}{n},\|\Sigma\|\right)+\sum_{\lambda\in\Lambda}h_\lambda^\star\xi_\lambda^2+\sum_{\lambda\in\Lambda}2f_\lambda h_\lambda^\star\xi_\lambda\\ +Pen(h^\star)-2Pen(h^\star).\end{aligned} \tag{6.7}$$

Therefore, using this equation and (6.5, 6.6), we obtain

$$\begin{aligned}\mathbf{E}_f\|\hat f(h^\star)-f\|^2\leqslant&\mathbf{E}_f\Big[\min_{h\in\mathcal{H}}\bar R_{pen}(X,h)\Big]+\|f\|^2+\sqrt{2}\min\left(\tfrac{1}{n},\|\Sigma\|\right)\\ &+2\mathbf{E}_f\sum_{\lambda\in\Lambda}h_\lambda^\star f_\lambda\xi_\lambda+\mathbf{E}_f\left[\sum_{\lambda\in\Lambda}h_\lambda^\star\xi_\lambda^2-Pen(h^\star)\right].\end{aligned} \tag{6.8}$$

Remark now that for any deterministic index set $\Lambda'\subseteq\Lambda$

$$\mathbf{E}_f\sum_{\lambda\in\Lambda'}2h_\lambda^\star f_\lambda\xi_\lambda+\mathbf{E}_f\sum_{\lambda\in\Lambda'}2(1-h_\lambda^\star)f_\lambda\xi_\lambda=\mathbf{E}_f\sum_{\lambda\in\Lambda'}2f_\lambda\xi_\lambda=0. \tag{6.9}$$

This implies for $\Lambda_1:=\{\lambda\in\Lambda:f_\lambda^2>\sigma_\lambda^2\}$

$$\mathbf{E}_f\sum_{\lambda\in\Lambda}2h_\lambda^\star f_\lambda\xi_\lambda=-\mathbf{E}_f\sum_{\lambda\in\Lambda_1}2(1-h_\lambda^\star)f_\lambda\xi_\lambda+\mathbf{E}_f\sum_{\lambda\in\Lambda_1^0}2h_\lambda^\star f_\lambda\xi_\lambda. \tag{6.10}$$

Then, by the general inequality $2AB\leqslant\frac{\delta}{2}A^2+\frac{2}{\delta}B^2$ for $A,B,\delta>0$ we obtain

$$\left|\mathbf{E}_f\sum_{\lambda\in\Lambda_1}2(1-h_\lambda^\star)\xi_\lambda f_\lambda\right|\leqslant\tfrac{\delta}{2}\mathbf{E}_f\sum_{\lambda\in\Lambda}(1-h_\lambda^\star)f_\lambda^2+\tfrac{2}{\delta}\mathbf{E}_f\sum_{\lambda\in\Lambda_1}(1-h_\lambda^\star)\xi_\lambda^2. \tag{6.11}$$

Note that in terms of the trace norm $\|M\|_{tr}=\sum_i M_{ii}$ for positive-definite matrices

$$\frac{2}{\delta}\mathbf{E}_f\sum_{\lambda\in\Lambda_1}(1-h_\lambda^\star)\xi_\lambda^2\leqslant\tfrac{2}{\delta}\|\Sigma_{\Lambda_1}\|_{tr} \tag{6.12}$$

since $|1-h^\star_\lambda| \leqslant 1$. By (6.11) and (6.12) we obtain

$$\left|\mathbf{E}_f \sum_{\lambda\in\Lambda_1} 2(1-h^\star_\lambda) f_\lambda \xi_\lambda\right| \leqslant \tfrac{2}{\delta}\|\Sigma_{\Lambda_1}\|_{tr} + \tfrac{\delta}{2}\mathbf{E}_f \sum_{\lambda\in\Lambda}(1-h^\star_\lambda) f^2_\lambda. \tag{6.13}$$

In a similar way, we obtain

$$\left|\mathbf{E}_f \sum_{\lambda\in\Lambda_1^{\complement}} 2h^\star_\lambda \xi_\lambda f_\lambda\right| \leqslant \tfrac{\delta}{2}\mathbf{E}_f \sum_{\lambda\in\Lambda} h^\star_\lambda \xi^2_\lambda + \tfrac{2}{\delta}\mathbf{E}_f \sum_{\lambda\in\Lambda_1^{\complement}} h^\star_\lambda f^2_\lambda. \tag{6.14}$$

Note that

$$\tfrac{2}{\delta}\mathbf{E}_f \sum_{\lambda\in\Lambda_1^{\complement}} h^\star_\lambda f^2_\lambda \leqslant \tfrac{2}{\delta}\sum_{\lambda\in\Lambda_1^{\complement}} f^2_\lambda \tag{6.15}$$

since $|h^\star_\lambda| \leqslant 1$. Using (6.14) and (6.15) one has

$$\left|\mathbf{E}_f \sum_{\lambda\in\Lambda_1^{\complement}} 2h^\star_\lambda f_\lambda \xi_\lambda\right| \leqslant \tfrac{2}{\delta}\sum_{\lambda\in\Lambda_1^{\complement}} f^2_\lambda + \tfrac{\delta}{2}\mathbf{E}_f \sum_{\lambda\in\Lambda} h^\star_\lambda \xi^2_\lambda. \tag{6.16}$$

Note also that, since $h_\lambda \in \{0,1\}$, we have

$$\mathbf{E}_f\|\hat f(h^\star) - f\|^2 = \mathbf{E}_f \sum_{\lambda\in\Lambda}(1-h^\star_\lambda) f^2_\lambda + \mathbf{E}_f \sum_{\lambda\in\Lambda} h^\star_\lambda \xi^2_\lambda.$$

Insertion of (6.13) and (6.16) into (6.10) yields

$$\left|\mathbf{E}_f \sum_{\lambda\in\Lambda} 2h^\star_\lambda f_\lambda \xi_\lambda\right| \leqslant \tfrac{\delta}{2}\mathbf{E}_f\|\hat f(h^\star) - f\|^2 + \tfrac{2}{\delta}\|\Sigma_{\Lambda_1}\|_{tr} + \tfrac{2}{\delta}\sum_{\lambda\in\Lambda_1^{\complement}} f^2_\lambda. \tag{6.17}$$

By using the risk hull as in Lemma 2.1, one obtains

$$\mathbf{E}_f\left[\sum_{\lambda\in\Lambda} h^\star_\lambda \xi^2_\lambda - Pen(h^\star)\right] \leqslant \sqrt{2}\min\left(\tfrac{1}{n}, \|\Sigma\|\right). \tag{6.18}$$

Inserting (6.13), (6.16) and (6.18) into (6.8) yields

$$\begin{aligned}\mathbf{E}_f\|\hat f(h^\star) - f\|^2 \leqslant{}& \mathbf{E}_f\left[\min_{h\in\mathcal{H}} \bar R_{pen}(X,h)\right] + \|f\|^2 + \sqrt{2}\min\left(\tfrac{1}{n}, \|\Sigma\|\right)\\ &+ \tfrac{2}{\delta}\sum_{\lambda\in\Lambda}\min(f^2_\lambda, \sigma^2_\lambda) + \sqrt{2}\min\left(\tfrac{1}{n}, \|\Sigma\|\right) + \frac{\delta}{2}\mathbf{E}_f\|\hat f(h^\star) - f\|^2.\end{aligned} \tag{6.19}$$

Using (6.19) we obtain,

$$\begin{aligned}&(1-\tfrac{\delta}{2})\mathbf{E}_f\|\hat f(h^\star) - f\|^2\\ &\leqslant \mathbf{E}_f\left[\min_{h\in\mathcal{H}} \bar R_{pen}(X,h) + \|f\|^2\right] + 2\sqrt{2}\min\left(\tfrac{1}{n}, \|\Sigma\|\right) + \frac{2}{\delta}\sum_{\lambda\in\Lambda}\min(f^2_\lambda, \sigma^2_\lambda).\end{aligned}$$

Finally, we let the bias explicitly appear in

$$\bar{R}_{pen}(X,h) + \|f\|^2 = \sum_{\lambda\in\Lambda}(1-h_\lambda)f_\lambda^2 - \sum_{\lambda\in\Lambda} h_\lambda(X_\lambda^2 - f_\lambda^2) + 2Pen(h)$$

and the result follows from $(1-\frac{\delta}{2})^{-1} \leqslant 1+\delta$ for $\delta \in [0,1]$.

6.3. *Proof of Theorem 3.1*

For $f \in \mathcal{F}_0(\gamma_n)$ the right-hand side in Theorem 2.2 can be bounded by considering the oracle $h^f = \mathbf{1}(\{\lambda : f_\lambda \neq 0\})$ such that

$$(1+\delta)\mathbf{E}_f\left[\Big(-\sum_{\lambda\in\Lambda} h_\lambda^f(X_\lambda^2 - f_\lambda^2) + 2Pen(h^f)\Big)\right] + \Omega_\delta \leqslant (1+\delta)2Pen(h^f) + \Omega_\delta. \tag{6.20}$$

We will use the following inequality, as $J\to\infty$,

$$\sum_{j=1}^{J}(\log(ne/j) + j^{-1}\log_+(n\|\Sigma\|)) \leqslant (J\log(ne/J) + \log(J)\log_+(n\|\Sigma\|))(1+o(1)), \tag{6.21}$$

by comparison with the integral. Since $|h^f| \leqslant n\gamma_n$, we obtain that

$$Pen(h^f) \leqslant 2\|\Sigma_{h^f}\|\left(\sum_{j=1}^{|h^f|}(\log(ne/j) + j^{-1}\log_+(n\|\Sigma\|))\right)$$

$$\leqslant 2\|\Sigma_{h^f}\|\left(n\gamma_n\log(\gamma_n^{-1}) + \log(n\gamma_n)\log_+(n\|\Sigma\|)\right)(1+o(1)),$$

as $n\to\infty$. On the other hand, we have

$$\Omega_\delta = 4\sqrt{2}\min\left(\tfrac{1}{n}, \|\Sigma\|\right) + \frac{2}{\delta}\sum_{\lambda: f_\lambda\neq 0}\min(\sigma_\lambda^2, f_\lambda^2).$$

We use $\sum_{\lambda:f_\lambda\neq 0}\sigma_\lambda^2 \leqslant n\gamma_n\|\Sigma_{h^f}\|$ which shows

$$\Omega_\delta \leqslant \frac{4\sqrt{2}}{n} + \frac{2}{\delta}n\gamma_n\|\Sigma_{h^f}\|.$$

Choosing $\delta\to 0$ such that $\delta^{-1} = o(\log(\gamma_n^{-1}))$, e.g. $\delta = 1/\log\log(\gamma_n^{-1})$, we thus find, as $n\to\infty$,

$$\frac{2}{\delta}n\gamma_n\|\Sigma_{h^f}\| = o\Big(\|\Sigma_{h^f}\|n\gamma_n\log(\gamma_n^{-1})\Big). \tag{6.22}$$

Using Theorem 2.2, Equation (6.22) we have (3.2). Moreover, using the bounds on $\|\Sigma_{h^f}\|$ and $\|\Sigma\|$ we obtain (3.3).

6.4. Proof of Theorem 3.2

Let us now evaluate the right-hand side of the oracle inequality in Theorem 2.2 for the threshold selection rules with arbitrary threshold values $t > 0$ defined in (3.4). Given an oracle parameter $t^0 > 1$ (to be determined below), we set $\tau_\lambda := \sigma_\lambda t^0$. We obtain with R_λ denoting the (inverse) rank of the coefficient with index λ among $(\sigma_\lambda^2 1(|X_\lambda| > \tau_\lambda))_{\lambda\in\Lambda}$

$$\mathbf{E}_f\left[\inf_{h\in\mathcal{H}}\Big(\sum_{\lambda\in\Lambda}(1-h_\lambda)f_\lambda^2 - \sum_{\lambda\in\Lambda}h_\lambda(X_\lambda^2 - f_\lambda^2) + 2Pen(h)\Big)\right] \tag{6.23}$$

$$\leqslant \mathbf{E}_f\Big[\sum_{\lambda\in\Lambda}\Big(\mathbf{1}(|X_\lambda|\leqslant\tau_\lambda)f_\lambda^2 - \mathbf{1}(|X_\lambda|>\tau_\lambda)(X_\lambda^2 - f_\lambda^2) \tag{6.24}$$

$$+ 4\sigma_\lambda^2\mathbf{1}(|X_\lambda|>\tau_\lambda)(\log(en/R_\lambda) + R_\lambda^{-1}\log_+(n\|\Sigma\|))\Big)\Big]. \tag{6.25}$$

Let us first show that $\mathbf{E}_f[\mathbf{1}(|X_\lambda| > \tau_\lambda)(X_\lambda^2 - f_\lambda^2)]$ is always non-negative. By symmetry $X'_\lambda := f_\lambda - \xi_\lambda$ has the same law as X_λ. Defining the function $g(\xi) := \mathbf{1}(|f_\lambda + \xi| > \tau_\lambda)((f_\lambda+\xi)^2 - f_\lambda^2)$, we check by considering the different cases that $g(\xi) + g(-\xi) \geqslant 0$ holds. We conclude

$$\mathbf{E}_f[\mathbf{1}(|X_\lambda|>\tau_\lambda)(X_\lambda^2 - f_\lambda^2)] = \tfrac{1}{2}\mathbf{E}_f[g(\xi_\lambda) + g(-\xi_\lambda)] \geqslant 0.$$

Hence, the term with a minus sign in (6.23) can be discarded for an upper bound.

Let us now consider the coefficients that contain a signal part (i.e. with $f_\lambda \neq 0$). The following inequality will be helpful to obtain a bound independent of the size of $|f_\lambda|$. Let us denote by r_λ^f the corresponding inverse rank within $(\sigma_\lambda^2\mathbf{1}(f_\lambda\neq 0))_{\lambda\in\Lambda}$. With $f_\lambda^2 \leqslant (|\xi_\lambda| + \tau_\lambda)^2$ on the event $\{|X_\lambda| \leqslant \tau_\lambda\}$ we obtain

$$\begin{aligned}
&\sum_{\lambda\in\Lambda, f_\lambda\neq 0}\Big(\mathbf{1}(|X_\lambda|\leqslant\tau_\lambda)f_\lambda^2 + 4\sigma_\lambda^2\mathbf{1}(|X_\lambda|>\tau_\lambda)(\log(en/R_\lambda) + R_\lambda^{-1}\log_+(n\|\Sigma\|))\Big)\\
&\leqslant \sum_{\lambda\in\Lambda, f_\lambda\neq 0}\max\Big((|\xi_\lambda|+\tau_\lambda)^2, 4\sigma_\lambda^2(\log(en/R_\lambda) + R_\lambda^{-1}\log_+(n\|\Sigma\|))\Big)\\
&\leqslant \sum_{\lambda\in\Lambda, f_\lambda\neq 0}\max\Big((|\xi_\lambda|+\tau_\lambda)^2, 4\sigma_\lambda^2(\log(en/r_\lambda^f) + (r_\lambda^f)^{-1}\log_+(n\|\Sigma\|))\Big), \qquad (6.26)
\end{aligned}$$

where for the last inequality we have used that for $n\gamma_n$ distinct values $R_\lambda \in \mathbb{N}$ the expression is maximal in the case $R_\lambda = r_\lambda^f$.

The general identity $\mathbf{E}[\max(Z,c)] = c + \int_c^\infty P(Z \geqslant z)dz$ applied to $Z = (|\xi_\lambda| + \tau_\lambda)^2$ and deterministic $c_\lambda \geqslant \tau_\lambda^2$ yields

$$\begin{aligned}
\mathbf{E}[\max((|\xi_\lambda|+\tau_\lambda)^2, c_\lambda)] &\leqslant c_\lambda + \int_{c_\lambda}^\infty P(|\xi_\lambda| \geqslant \sqrt{z} - \tau_\lambda)\,dz\\
&\leqslant c_\lambda + 2e^{-(\sqrt{c_\lambda}-\tau_\lambda)^2/(2\sigma_\lambda^2)}. \qquad (6.27)
\end{aligned}$$

In order to ensure $\tau_\lambda^2 \leqslant c_\lambda := 4\sigma_\lambda^2(\log(en/r_\lambda^f) + (r_\lambda^f)^{-1}\log_+(n\|\Sigma\|))$ whenever $f_\lambda \neq 0$, we are lead to choose

$$t^0 = \sqrt{4\log(e/\gamma_n)}. \tag{6.28}$$

In the sequel we bound σ_λ^2 simply by $\|\Sigma_{h_f}\|$ in the case $f_\lambda \neq 0$. Then using again the bound on sums of logarithms (6.21) and $\#\{f_\lambda \neq 0\} \leqslant n\gamma_n$ as well as the concavity of e^{-x} for bounding the sum of exponentials, we obtain that (6.23) over the signal part satisfies

$$\begin{aligned}
&\mathbf{E}_f\Big[\sum_{\lambda: f_\lambda \neq 0}\Big(\mathbf{1}(|X_\lambda| \leqslant \tau_\lambda)f_\lambda^2 + 4\sigma_\lambda^2\mathbf{1}(|X_\lambda| > \tau_\lambda)(\log(en/R_\lambda) + R_\lambda^{-1}\log_+(n\|\Sigma\|))\Big)\Big] \\
&\leqslant \sum_{\lambda\in\Lambda, f_\lambda\neq 0}(c_\lambda + 2e^{-(\sqrt{c_\lambda}-\tau_\lambda)^2/(2\sigma_\lambda^2)}) \leqslant n\gamma_n(C_n\|\Sigma_{h_f}\| + 2e^{-(C_n-(t^0)^2)/2}),
\end{aligned}$$

where

$$C_n = (4 + o(1))(\log(\gamma_n^{-1}) + \log_+(n\|\Sigma\|)\log(n\gamma_n)/(n\gamma_n)). \tag{6.29}$$

Owing to $C_n\|\Sigma_{h_f}\| \to \infty$ we even have

$$\begin{aligned}
&\mathbf{E}_f\Big[\sum_{\lambda: f_\lambda \neq 0}\Big(\mathbf{1}(|X_\lambda| \leqslant \tau_\lambda)f_\lambda^2 + 4\sigma_\lambda^2\mathbf{1}(|X_\lambda| > \tau_\lambda)(\log(en/R_\lambda) + R_\lambda^{-1}\log_+(n\|\Sigma\|))\Big)\Big] \\
&\leqslant \|\Sigma_{h_f}\|n\gamma_n C_n(1 + o(1)).
\end{aligned} \tag{6.30}$$

On the other hand, for the non-signal part $f_\lambda = 0$, we introduce $N_\tau := \sum_{\lambda\in\Lambda}\mathbf{1}(|\xi_\lambda| > \tau_\lambda)$ and we use the large deviation bound:

$$\mathbf{E}[N_\tau] = nP(|\xi_\lambda| > \tau_\lambda) \leqslant 2n(t^0)^{-1}e^{-(t^0)^2/2}.$$

Again by considering worst case permutations instead of the ranks, using (6.21) and by Jensen's inequality for the concave functions $\log(x), x\log(en/x)$ we infer:

$$\begin{aligned}
&\mathbf{E}_f\Big[\sum_{\lambda: f_\lambda = 0}\Big(\mathbf{1}(|X_\lambda| \leqslant \tau_\lambda)f_\lambda^2 + 4\sigma_\lambda^2\mathbf{1}(|X_\lambda| > \tau_\lambda)(\log(en/R_\lambda) + R_\lambda^{-1}\log_+(n\|\Sigma\|))\Big)\Big] \\
&\leqslant 4\|\Sigma\|\mathbf{E}_f\Big[\sum_{\lambda\in\Lambda}\mathbf{1}(|\xi_\lambda| > \tau_\lambda)(\log(en/R_\lambda) + R_\lambda^{-1}\log_+(n\|\Sigma\|))\Big] \\
&\leqslant 4\|\Sigma\|\mathbf{E}\Big[\sum_{j=1}^{N_\tau}(\log(en/j) + j^{-1}\log_+(n\|\Sigma\|))\Big] \\
&\leqslant 4\|\Sigma\|\mathbf{E}\Big[(N_\tau\log(en/N_\tau) + \log(N_\tau)\log_+(n\|\Sigma\|))\Big](1 + o(1)) \\
&\leqslant 4\|\Sigma\|(2n(t^0)^{-1}e^{-(t^0)^2/2}(1 + t_0^2/2) + (\log n - (t^0)^2/2)\log_+(n\|\Sigma\|))(1 + o(1)) \\
&\leqslant 2\|\Sigma\|(2ne^{-(t^0)^2/2}t^0 + (2\log n - (t^0)^2)\log_+(n\|\Sigma\|))(1 + o(1)).
\end{aligned} \tag{6.31}$$

For the t^0 chosen, the total bound over (6.23) is thus, by (6.30), (6.31) and by definition of C_n in (6.29),

$$
\begin{aligned}
&n\gamma_n(1+o(1))\Big(\|\Sigma_{h_f}\|C_n + 2\|\Sigma\|(2e^{-(t^0)^2/2}t^0 + (2\log n - (t^0)^2)\frac{\log_+(n\|\Sigma\|)}{n\gamma_n})\Big)\\
&= n\gamma_n(1+o(1))\Big(4\|\Sigma_{h_f}\|(\log(\gamma_n^{-1}) + \log_+(n\|\Sigma\|)\log(n\gamma_n)/(n\gamma_n))\\
&\quad + 2\|\Sigma\|(4\gamma_n\sqrt{\log(\gamma_n^{-1})} + 2\log(n\gamma_n^2)\log_+(n\|\Sigma\|)/(n\gamma_n))\Big).
\end{aligned}
$$

This yields the asserted general bound and inserting the bound for $\log_+(n\|\Sigma\|)$ gives directly the second bound.

6.5. Proof of Theorem 4.1

Consider for each coefficient f_λ the following Bayesian prior, which turns out to be asymptotically least favorable:

$$
\pi_\lambda = (1-\alpha_{\lambda,n})\delta_0 + \alpha_{\lambda,n}\delta_{\mu_{\lambda,n}}, \quad \lambda \in \Lambda,
$$

with some $\mu_{\lambda,n} \geqslant 0$. Without loss of generality we may assume $c_n \downarrow 0$ so slowly that $c_n\sqrt{n\gamma_n} \to \infty$. Introducing the number of non-zero entries $N := \sum_\lambda \mathbf{1}(f_\lambda \neq 0)$ and writing P for the joint law of prior and observations, we deduce by Chebyshev inequality

$$
\begin{aligned}
P(f \notin \mathcal{F}_0(\gamma_n)) &= P(N > n\gamma_n) = P(N - n\gamma_n(1-c_n) > n\gamma_n c_n)\\
&\leqslant \frac{\mathrm{Var}(N)}{(c_n n\gamma_n)^2} \leqslant \frac{n\gamma_n}{(c_n n\gamma_n)^2} \to 0.
\end{aligned}
$$

The property $P(f \in \mathcal{F}_0(\gamma_n)) \to 1$ then implies that the Bayes-optimal risk, derived below, will be an asymptotic minimax lower bound over $\mathcal{F}_0(\gamma_n)$.

We need to calculate the Bayes risk and find the posterior law of $f_\lambda \in \{0, \mu_{\lambda,n}\}$ for each coordinate λ:

$$
P(f_\lambda = \mu_{\lambda,n} | X_\lambda = x) = \frac{\alpha_{\lambda,n}\varphi_{\mu_{\lambda,n},\sigma_\lambda^2}(x)}{(1-\alpha_{\lambda,n})\varphi_{0,\sigma_\lambda^2}(x) + \alpha_{\lambda,n}\varphi_{\mu_{\lambda,n},\sigma_\lambda^2}(x)}.
$$

Since we deal with quadratic loss, the Bayes estimator $\hat{f}_\lambda$ equals the conditional expectation $\mathbf{E}[f_\lambda|X_\lambda]$ and the Bayes risk the expectation of the conditional variance, which is calculated as

$$
\begin{aligned}
\mathbf{E}[\mathrm{Var}(f_\lambda \,|\, X_\lambda)] &= \mathbf{E}[f_\lambda^2] - \mathbf{E}[\mathbf{E}[f_\lambda|X_\lambda]^2]\\
&= \mu_{\lambda,n}^2\Big(\alpha_{\lambda,n} - \int \frac{\alpha_{\lambda,n}^2\varphi_{\mu_{\lambda,n},\sigma_\lambda^2}(x)^2}{(1-\alpha_{\lambda,n})\varphi_{0,\sigma_\lambda^2}(x) + \alpha_{\lambda,n}\varphi_{\mu_{\lambda,n},\sigma_\lambda^2}(x)}\,dx\Big).
\end{aligned}
\tag{6.32}
$$

The integral can be transformed into an expectation with respect to $Z \sim \mathcal{N}(0,1)$ and bounded by Jensen's inequality:

$$\begin{aligned}
&\int \frac{\alpha_{\lambda,n}^2 \varphi_{\mu_{\lambda,n},\sigma_\lambda^2}(x)^2}{(1-\alpha_{\lambda,n})\varphi_{0,\sigma_\lambda^2}(x) + \alpha_{\lambda,n}\varphi_{\mu_{\lambda,n},\sigma_\lambda^2}(x)}\,dx \\
&\quad = \alpha_{\lambda,n}\,\mathbf{E}\left[\Big(1+\alpha_{\lambda,n}^{-1}(1-\alpha_{\lambda,n})\exp(\sigma_\lambda^{-1}Z - \mu_{\lambda,n}^2/(2\sigma_\lambda^2))\Big)^{-1}\right] \\
&\quad \leqslant \alpha_{\lambda,n}\Big(1+\alpha_{\lambda,n}^{-1}(1-\alpha_{\lambda,n})\,\mathbf{E}[\exp(\sigma_\lambda^{-1}Z - \mu_{\lambda,n}^2/(2\sigma_\lambda^2))]\Big)^{-1} \\
&\quad = \alpha_{\lambda,n}\Big(1+\alpha_{\lambda,n}^{-1}(1-\alpha_{\lambda,n})\exp((1-\mu_{\lambda,n}^2)/(2\sigma_\lambda^2))\Big)^{-1}.
\end{aligned}$$

Since $\alpha_{\lambda,n} \to 0$ uniformly by the choice of $c_n \to 0$, we just select

$$\mu_{\lambda,n} = \sigma_\lambda \sqrt{2(1-(\log c_n^{-1})^{-1/2})\log(\alpha_{\lambda,n}^{-1})}$$

such that $\mathbf{E}[\mathrm{Var}(f_\lambda \,|\, X_\lambda)]$ is larger or equal to

$$2\sigma_\lambda^2\alpha_{\lambda,n}(1-(\log c_n^{-1})^{-1/2})\log(\alpha_{\lambda,n}^{-1})(1-((1+(1-\alpha_{\lambda,n})\alpha_{\lambda,n}^{-(\log c_n^{-1})^{-1/2}} e^{1/(2\sigma_\lambda^2)}))^{-1}).$$

Noting $\alpha_{\lambda,n}^{-(\log c_n^{-1})^{-1/2}} \to \infty$ uniformly over λ, the overall Bayes risk is hence uniformly lower bounded by

$$2\big(1+o(1)\big)\Big(\sum_{\lambda\in\Lambda}\sigma_\lambda^2\alpha_{\lambda,n}\log(\alpha_{\lambda,n}^{-1})\Big).$$

The supremum at n is attained for

$$\alpha_{\lambda,n} = \exp\Big(\frac{\bar\sigma_n^2}{\sigma_\lambda^2}\log(e\gamma_n(1-c_n)) - 1\Big) = e^{-1}(e\gamma_n(1-c_n))^{\bar\sigma_n^2/\sigma_\lambda^2},$$

where $\bar\sigma_n > 0$ is such that $\sum_\lambda \alpha_{\lambda,n} = n\gamma_n(1-c_n)$ holds, provided $\alpha_{\lambda,n} \leqslant c_n$ for all λ. The latter condition is fulfilled if $\bar\sigma_n^2 \gtrsim \max_\lambda \sigma_\lambda^2$.

6.6. Proof of Corollary 4.2

Let us write $\alpha_{\lambda,n} = n\gamma_n(1-c_n)w_{\lambda,n}$ and the entropy expression in Theorem 4.1 becomes

$$2\big(1+o(1)\big)n\gamma_n \sup_{w_{\lambda,n}}\Big(\sum_{\lambda\in\Lambda}\sigma_\lambda^2 w_{\lambda,n}\big(\log(w_{n,\lambda}^{-1}) - \log(n\gamma_n)\big)\Big)$$

where the $w_{\lambda,n} \in [0,(n\gamma_n(1-c_n))^{-1}]$ sum up to one: $\sum_\lambda w_{\lambda,n} = 1$. From this representation we immediately infer the lower bound

$$2n\gamma_n\log(\gamma_n^{-1})\big(1+o(1)\big)\frac{1}{n}\sum_{\lambda\in\Lambda}\sigma_\lambda^2$$

using the uniform weights $w_{\lambda,n} = 1/n$.

Note that for polynomial growth $\sigma^2_{(i)} \sim (n-i)^\beta$, $\beta > 0$, and for $r_n = o(n)$, we have $\sigma^2_{(r_n)}/\sigma^2_{(1)} \to 1$ and the lower bound is indeed

$$\sup_{f \in \mathcal{F}_0(\gamma_n)} \mathbf{E}_f[\|\hat{f}_n - f\|^2] \geqslant 2\big(1+o(1)\big)\|\Sigma\| n\gamma_n \log(\gamma_n^{-1}).$$

6.7. Proof of Theorem 4.3

Introduce the threshold value $\tau_{\lambda,n} = \sqrt{2\log(\alpha^{-1}_{\lambda,n})}$ and note $\max_\lambda \alpha_{\lambda,n} \to 0$. We can split the error as follows:

$$\mathbf{E}[(\hat{f}_\lambda - f_\lambda)^2] = f_\lambda^2\, \mathbb{P}((\xi_\lambda + f_\lambda/\sigma_\lambda)^2 \leqslant \tau^2_{\lambda,n}) + \mathbf{E}[\sigma^2_\lambda \xi^2_\lambda \mathbf{1}_{\{(\xi_\lambda + f_\lambda/\sigma_\lambda)^2 > \tau^2_{\lambda,n}\}}] =: I + II.$$

For $f_\lambda > \tau_{\lambda,n}\sigma_\lambda$ term I is estimated by

$$I \leqslant f_\lambda^2\, \mathbb{P}(\xi_\lambda \leqslant \tau_{\lambda,n} - f_\lambda/\sigma_\lambda) \leqslant f_\lambda^2 \exp(-(\tau_{\lambda,n} - f_\lambda/\sigma_\lambda)^2/2).$$

Together with a symmetric argument for $f_\lambda < -\tau_{\lambda,n}\sigma_\lambda$ and a direct bound for $f_\lambda^2 \leqslant \tau^2_{\lambda,n}\sigma^2_\lambda$, we thus obtain a bound for general f_λ:

$$I \leqslant \big(f_\lambda^2 \exp(-(\tau_{\lambda,n} - |f_\lambda|/\sigma_\lambda)^2/2)\big) \vee \tau^2_{\lambda,n}\sigma^2_\lambda.$$

Since for $\tau_{\lambda,n} \to \infty$ we have $\sup_{x \geqslant 1} x^2 e^{-\tau^2_{\lambda,n}(x-1)^2/2} \to 1$, we consider $x = |f_\lambda|/(\tau_{\lambda,n}\sigma_\lambda)$ and infer

$$I \leqslant \sigma^2_\lambda \tau^2_{\lambda,n}(1+o(1)) \text{ uniformly in } \lambda.$$

Inserting the choice of the thresholds, we conclude

$$I \leqslant \sigma^2_\lambda \tau^2_{\lambda,n}(1+o(1))\mathbf{1}_{\{f_\lambda \neq 0\}} = 2\sigma^2_\lambda \log(\alpha^{-1}_{\lambda,n})(1+o(1))\mathbf{1}_{\{f_\lambda \neq 0\}}.$$

For term II and $f_\lambda \neq 0$ the immediate estimate $II \leqslant \sigma^2_\lambda$ suffices, while for $f_\lambda = 0$ we integrate out explicitly and obtain:

$$II = \sigma^2_\lambda\, \mathbf{E}[\xi^2_\lambda \mathbf{1}_{\{\xi^2_\lambda > \tau^2_{\lambda,n}\}}] = \sigma^2_\lambda 2(\tau_{\lambda,n}+1)e^{-\tau^2_{\lambda,n}/2} = 2\sigma^2_\lambda \sqrt{2\log(\alpha^{-1}_{\lambda,n})}\alpha_{\lambda,n}(1+\tau^{-1}_{\lambda,n}).$$

The overall risk of our estimator is therefore bounded by

$$\begin{aligned}
&\sum_{\lambda \in \Lambda} \mathbf{E}_f[(\hat{f}_\lambda - f_\lambda)^2] \\
&\leqslant \sum_{\lambda: f_\lambda \neq 0} \Big(2\sigma^2_\lambda \log(\alpha^{-1}_{\lambda,n})(1+o(1)) + \sigma^2_\lambda\Big) + \sum_{\lambda: f_\lambda = 0} 2\sigma^2_\lambda \sqrt{2\log(\alpha^{-1}_{\lambda,n})}\alpha_{\lambda,n}(1+o(1)) \\
&\leqslant (2+o(1))\Big(\sum_{\lambda: f_\lambda \neq 0} \log(\alpha^{-1}_{\lambda,n})\sigma^2_\lambda + \sqrt{2}\max_\lambda \frac{\alpha_{\lambda,n}}{\sqrt{\log(\alpha^{-1}_{\lambda,n})}} \sum_{\lambda: f_\lambda = 0} \sigma^2_\lambda \log(\alpha^{-1}_{\lambda,n})\Big).
\end{aligned}$$

Choosing $\alpha_{\lambda,n} = e^{-\beta_n/\sigma_\lambda^2}$, with $\beta_n > 0$ satisfying $\sum_{\lambda\in\Lambda}\alpha_{\lambda,n} = n\gamma_n$, minimises the last bound (asymptotically) and yields

$$\sum_{\lambda\in\Lambda}\mathbf{E}_f[(\hat{f}_\lambda - f_\lambda)^2] \leqslant (2+o(1))n\gamma_n\beta_n$$

because by $\max_\lambda(\log(\alpha_{\lambda,n}^{-1}))^{-1/2}\alpha_{\lambda,n} \to 0$ the second term is of smaller order. The last result is a direct consequence. Indeed, we always have $\beta_n \leqslant \log(\gamma_n^{-1})\|\Sigma\|$ by bounding $\sigma_\lambda^2 \leqslant \|\Sigma\|$, which is minimax optimal for at most polynomial growth in (σ_λ^2) by the lower bound in Theorem 4.1.

References

Abramovich F., Benjamini Y., Donoho D.L. and Johnstone I.M. (2006). Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist.* **34**, 584–653. MR2281879

Abramovich F. and Silverman B.W. (1998). Wavelet decomposition approaches to statistical inverse problems. *Biometrika* **85**, 115–129. MR1627226

Akaike H. (1973). *Information theory and an extension of the maximum likelihood principle.* Proc. 2nd Intern. Symp. Inf. Theory, Petrov P.N. and Csaki F. eds. Budapest, 267–281. MR0483125

Benjamini Y. and Hochberg Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Royal Stat. Soc. Ser. B* **57**, 289–300. MR1325392

Birgé L. and Massart P. (2001). Gaussian model selection. *J. Eur. Math. Soc.* **3** 203–268. MR1848946

Cavalier L. (2004). Estimation in a problem of fractional integration. *Inverse Problems* **20**, 1–10. MR2109128

Cavalier L. (2011). Inverse problems in statistics. Inverse problems and high-dimensional estimation, *Lecture Notes in Statistics*, Springer. MR2868199

Cavalier L. and Golubev Yu. (2006). Risk hull method and regularization by projections of ill-posed inverse problems. *Ann. Statist.* **34**, 1653–1677. MR2283712

Cavalier L., Golubev G.K., Picard D. and Tsybakov A.B. (2002). Oracle inequalities in inverse problems. *Ann. Statist.* **30**, 843–874. MR1922543

Cavalier L. and Raimondo M. (2007). Wavelet deconvolution with noisy eigenvalues. *IEEE Trans. Signal Proc.* **55**, 2414–2424. MR1500172

Cohen A., Hoffmann M. and Reiss M. (2004). Adaptive wavelet Galerkin method for linear inverse problems. *SIAM J. Numer. Anal.* **42**, 1479–1501. MR2114287

Comte F. and Renault E. (1996). Long memory continuous time models. *Journal of Econometrics* **73**, 101–149. MR1410003

Donoho D.L. (1995). Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition. *Appl. and Comput. Harmon. Anal.* **2**, 101–126. MR1325535

Golubev Y. (2002). Reconstruction of sparse vectors in white Gaussian noise. *Probl. Inf. Transm.* **1** 65–79. MR2101314

Golubev Y. (2011). On oracle inequalities related to data-driven hard thresholding. *Probab. Theory Related Fields* **150**, 435–469. MR2824863

Hoffmann M. and Reiss M. (2008). Nonlinear estimation for linear inverse problems with error in the operator. *Ann. Statist.* **36**, 310–336. MR2387973

Johnstone I.M. (2011). *Gaussian estimation: Sequence and wavelets models.* Book to appear.

Johnstone I.M. and D. Paul (2013). Adaptation in a class of linear inverse problems. arXiv:1310.7149.

Johnstone I.M. and Silverman B.W. (1997). Wavelet threshold estimators for data with correlated noise. *J. Royal Stat. Soc. Ser. B* **59**, 300–351. MR1440585

Massart P. (2007). *Concentration inequalities and model selection.* Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003. Lecture Notes in Mathematics, Springer, Berlin. MR2319879

Rochet P. (2013). Adaptive hard-thresholding for linear inverse problems. To appear in *ESAIM*. MR3070888

Sowell F. (1990). *The fractional unit root distribution.* Econometrica **58**, 495–505. MR1046932

Wang Y. (1996). Function estimation via wavelet shrinkage for long-memory data. *Annals of Statist.* **24**, 466–484. MR1394972

Wu Z. and Zhou H.H. (2013). Model selection and sharp asymptotic minimaxity. *Probab. Theory Related Fields* **156**, 193–227. MR3055256

Zygmund A. (1959). *Trigonometric series.* 2nd ed. Vols. I, II. Cambridge University Press, New York. MR0107776