Biometrika (2013), **100**, 2, *pp.* 355–370
doi: 10.1093/biomet/ass082
Advance Access publication 3 March 2013

Estimation of a sparse group of sparse vectors

BY FELIX ABRAMOVICH

Department of Statistics and Operations Research, Tel Aviv University, Tel Aviv 69978, Israel
felix@post.tau.ac.il

AND VADIM GRINSHTEIN

Department of Mathematics and Computer Science, The Open University of Israel, Raanana 43537, Israel
vadimg@openu.ac.il

SUMMARY

We consider estimating a sparse group of sparse normal mean vectors, based on penalized likelihood estimation with complexity penalties on the number of nonzero mean vectors and the numbers of their significant components, which can be performed by a fast algorithm. The resulting estimators are developed within a Bayesian framework and can be viewed as maximum a posteriori estimators. We establish their adaptive minimaxity over a wide range of sparse and dense settings. A simulation study demonstrates the efficiency of the proposed approach, which successfully competes with the sparse group lasso estimator.

Some key words: Adaptive minimaxity; Complexity penalty; Maximum a posteriori rule; Sparsity; Thresholding.

1. INTRODUCTION

Suppose we observe m independent n-dimensional Gaussian vectors $y_1, \ldots, y_m$ with independent components and common variance:

$$y_j = \mu_j + \epsilon_j \quad (j = 1, \ldots, m), \tag{1}$$

where $\epsilon_j \sim \mathcal{N}_n(0, \sigma_n^2 I_n)$ and are independent. The variance $\sigma_n^2 > 0$, which may depend on n, is assumed to be known, and the goal is to estimate the unknown mean vectors $\mu_1, \ldots, \mu_m$.

The key extra assumption on (1) is sparsity both within and between vectors. Hereafter we will refer to them as within- and between-sparsity for brevity. More specifically, between-sparsity assumes that some of the vectors μ_j are identically zero vectors and the entire information in the noisy data is contained only in a small fraction of them, while within-sparsity means that even within nonzero vectors μ_j, most of their components are zero or at least negligible. Neither the indices of nonzero vectors μ_j nor the locations of their significant components are known in advance.

Such a model appears in a variety of statistical applications.

Example 1. Consider a regression model $y_i = f(x_{1i}, \ldots, x_{m_i}) + \epsilon_i$ $(i = 1, \ldots, n)$, where $f : \mathbb{R}^m \to \mathbb{R}$ is the unknown regression function assumed to belong to some class of functions, e.g., Hólder, Sobolev or Besov classes, and the ϵ_i are independent $\mathcal{N}(0, \sigma_n^2)$ variates. Estimating f in such a general set-up suffers from a severe curse of dimensionality, where typically the

sample size n should grow exponentially with the dimension m to achieve consistent estimation. It is then essential to place extra restrictions on the complexity of f. One of the most common approaches is to consider the additive models, where $f(x_1, \ldots, x_m) = f_1(x_1) + \cdots + f_m(x_m)$ and each component f_j lies in some smoothness class. In addition, similar to sparse linear regression models, it is often reasonable to assume that only some of the predictors among the $x_1, \ldots, x_m$ are really significant, while the impact of others is negligible. Such sparse additive models are especially relevant for $m \sim n$ and $m \gg n$ set-ups and have been considered in Lin & Zhang (2006), Meier et al. (2009), Ravikumer et al. (2009) and Raksutti et al. (2012).

Expand each f_j into univariate orthonormal series (ψ_{ij}) as $\sum \mu_{ij}\psi_{ij}(x_j)$, where $\mu_{ij} = \int f_j(x_j)\psi_{ij}(x_j)\,\mathrm{d}x_j$. The original nonparametric additive model is then transformed into the equivalent problem (1) of estimating vectors of corresponding coefficients $\mu_1, \ldots, \mu_m$ within Gaussian noise, where for sparse additive models, most of the vectors μ_j are zero. Moreover, for properly chosen bases (ψ_{ij}), e.g., Fourier series for Sobolev, or wavelets for more general Besov classes, the nonzero μ_j will also be sparse.

Example 2. In time-course microarray experiments the data consist of measurements of differences in the expression levels between treated and control samples of m genes recorded at different times. A record on the jth gene at time-point t_i is modelled as a measurement of an unknown expression profile function $f_j(t)$ at time t_i, corrupted by Gaussian noise. The expressions of most genes are the same in both groups, that is, $f_j \equiv 0$, and the goal is to identify the differentially expressed genes and estimate the corresponding profile functions f_j. Similarly to the previous example, each f_j is commonly expanded into some parsimonious orthonormal basis, e.g., Legendre polynomials, harmonic functions or wavelets, as $f_j(t) = \sum_i \mu_{ij}\psi_{ij}(t)$, and in the coefficient domain the functional model becomes

$$y_{ij} = \mu_{ij} + z_{ij} \quad (j = 1, \ldots, m;\ i = 1, \ldots, n),$$

where the y_{ij} are empirical coefficients of the data on the jth gene and the z_{ij} are Gaussian variates (see, e.g., Angelini et al., 2007). For most genes, $\mu_j = 0$, while due to the parsimony of the chosen basis, for differentially expressed genes, the μ_j will still have a sparse representation.

To estimate $\mu_1, \ldots, \mu_m$ in (1) under the assumptions of between- and within-sparsity we proceed as follows. From Donoho & Johnstone (1994a, b), it is known that the optimal strategy for estimating a single sparse vector μ_j from y_j is thresholding. Various threshold estimators $\hat{\mu}_j$ can be considered as penalized likelihood estimators, where

$$\hat{\mu}_j = \arg\min_{\tilde{\mu}_j \in \mathbb{R}^n} \|y_j - \tilde{\mu}_j\|_2^2 + \mathrm{P}_j(\tilde{\mu}_j),$$

corresponding to different choices of penalties $\mathrm{P}_j(\tilde{\mu})$. In particular, the l_1 penalty $\mathrm{P}_j(\tilde{\mu}_j) = \lambda\|\tilde{\mu}_j\|_1$ leads to soft thresholding of components of $\tilde{\mu}_j$ with the constant threshold $\lambda/2$ that coincides with the lasso estimator of Tibshirani (1996). Wider classes of penalties on the magnitudes of components $\tilde{\mu}_{ij}$ are discussed in Antoniadis & Fan (2001). In this paper we consider l_0 or complexity penalties $\mathrm{P}_j(\|\tilde{\mu}_j\|_0)$ on the number of nonzero components $\tilde{\mu}_{ij}$, where $\|\tilde{\mu}_j\|_0 = \#\{i : \tilde{\mu}_{ij} \neq 0\}$, which yield hard thresholding rules. In the simplest case, where $\mathrm{P}_j(\|\tilde{\mu}_j\|_0) = \lambda\|\tilde{\mu}_j\|_0$, the resulting constant threshold is $\sqrt{\lambda}$. More general complexity penalties were studied in Birgé & Massart (2001), Abramovich et al. (2007, 2010) and Wu & Zhou (2013).

Penalizing each $\tilde{\mu}_j$ separately, however, ignores the between-sparsity, where it is assumed that most of the μ_j are identically zero and should be estimated by $\hat{\mu}_j = 0$. Thus, simultaneous estimation of all m mean vectors in model (1) should involve an additional penalty $\mathrm{P}_0(\cdot)$ on the

number of nonzero $\hat{\mu}_j$. The estimators $\hat{\mu}_j$ are then defined as solutions of

$$\min_{\tilde{\mu}_1,\ldots,\tilde{\mu}_m \in \mathbb{R}^n} \left[\sum_{j=1}^{m} \{\|y_j - \tilde{\mu}_j\|_2^2 + \mathrm{P}_j(\|\tilde{\mu}_j\|_0)\} + \mathrm{P}_0(k) \right], \tag{2}$$

where $k = |\{j : \tilde{\mu}_j \neq 0\}|$. In this paper we investigate the optimality of such an approach for estimating $\mu_1, \ldots, \mu_m$ under various within- and between-sparsity set-ups. In particular, we specify the classes of complexity penalties $\mathrm{P}_j(\|\tilde{\mu}_j\|_0)$ and $\mathrm{P}_0(k)$ on within- and between-sparsity respectively for which the resulting estimators $\hat{\mu}_1, \ldots, \hat{\mu}_m$ achieve asymptotically minimax rates simultaneously for a wide range of sparse and dense cases. Such types of penalties naturally arise within a Bayesian model selection framework. In this sense, this paper extends the results of the Bayesian maximum a posteriori approach of Abramovich et al. (2007, 2010) to simultaneous estimation of a group of m vectors in model (1).

It is interesting to compare the proposed complexity penalization (2) with lasso-type procedures. Similar to l_0-type penalization, the vector-wise use of the original lasso of Tibshirani (1996) for estimating each μ_j in (1) results in per-component soft thresholding within each y_j that handles within-sparsity but ignores between-sparsity. To address the latter, Yuan & Lin (2006) proposed a group lasso that for model (1) solves

$$\min_{\tilde{\mu}_1,\ldots,\tilde{\mu}_m \in \mathbb{R}^n} \sum_{j=1}^{m} (\|y_j - \tilde{\mu}_j\|_2^2 + \lambda \|\tilde{\mu}_j\|_2).$$

It can be shown that in such a set-up, the group lasso estimator is available in closed form, namely,

$$\hat{\mu}_j = y_j \left(1 - \frac{\lambda/2}{\|y_j\|_2}\right)_+ \quad (j = 1, \ldots, m),$$

which is vector-level shrink-or-kill thresholding with the threshold $\lambda/2$. The $\hat{\mu}_j$ are, therefore, either entirely zero or do not have zero components at all. As a result, the group lasso does not handle within-sparsity. To combine both types of sparsity, Friedman, Hastie & Tibshirani in an unpublished 2010 Stanford University manuscript entitled 'A note on the group lasso and a sparse group lasso' introduced the sparse group lasso that for model (1) is defined as

$$\min_{\tilde{\mu}_1,\ldots,\tilde{\mu}_m \in \mathbb{R}^n} \sum_{j=1}^{m} (\|y_j - \tilde{\mu}_j\|_2^2 + \lambda_1 \|\tilde{\mu}_j\|_2 + \lambda_2 \|\tilde{\mu}_j\|_1) \tag{3}$$

yielding

$$\hat{\mu}_j = \tilde{y}_j \left(1 - \frac{\lambda_1/2}{\|\tilde{y}_j\|_2}\right)_+ \quad (j = 1, \ldots, m),$$

where the $\tilde{y}_{ij} = \mathrm{sign}(y_{ij})(|y_{ij}| - \lambda_2/2)_+$ $(i = 1, \ldots, n)$ is the result of component-level soft thresholding of each y_j with the threshold $\lambda_2/2$.

To the best of our knowledge, there are no theoretical results on optimality of the sparse group lasso similar to those presented here for the complexity penalized estimators (2). Moreover, we believe that, generally, l_0-type penalties are more natural for representing sparsity and the main reasons for other types of penalties are mostly computational. For a general regression model, complexity penalties imply a combinatorial search over all possible models, while, for example, the sparse group lasso estimator can be efficiently computed by numerical iterative algorithms

(Simon et al., 2013). However, for model (1), which can be viewed as a special case of a general regression set-up, (2) can also be solved by fast algorithms.

2. BAYESIAN SPARSE GROUP MAXIMUM A POSTERIORI ESTIMATION

Consider again model (1). If we knew the indices of the nonzero vectors μ_j and the locations of their significant entries μ_{ij}, we would estimate them by the corresponding y_{ij} and set the others to zero. Hence, the original problem is reduced to finding an $n \times m$ indicator matrix D, where d_{ij} indicates whether μ_{ij} is significant or not, and can be viewed as a model selection problem. Due to between- and within-sparsity assumptions, the matrix D should be sparse in the double sense: only some of the columns of D are nonzero, and even nonzero columns are sparse.

We first introduce some notation. Let $\mathcal{J}_0$ and $\mathcal{J}_0^c$ be the sets of indices corresponding respectively to the zero and nonzero μ_j, and let $m_0 = |\mathcal{J}_0^c| = |\{j : \mu_j \neq 0,\ j = 1, \ldots, m\}|$. Let $h_j = \sum_{i=1}^n d_{ij} = |\{i : \mu_{ij} \neq 0,\ i = 1, \ldots, n\}|$ denote the number of nonzero components in μ_j, where evidently $h_j = 0$ for $j \in \mathcal{J}_0$.

Consider the following Bayesian model selection procedure for identifying the nonzero components μ_{ij} or, equivalently, the indicator matrix D. To capture the between- and within-sparsity assumptions we place a hierarchical prior on D. We first assume some prior distribution on the number of nonzero mean vectors, $m_0 \sim \pi_0(m_0) > 0$ $(m_0 = 0, \ldots, m)$. For a given m_0, assume that all different configurations of the zero and nonzero μ_j are equally likely, that is, conditionally on m_0,

$$\text{pr}\left(\mathcal{J}_0^c \mid |\mathcal{J}_0^c| = m_0\right) = \binom{m}{m_0}^{-1}.$$

Obviously, $h_j \mid (j \in \mathcal{J}_0) \sim \delta(0)$ and, thus, $d_j \mid (j \in \mathcal{J}_0) \sim \delta(0)$ and $\mu_j \mid (j \in \mathcal{J}_0) \sim \delta(0)$. For the nonzero μ_j we place independent priors $\pi_j(\cdot)$ on the number of their nonzero components, that is, $h_j \mid (j \in \mathcal{J}_0^c) \sim \pi_j(h_j) > 0$ $(h_j = 1, \ldots, n)$. In this case, we again assume that for a given h_j, all possible indicator vectors d_j with h_j nonzero components have the same prior probabilities and, therefore,

$$\text{pr}\left(d_j \mid \|d_j\|_0 = h_j, j \in \mathcal{J}_0^c\right) = \binom{n}{h_j}^{-1}.$$

Finally, to complete the prior for model (1), we have $\mu_{ij} \mid \{d_{ij} = 0\} \sim \delta(0)$, while the nonzero μ_{ij} are assumed to be independent identically distributed $\mathcal{N}(0, \gamma\sigma_n^2)$ variates, where $\gamma > 0$.

The posterior probability for a given indicator matrix D is therefore

$$\text{pr}(D \mid y) \propto \pi_0(m_0)\binom{m}{m_0}^{-1} \prod_{j \in \mathcal{J}_0^c} \left\{ \pi_j(h_j)\binom{n}{h_j}^{-1} (1+\gamma)^{-h_j/2} \exp\left(\frac{\gamma}{\gamma+1}\frac{\sum_{i=1}^n y_{ij}^2 d_{ij}}{2\sigma_n^2}\right)\right\}.$$

Given the posterior distribution $\text{pr}(D \mid y)$ we apply the maximum a posteriori rule to choose the most likely configuration of zero and nonzero μ_{ij}, yielding the following criterion:

$$\max_D \left(\sum_{j \in \mathcal{J}_0^c} \left[\sum_{i=1}^n y_{ij}^2 d_{ij} + 2\sigma_n^2(1+1/\gamma) \log\left\{ \pi_j(h_j)\binom{n}{h_j}^{-1} (1+\gamma)^{-h_j/2}\right\}\right] + 2\sigma_n^2(1+1/\gamma)\log\left\{\pi_0(m_0)\binom{m}{m_0}^{-1}\right\}\right). \quad (4)$$

From (4) it follows that for a given $h_j > 0$ the optimal choice $\hat{d}_j(h_j)$ for d_j is $\hat{d}_{ij}(h_j) = 1$ for the h_j largest $|y_{ij}|$ and zero otherwise. Criterion (4) then reduces to

$$\max_D \left(\sum_{j \in \mathcal{J}_0^c} \left[\sum_{i=1}^{h_j} y_{(i)j}^2 + 2\sigma_n^2(1+1/\gamma) \log \left\{ \pi_j(h_j) \binom{n}{h_j}^{-1} (1+\gamma)^{-h_j/2} \right\} \right] + 2\sigma_n^2(1+1/\gamma) \log \left\{ \pi_0(m_0) \binom{m}{m_0}^{-1} \right\} \right), \tag{5}$$

where $|y_{(1)j}| \geqslant \cdots \geqslant |y_{(n)j}|$. For each $j = 1, \ldots, m$ define

$$\begin{aligned} \hat{h}_j &= \arg \min_{1 \leqslant h_j \leqslant n} \left[\sum_{i=h_j+1}^{n} y_{(i)j}^2 + 2\sigma_n^2(1+1/\gamma) \log \left\{ \pi_j^{-1}(h_j) \binom{n}{h_j} (1+\gamma)^{h_j/2} \right\} \right] \\ &= \arg \min_{1 \leqslant h_j \leqslant n} \left[-\sum_{i=1}^{h_j} y_{(i)j}^2 + 2\sigma_n^2(1+1/\gamma) \log \left\{ \pi_j^{-1}(h_j) \binom{n}{h_j} (1+\gamma)^{h_j/2} \right\} \right]. \end{aligned} \tag{6}$$

Then, (5) is equivalent to minimizing

$$\sum_{j \in \mathcal{J}_0^c} \left[-\sum_{i=1}^{\hat{h}_j} y_{(i)j}^2 + 2\sigma_n^2(1+1/\gamma) \log \left\{ \pi_j^{-1}(\hat{h}_j) \binom{n}{\hat{h}_j} (1+\gamma)^{\hat{h}_j/2} \right\} \right] + 2\sigma_n^2(1+1/\gamma) \log \left\{ \pi_0^{-1}(m_0) \binom{m}{m_0} \right\} \tag{7}$$

over all subsets of indices $\mathcal{J}_0 \subseteq \{1, \ldots, m\}$. Define

$$W_j = -\sum_{i=1}^{\hat{h}_j} y_{(i)j}^2 + 2\sigma_n^2(1+1/\gamma) \log \left\{ \pi_j^{-1}(\hat{h}_j) \binom{n}{\hat{h}_j} (1+\gamma)^{\hat{h}_j/2} \right\}. \tag{8}$$

Then, (7) reduces to

$$\min_{0 \leqslant m_0 \leqslant m} \left[\sum_{j=1}^{m_0} W_{(j)} + 2\sigma_n^2(1+1/\gamma) \log \left\{ \pi_0^{-1}(m_0) \binom{m}{m_0} \right\} \right], \tag{9}$$

where $W_{(1)} \leqslant \cdots \leqslant W_{(m)}$, and the sums $\sum_{j \in \mathcal{J}_0^c}$ in (7) and $\sum_{j=1}^{m_0}$ in (9) do not appear for $m_0 = 0$.

The resulting algorithm for finding the proposed sparse group maximum a posteriori estimators of $\mu_1, \ldots, \mu_m$ in (1) is as follows:

Algorithm 1. Sparse group maximum a posteriori estimation algorithm.

Step 1. For $j = 1$ to m, find $\hat{h}_j$ in (6) and calculate the corresponding W_j in (8). Order W_j in ascending order $W_{(1)} \leqslant \cdots \leqslant W_{(m)}$ and find

$$\hat{m}_0 = \arg \min_{0 \leqslant m_0 \leqslant m} \left[\sum_{j=1}^{m_0} W_{(j)} + 2\sigma_n^2 (1 + 1/\gamma) \log \left\{ \pi_0^{-1}(m_0) \binom{m}{m_0} \right\} \right].$$

Step 2. Let $\hat{\mathcal{J}}_0^c$ be the set of indices corresponding to the $\hat{m}_0$ smallest W_j. Set $\hat{\mu}_j = 0$ for all $j \in \hat{\mathcal{J}}_0$. For $j \in \hat{\mathcal{J}}_0^c$, take the $\hat{h}_j$ largest $|y_{ij}|$ and threshold the others, that is, $\hat{\mu}_{ij} = y_{ij} \mathbb{I}\{|y_{ij}| \geqslant |y_{(\hat{h}_j)j}|\}$ $(i = 1, \ldots, n)$, where $|y_{(1)j}| \geqslant \cdots \geqslant |y_{(n)j}|$.

The resulting estimation procedure therefore combines vector-wise and componentwise thresholding. It is easily verified that the minimizer of (7) is, in fact, the penalized likelihood estimator (2) with the complexity penalties

$$\mathrm{P}_j(0) = 0, \mathrm{P}_j(h_j) = 2\sigma_n^2 (1 + 1/\gamma) \log \left\{ \pi_j^{-1}(h_j) \binom{n}{h_j} (1 + \gamma)^{h_j/2} \right\} \quad (h_j = 1, \ldots, m) \tag{10}$$

and

$$\mathrm{P}_0(m_0) = 2\sigma_n^2 (1 + 1/\gamma) \log \left\{ \pi_0^{-1}(m_0) \binom{m}{m_0} \right\} \quad (m_0 = 0, \ldots, m). \tag{11}$$

The penalties $\mathrm{P}_j(\cdot)$ and $\mathrm{P}_0(\cdot)$ depend on the priors $\pi_j(\cdot)$ and $\pi_0(\cdot)$. For example, binomial priors $m_0 \sim B(m, \xi_0)$ and $h_j \sim B(n, \xi_j)$ yield linear type penalties $\mathrm{P}_0(m_0) = 2\sigma_n^2 \lambda_0^2 m_0$ and $\mathrm{P}_j(h_j) = 2\sigma_n^2 \lambda_j^2 h_j$ respectively, where $\lambda_0^2 = (1 + 1/\gamma) \log\{(1 - \xi_0)/\xi_0\}$ and $\lambda_j^2 = (1 + 1/\gamma) \log\{(1 + \gamma)^{1/2}(1 - \xi_j)/\xi_j\}$. For such a choice of $\pi_j(\cdot)$, W_j in (8) is obtained by hard thresholding of the y_j with the constant threshold $\sqrt{2}\sigma_n \lambda_j$. In particular, $\xi_j = (\gamma + 1)^{1/2}/\{(\gamma + 1)^{1/2} + n^{\gamma/(\gamma+1)}\}$ leads to the universal thresholding of Donoho & Johnstone (1994a) with $\lambda_j = \sqrt{\log n}$. The truncated geometric priors $\pi_j(h_j) \propto q_j^{h_j}$ $(h_j = 1, \ldots, n)$, for some $0 < q_j < 1$, imply the nonlinear so-called $2k \log(n/k)$-type penalties. The optimality of the resulting hard thresholding estimator with a data-driven threshold for estimating a single normal mean vector has been shown in Abramovich et al. (2007, 2010) and Wu & Zhou (2013).

3. ADAPTIVE MINIMAXITY OF SPARSE GROUP MAXIMUM A POSTERIORI ESTIMATORS

In this section we investigate the goodness of the proposed sparse group maximum a posteriori estimators (2) with the penalties (10)–(11), where the goodness of fit is measured by the global quadratic risk $\sum_{j=1}^m E(\|\hat{\mu}_j - \mu_j\|_2^2)$. We establish their asymptotic minimaxity over a wide range of sparse and dense settings. To derive these results we need the following assumption on the priors $\pi_j(\cdot)$.

Assumption 1. Assume that

$$\pi_j(h) \leqslant \binom{n}{h} e^{-c(\gamma)h} \quad (h = 1, \ldots, n;\ j = 1, \ldots, m), \tag{12}$$

where $c(\gamma) = 8(\gamma + 3/4)^2 > 9/2$.

Assumption 1 is not restrictive. Indeed, the obvious inequality

$$\binom{n}{h} \geqslant \left(\frac{n}{h}\right)^h$$

implies that for any $\pi_j(\cdot)$, (12) holds for all $h \leqslant ne^{-c(\gamma)}$. In particular, Assumption 1 is satisfied for binomial priors $B(n, \xi_j)$ with $\xi_j \leqslant e^{-c(\gamma)}/(1 + e^{-c(\gamma)})$ and truncated geometric priors.

First, we obtain a general upper bound for the quadratic risk of the sparse group maximum a posteriori estimator that will be the key for deriving its asymptotic minimaxity.

THEOREM 1. *Consider model* (1). *Let* $\hat{\mu}_1, \ldots, \hat{\mu}_m$ *be the sparse group maximum a posteriori estimators* (2) *of* $\mu_1, \ldots, \mu_m$ *with the complexity penalties* (10)–(11). *Under Assumption* 1 *we have*

$$\sum_{j=1}^{m} E\left(\|\hat{\mu}_j - \mu_j\|_2^2\right) \leqslant c_1(\gamma) \min_{\mathcal{J}_0 \subseteq \{1,\ldots,m\}} \left[\sum_{j \in \mathcal{J}_0^c} \min_{1 \leqslant h_j \leqslant n} \left\{ \sum_{i=h_j+1}^{n} \mu_{(i)j}^2 + \mathrm{P}_j(h_j) \right\} + \sum_{j \in \mathcal{J}_0} \sum_{i=1}^{n} \mu_{ij}^2 + \mathrm{P}_0(|\mathcal{J}_0^c|) \right] + c_2(\gamma)\sigma_n^2\{1 - \pi_0(0)\}, \tag{13}$$

where $|\mu_{(1)j}| \geqslant \cdots \geqslant |\mu_{(n)j}|$ *and* $c_1(\gamma)$, $c_2(\gamma)$ *depend only on* γ.

Theorem 1 holds for any normal mean vectors $\mu_1, \ldots, \mu_m$. Now we consider model (1) under the extra within- and between-sparsity assumptions defined rigorously below.

Between-sparsity is measured by the number m_0 of the nonzero μ_j. Within-sparsity can be introduced in several ways. An obvious measure for a single normal mean vector $\mu \in \mathbb{R}^n$ is the number of its nonzero components, that is, its l_0 quasi-norm $\|\mu\|_0$. Define an l_0-ball $l_0(\eta)$ of standardized radius η as a set of μ with at most a proportion η of nonzero entries, that is

$$l_0(\eta) = \{\mu \in \mathbb{R}^n : \|\mu\|_0 \leqslant \eta n\}.$$

One can argue that in many practical settings, it is more reasonable to assume that the components μ_i of μ are not exactly zero but are small. In a wider sense the within-sparsity of μ can then be defined by the proportion of its large entries. Formally, define a weak l_p-ball $m_p(\eta)$ with a standardized radius η as

$$m_p(\eta) = \left\{\mu \in \mathbb{R}^n : |\mu|_{(i)} \leqslant \sigma_n \eta (n/i)^{1/p}, \;\; i = 1, \ldots, n\right\},$$

where $\mu_{(1)} \geqslant \cdots \geqslant \mu_{(n)}$ are the ordered components of μ. For $\mu \in m_p(\eta)$, the proportion of the $|\mu_i|$ larger than $\sigma_n \delta$ for some $\delta > 0$ is at most $(\eta/\delta)^p$.

Within-sparsity can also be measured in terms of the l_p-norm of μ, where a strong l_p-ball $l_p(\eta)$ with a standardized radius η is defined as

$$l_p(\eta) = \left\{\mu \in \mathbb{R}^n : \frac{1}{n}\sum_{i=1}^{n} |\mu_i|^p \leqslant \sigma_n^p \eta^p\right\}.$$

Table 1. *Minimax rates over various $l_0(\eta_n)$, $l_p(\eta_n)$, and $m_p(\eta_n)$-balls. The rates are the same for $l_p(\eta_n)$ and $m_p(\eta_n)$ except for $p=2$, where for $m_p(\eta_n)$ there appears an additional log term not presented for brevity*

Case	$p=0$	$0<p<2$	$p\geqslant 2$
Dense	$\sigma_n^2 n$	$\sigma_n^2 n$	$\sigma_n^2 n$
Sparse	$\sigma_n^2 n\eta_n(\log \eta_n^{-1})$	$\sigma_n^2 n\eta_n^p(\log \eta_n^{-p})^{1-p/2}$	$\sigma_n^2 n\eta_n^2$
Super-sparse	–	$\sigma_n^2 n^{2/p}\eta_n^2$	$\sigma_n^2 n\eta_n^2$

There are well-known relationships between these types of balls. The l_p-norm approaches l_0 as p decreases, while a weak l_p-ball contains the corresponding strong l_p-ball, but only just:

$$l_p(\eta)\subset m_p(\eta)\not\subset l_{p'}(\eta)\quad (p'>p). \tag{14}$$

First we recall the known results on minimax rates for estimating a single normal mean vector μ over different types of balls introduced above. Let $\Theta(\eta_n)\subset\mathbb{R}^n$ be any of $l_0(\eta_n)$, $l_p(\eta_n)$ or $m_p(\eta_n)$, where the standardized radius η might depend on n. The corresponding minimax quadratic risk for estimating a single μ over $\Theta(\eta_n)$ in model (1) is $R\{\Theta(\eta_n)\}=\inf_{\tilde\mu}\sup_{\mu\in\Theta(\eta_n)}E(\|\tilde\mu-\mu\|_2^2)$, where the infimum is taken over all estimates $\tilde\mu$ of μ. For $p>0$ define $\eta_{0n}=n^{-1/\min(p,2)}\sqrt{\log n}$. Depending on the behaviour of η_n as n increases, we distinguish between three cases for $p>0$ and two cases for $p=0$:

(a) dense, where $\eta_n\not\to 0$ for $p\geqslant 0$;
(b) sparse, where $\eta_n\to 0$, $\eta_n/\eta_{0n}\not\to 0$ for $p>0$ and $\eta_n\geqslant n^{-1}$ for $p=0$;
(c) super-sparse, where $\eta_n/\eta_{0n}\to 0$ for $p>0$.

The corresponding minimax convergence rates $R\{\Theta(\eta_n)\}$ for various cases and p are summarized in Table 1 (Donoho et al., 1992; Johnstone, 1994; Donoho & Johnstone, 1994b). The rates for $m_p(\eta_n)$ are the same as for $l_p(\eta_n)$ except for $p=2$, where there is an additional log term. Table 1 defines dense and sparse zones for $p=0$ and $p\geqslant 2$, and dense, sparse, and super-sparse zones for $0<p<2$ of different minimax rates.

Consider now model (1) for $m\geqslant 1$. Recall that $m_0=|\{j:\mu_j\neq 0\}|$ and $\mathcal{J}_0^c$ is the set of indices for the nonzero μ_j. In what follows we assume that $\mu_j\in\Theta_j(\eta_{jn})$ for $j\in\mathcal{J}_0^c$, where the types and the parameters p of the corresponding balls are not necessarily the same for all j. Furthermore, we allow the priors $\pi_0(\cdot)$ and $\pi_j(\cdot)$ to depend respectively on m and n.

Theorem 2 below defines the upper bounds for the quadratic risks of the sparse group maximum a posteriori estimator in model (1) under within- and between-sparsity assumptions.

THEOREM 2. *Consider model (1), where $\mathcal{J}_0^c\neq\emptyset$. Assume that $\mu_j\in\Theta_j(\eta_{jn})$ for all $j\in\mathcal{J}_0^c$, where $\eta_{jn}\geqslant n^{-1/\min(p_j,2)}\sqrt{\log n}$ for all $p_j>0$, thus excluding super-sparse cases.*

Let $\hat\mu_1,\ldots,\hat\mu_m$ be the sparse group maximum a posteriori estimators (2) with the complexity penalties (10)–(11), where we assume that there exist constants $c_0,c_1>0$ and $c_2>c(\gamma)$ such that

1. *$\pi_0(k)\geqslant(k/m)^{c_0k}$ $(k=1,\ldots,\lfloor m/e\rfloor)$ and $\pi_0(m)\geqslant e^{-c_0m}$;*
2. *for all $j=1,\ldots,m$, the $\pi_j(\cdot)$ satisfies Assumption 1 and, in addition, $\pi_j(h)\geqslant(h/n)^{c_1h}$ $(h=1,\ldots,\lfloor ne^{-c(\gamma)}\rfloor)$; $\pi_j(n)\geqslant e^{-c_2n}$.*

Then, for any $\mathcal{J}_0^c \subseteq \{1, \ldots, m\}$ *with* $|\mathcal{J}_0^c| = m_0$ *and all* $\Theta_j(\eta_{jn})$, $j \in \mathcal{J}_0^c$,

$$\sup_{\mu_j \in \Theta_j(\eta_{jn}), j \in \mathcal{J}_0^c} \sum_{j=1}^{m} E\left(\|\hat{\mu}_j - \mu_j\|_2^2\right) \leqslant C_1(\gamma) \max\left[\sum_{j \in \mathcal{J}_0^c} R\{\Theta_j(\eta_{jn})\}, \sigma_n^2 m_0 \log(m/m_0)\right] \tag{15}$$

for some constant $C_1(\gamma)$ *depending only on* γ, *where up to multiplying constants, the corresponding* $R\{\Theta_j(\eta_n)\}$ *are given in Table 1.*

Theorem 2 shows that as both m and n increase, the asymptotic convergence rates in (15) are either of order $\sum_{j \in \mathcal{J}_0^c} R\{\Theta_j(\eta_{jn})\}$ or $\sigma_n^2 m_0 \log(m/m_0)$. The former is associated with the optimal rates of estimating m_0 single sparse vectors in $\Theta_j(\eta_{jn})$ ($j \in \mathcal{J}_0^c$), while the latter appears in the optimal rates in the model selection and corresponds to the error of selecting a subset of m_0 nonzero elements out of m (Abramovich & Grinshtein, 2010; Raskutti et al., 2011; Rigollet & Tsybakov, 2011). From Table 1 it follows that for all within-dense and within-sparse cases, $C_1\sigma_n^2 \log n \leqslant R\{\Theta_j(\eta_{jn})\} \leqslant C_2\sigma_n^2 n$ ($j \in \mathcal{J}_0^c$) for some $C_1,\ C_2 > 0$ and, therefore, the first term $\sum_{j \in \mathcal{J}_0^c} R\{\Theta_j(\eta_n)\}$ in the upper bound (15) is always dominating for $m_0/m > n^{-1}$, while the second term $\sigma_n^2 m_0 \log(m/m_0)$ is necessarily the main one for $m_0/m < e^{-n}$.

One can verify that the conditions on the priors $\pi_0(\cdot)$ and $\pi_j(\cdot)$ required in Theorem 2 are satisfied, for the truncated geometric priors from § 2. On the other hand, no binomial priors $\pi_0 = B(m, \xi_0)$ or $\pi_j = B(n, \xi_j)$ can satisfy all of them: the requirement $\pi_j(n) = \xi_j^n \geqslant e^{-c_2 n}$ yields $\xi_j \geqslant e^{-c_2}$, while one needs $\xi_j \to 0$ as n increases to have $\pi_j(1) = n\xi_j(1-\xi_j)^{n-1} \geqslant n^{-c_1}$.

To establish the corresponding lower bound for the minimax risk, for simplicity of exposition we consider only the two cases where the p_j for $j \in \mathcal{J}_0^c$ are either all zeros or all positive. These are the two main scenarios appearing in various set-ups. Similar results for minimax lower bounds in the particular context of sparse nonparametric additive models appear in Raksutti et al. (2012).

THEOREM 3. *Consider model* (1), *where* $\mu_j \in l_0(\eta_{jn})$ *for* $j \in \mathcal{J}_0^c$. *Assume that* $|\mathcal{J}_0^c| = m_0 > 0$. *Then there exists a constant* $C_2 > 0$ *such that*

$$\inf_{\tilde{\mu}_1, \ldots, \tilde{\mu}_m} \sup_{\mu_j \in l_0(\eta_{jn}), j \in \mathcal{J}_0^c} \sum_{j=1}^{m} E\left(\|\tilde{\mu}_j - \mu_j\|_2^2\right) \geqslant C_2 \max\left[\sum_{j \in \mathcal{J}_0^c} R\{l_0(\eta_{jn})\}, \sigma_n^2 m_0 \log(m/m_0)\right], \tag{16}$$

where the infimum is taken over all estimators $\tilde{\mu}_1, \ldots, \tilde{\mu}_m$ *of* $\mu_1, \ldots, \mu_m$.

Theorem 3 shows that, as m and n increase, the rates in (15) cannot be improved for l_0-balls. The proposed sparse group maximum a posteriori estimator in this case is, therefore, adaptive to the unknown degrees of within- and between-sparsity and is simultaneously minimax rate-optimal over the entire range of dense and sparse l_0-ball settings.

The analysis of the case $p_j > 0$ is slightly more delicate. Note first that due to the embedding properties of l_p-balls for $p > 0$ in (14), it is sufficient to establish the minimax lower bounds for strong l_p-ball settings.

THEOREM 4. *Consider model* (1), *where* $\mu_j \in l_{p_j}(\eta_{jn})$ *for* $j \in \mathcal{J}_0^c$ *and* $|\mathcal{J}_0^c| = m_0 > 0$. *In addition, assume that* $\eta_{jn}^2 \geqslant n^{-2/\min(p_j, 2)} \max\{\log n, \log(m/m_0)\}$. *Under this additional*

constraint, there exists a constant $C_2 > 0$ *such that*

$$\inf_{\tilde{\mu}_1,\ldots,\tilde{\mu}_m} \sup_{\mu_j \in l_{p_j}(\eta_{jn}), j \in \mathcal{J}_0^c} \sum_{j=1}^{m} E\left(\|\tilde{\mu}_j - \mu_j\|_2^2\right) \geqslant C_2 \max\left[\sum_{j \in \mathcal{J}_0^c} R\{l_{p_j}(\eta_{jn})\}, \sigma_n^2 m_0 \log(m/m_0)\right], \tag{17}$$

where the infimum is taken over all estimators $\tilde{\mu}_1, \ldots, \tilde{\mu}_m$ *of* $\mu_1, \ldots, \mu_m$.

Similar to Theorem 3, Theorem 4 implies simultaneous minimaxity of the sparse group maximum a posteriori estimator over strong and weak l_p-balls as both m and n increase but with the restriction on η_{jn} and m_0. In particular, it does not cover settings with within-super-sparsity but depending on m_0 might also exclude part of the corresponding within-sparse zone. Within- and between-sparsity cannot be simultaneously strong. In fact, the condition $\eta_{jn}^2 < n^{-2/\min(p_j,2)} \max\{\log n, \log(m/m_0)\}$ for $j \in \mathcal{J}_0^c$ can be viewed as an extended definition of super-sparsity for $m > 1$. For such a super-sparse case, the minimax bound (17) does not hold and can be reduced. Indeed, consider the trivial zero estimators $\tilde{\mu}_j = 0$ $(j = 1, \ldots, m)$, where, evidently,

$$\sup_{\mu_j \in l_{p_j}(\eta_{jn}), j \in \mathcal{J}_0^c} \sum_{j=1}^{m} E\left(\|\tilde{\mu}_j - \mu_j\|_2^2\right) = \sup_{\mu_j \in l_{p_j}(\eta_{jn}), j \in \mathcal{J}_0^c} \sum_{j \in \mathcal{J}_0^c} \|\mu_j\|_2^2. \tag{18}$$

The least favourable sequences that maximize $\|\mu_j\|_2^2$ over $l_{p_j}(\eta_{jn})$ are $(\sigma_n \eta_{jn}, \ldots, \sigma_n \eta_{jn})^{\mathrm{T}}$ and $(\sigma_n \eta_{jn} n^{1/p_j}, 0, \ldots, 0)^{\mathrm{T}}$ for $p_j \geqslant 2$ and $0 < p_j < 2$ respectively. Thus, $\sup_{\mu_j \in l_{p_j}(\eta_{jn})} \|\mu_j\|_2^2 = \sigma_n^2 \eta_{jn}^2 n^{2/\min(p_j,2)}$ and the right-hand side of (18) is less than $\sigma_n^2 m_0 \log(m/m_0)$ for $\eta_{jn}^2 < n^{-2/\min(p_j,2)} \log(m/m_0)$, $j \in \mathcal{J}_0^c$. This goes along the lines of the corresponding results for estimating a single normal mean vector, where a zero estimator is known to be rate-optimal for the super-sparse case (Donoho & Johnstone, 1994b).

4. SIMULATION STUDY

We generated data according to model (1) with $m = 10$, each vector μ_j of length $n = 100$. Five of the μ_j were identically zero, while the other five had respectively 100, 70, 50, 20, and 5 nonzero components randomly sampled from $\mathcal{N}(0, \tau^2)$ $(\tau = 1, 3, 5)$. Such a set-up covers various types of within-sparsity. Finally, independent standard Gaussian noise was added to all components of each μ_j. The corresponding variance ratios $\gamma = \tau^2/\sigma^2$ are therefore 1, 9, and 25.

We tried binomial and truncated geometric priors for sparse group maximum a posteriori estimators. For the binomial prior, we performed componentwise universal hard thresholding of Donoho & Johnstone (1994a) with the threshold $\lambda = \sigma(2 \log n)^{1/2}$ within each vector that essentially corresponds to $\xi_j = (\gamma + 1)^{1/2}/\{(\gamma + 1)^{1/2} + n^{\gamma/(\gamma+1)}\}$ and used $\xi_0 = 1/m$. For the truncated geometric prior we set $q_0 = q_j = 0{\cdot}3$. In addition, we compared the performances of sparse group maximum a posteriori estimators with the sparse group lasso estimator (3) from the unpublished 2010 manuscript of Friedman, Hastie & Tibshirani mentioned in § 1. They do not discuss the optimal choices for λ_1 and λ_2 in (3). Some heuristic arguments are given in Simon et al. (2013). In our simulation study we instead considered two oracle-based choices for these tuning parameters, thus giving a significant advantage to sparse group lasso estimators. Since in simulation examples we know the true mean vectors μ_j, they can be used for choosing λ_1 and λ_2 to minimize the mean squared error. In particular, we considered a semi-oracle sparse

Table 2. *Scaled mean squared errors, MSE/τ^2, averaged over* 1000 *replications for four sparse group estimators and various* τ

τ	Sparse group MAP (binomial)	Sparse group MAP (geometric)	Sparse group lasso (semi-oracle)	Sparse group lasso (fully oracle)
1	247·4	245·5	236·9	161·9
3	67·6	42·1	124·6	44·9
5	22·0	14·1	63·8	19·0

group lasso estimator, where we set $\lambda_2 = 2\sigma(2\log n)^{1/2}$ yielding universal soft thresholding within each vector to compare the sparse group lasso with the binomial sparse group maximum a posteriori estimator. The threshold λ_1 was chosen by minimizing the mean squared error $\sum_{j=1}^{m} E(\|\hat{\mu}_j(\lambda_1) - \mu_j\|_2^2)$ estimated by averaging over a series of 1000 replications for each value of λ_1 using a grid search. In addition, we applied a fully oracle sparse group lasso estimator, where both λ_1 and λ_2 were chosen to minimize the mean squared error by the two-dimensional grid. This can be considered as a benchmark for the performance of the sparse group lasso.

The resulting choices for the fully oracle sparse group lasso estimator were $\lambda_1 = 11{\cdot}8$, $\lambda_2 = 0{\cdot}9$ for $\tau = 1$; $\lambda_1 = 7{\cdot}2$, $\lambda_2 = 1{\cdot}1$ for $\tau = 3$; and $\lambda_1 = 4{\cdot}7$, $\lambda_2 = 1{\cdot}3$ for $\tau = 5$. For all τ, the oracle choice for λ_2 in the sparse group lasso is much smaller than the conservative universal threshold $2\sigma(2\log n)^{1/2} \approx 6{\cdot}06$. The oracle thresholding within each vector is thus significantly less severe and keeps many more coefficients. The oracle choices for λ_1 are also quite small and, as a result, for any τ, no single vector was thresholded by the fully oracle sparse group lasso estimator; all the $\hat{\mu}_j$ were nonzero. Thus, it was not really a between-sparse estimator for the considered set-up.

The parameter $\gamma = \tau^2/\sigma^2$ can be essentially viewed as a signal-to-noise ratio within individual vectors. Thus, to demonstrate the behaviour of estimators as n increases we compared them for increasing values of τ keeping n and σ fixed. The resulting mean squared errors should then be normalized by the corresponding values of τ^2 to be on the same scale. In Table 2 we present the resulting scaled mean squared errors averaged over 1000 replications for the four sparse group estimators for $\tau = 1, 3$ and 5.

For all estimators the scaled mean squared errors decrease as τ increases but at different rates. The semi-oracle sparse group lasso estimator has the slowest rate, while both sparse group maximum a posteriori estimators converge faster and successfully compete even with the fully oracle sparse group lasso estimator.

Small τ corresponds to a sparse setting, where only a few of the largest nonzero components can be distinguished from the noise. This explains the good performance of the binomial sparse group maximum a posteriori and the semi-oracle sparse group lasso estimators based respectively on universal hard and soft thresholding within each vector in this case. For larger τ, universal thresholding becomes over-conservative. The negative effect of its conservativeness is much stronger for the soft than for the hard version. The fully oracle sparse group lasso estimator strongly outperforms its semi-oracle counterpart especially for $\tau = 3$ and $\tau = 5$, also indicating that the universal thresholding is far from being optimal for the sparse group lasso, especially for moderate and large τ. See also our previous comments on the optimal choice of λ_2.

On the other hand, the geometric sparse group maximum a posteriori estimator corresponding to a nonlinear $2k\log(n/k)$-type penalty provides good results for all τ, following the theoretical results of § 3. Moreover, for $\tau = 3$ and $\tau = 5$, it outperforms even the fully oracle sparse group lasso estimator that was thought of as a benchmark rather than a fair competitor. This indicates that the sparse group lasso faces general problems. This is unsurprising since soft shrink-or-kill

thresholding inherent for the sparse group lasso is well known to be superior to hard keep-or-kill thresholding in sparse group maximum a posteriori estimation for small coefficients, but to be inferior for large ones due to the additional shrinkage. Moreover, the sparse group lasso in (3) involves a double amount of shrinkage: both within vectors and at each entire vector as a whole. It thus causes unnecessary extra bias that increases with τ, which outweighs the benefits of variance reduction. A similar phenomenon appears also for naïve elastic set estimation (Zou & Hastie, 2005).

We also analysed the performance of the four estimators for each individual μ_j. See the Supplementary Material for the results. The sparser the true mean vector, the better it was estimated by all methods. The main contribution to the overall errors always came from estimating dense μ_j. However, the general tendencies discussed above hold uniformly across all vectors. Thus, for all μ_j the fully oracle sparse group lasso estimator is much better than its semi-oracle counterpart for all τ; the binomial sparse group maximum a posteriori estimator and especially the semi-oracle sparse group lasso estimator perform worse for larger τ; and the geometric sparse group maximum a posteriori estimator provides good results for all τ and uniformly outperforms even the fully oracle sparse group lasso estimator for large τ. Simulation results indicate that a possible way to improve the performance of a sparse group lasso would be to consider different thresholds λ_{2j} and possibly λ_{1j} in (3).

In addition, we compared the four methods for estimating the vector supports even though this is a different problem from our original goal of estimating vectors in the l_2-norm. Indeed, for minimizing a quadratic risk it may be worthwhile to threshold small nonzero coefficients instead of paying a price in terms of the variance for their estimation. Being conservative, the semi-oracle sparse group lasso and the binomial sparse group maximum a posteriori estimators thresholded too many nonzero coefficients. As already mentioned, the thresholds λ_1 and λ_2 in the fully-oracle sparse group lasso were too small for a proper recovery of supports of sparse vectors and all the $\hat{\mu}_j$ were nonzero even for zero μ_j. The geometric sparse group maximum a posteriori estimator provided similar results with its binomial counterpart and the semi-oracle sparse group lasso estimator for sparse vectors but strongly outperformed them for dense cases. The results are given in the Supplementary Material.

ACKNOWLEDGEMENT

Both authors were supported by the Israel Science Foundation. We are grateful to Ofir Harari for his assistance in running simulation examples and Saharon Rosset for fruitful discussions.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online includes some additional results of the simulation study. The R-code used for the simulation study is available from the first author.

APPENDIX

Below we use C to denote a generic positive constant, not necessarily the same each time it is used, even within a single equation. Similarly, $C(\gamma)$ is a generic positive constant depending on γ.

Proof of Theorem 1. The proposed sparse group maximum a posteriori estimator can be viewed as a penalized likelihood estimator (2) with the complexity penalties (10) and (11). We first rewrite it in a different form that will allow us then to apply the general results of Birgé & Massart (2001) for complexity penalized estimators.

Let $y = (y_{11}, \ldots, y_{n1}, \ldots, y_{1m}, \ldots, y_{nm})^{\mathrm{T}}$ be an amalgamated $nm \times 1$ vector of data. Similarly, $\mu = (\mu_{11}, \ldots, \mu_{n1}, \ldots, \mu_{1m}, \ldots, \mu_{nm})^{\mathrm{T}}$, $\epsilon = (\epsilon_{11}, \ldots, \epsilon_{n1}, \ldots, \epsilon_{1m}, \ldots, \epsilon_{nm})^{\mathrm{T}}$ and the original model (1) can be rewritten as

$$y_i = \mu_i + \epsilon_i \quad (i = 1, \ldots, nm), \tag{A1}$$

where the $\epsilon_i \sim \mathcal{N}(0, \sigma_n^2)$ are independent. Define an indicator vector d, where $d_i = \mathbb{I}\{\mu_i \neq 0\}$ $(i = 1, \ldots, nm)$. In terms of the model (A1), $h_j = \sum_{i=n(j-1)+1}^{nj} d_i$ $(j = 1, \ldots, m)$ and $m_0 = \#\{j : h_j > 0\}$. For a given d, define $D_d = \sum_{j=1}^{m} h_j = \#\{i : d_i = 1,\ i = 1, \ldots, nm\}$ and

$$L_d = \frac{1}{D_d}\left[\sum_{j=1}^{m} \log\left\{\pi_j^{-1}(h_j)\binom{n}{h_j}\right\} + \log\left\{\pi_0^{-1}(m_0)\binom{m}{m_0}\right\}\right]$$

for $d \neq 0$ and $L_0 = 2\log \pi_0^{-1}(0)$, where we formally set $\pi_j(0) = 1$. Then, the sparse group maximum a posteriori estimator $\hat{\mu} = (\hat{\mu}_{11}, \ldots, \hat{\mu}_{n1}, \ldots, \hat{\mu}_{1m}, \ldots, \hat{\mu}_{nm})^{\mathrm{T}}$ is the penalized likelihood estimator of μ with the complexity penalty

$$\begin{aligned}\mathrm{P}(d) &= 2\sigma_n^2(1 + 1/\gamma)\left[\sum_{j=1}^{m} \log\left\{\pi_j^{-1}(h_j)\binom{n}{h_j}(1+\gamma)^{h_j/2}\right\} + \log\left\{\pi_0^{-1}(m_0)\binom{m}{m_0}\right\}\right] \\ &= \sigma_n^2(1 + 1/\gamma)D_d\{2L_d + \log(1+\gamma)\}\end{aligned}$$

for $d \neq 0$ and $\mathrm{P}(0) = \sigma_n^2(1 + 1/\gamma)L_0$.

One can verify that

$$\sum_{d \neq 0} e^{-D_d L_d} = \sum_{k=1}^{m} \pi_0(k) = 1 - \pi_0(0).$$

Straightforward computation similar to that in the proof of Theorem 1 of Abramovich et al. (2007) implies also that for any d under Assumption 1,

$$(1 + 1/\gamma)\{2L_{\mathbf{d}} + \log(1+\gamma)\} \geqslant C(\gamma)\{1 + (2L_d)^{1/2}\}^2,$$

where $C(\gamma) > 1$. One can then apply Theorem 2 of Birgé & Massart (2001) to get

$$\begin{aligned}\sum_{j=1}^{m} E\left(\|\hat{\mu}_j - \mu_j\|_2^2\right) \leqslant c_1(\gamma) \min_{\mathcal{J}_0 \subseteq \{1,\ldots,m\}} &\left[\sum_{j \in \mathcal{J}_0^c} \min_{1 \leqslant h_j \leqslant n}\left\{\sum_{i=h_j+1}^{n} \mu_{(i)j}^2 + \mathrm{P}_j(h_j)\right\}\right. \\ &\left. + \sum_{j \in \mathcal{J}_0}\sum_{i=1}^{n} \mu_{ij}^2 + \mathrm{P}_0(m_0)\right] + c_2(\gamma)\sigma_n^2\{1 - \pi_0(0)\}\end{aligned} \tag{A2}$$

and the proof is complete. □

Proof of Theorem 2. One can check from Table 1 that for $\eta_{jn} \geqslant n^{-1/\min(p_j,2)}\sqrt{\log n}$ if $p_j > 0$ and for $\eta_{jn} > n^{-1}$ if $p_j = 0$, the last term $c_2(\gamma)\sigma_n^2\{1 - \pi_0(0)\}$ in the right-hand side of (13) is of order $O(\sigma_n^2) = o[R\{\Theta_j(\eta_{jn})\}]$ for all nonzero μ_j.

Let $\mathcal{J}_0^{c*}$ be the true unknown subset of nonzero vectors μ_j and $m_0^* = |\mathcal{J}_0^{c*}|$.

I. Consider $m_0^* \leqslant \lfloor m/e \rfloor$. Apply Theorem 1 for $\mathcal{J}_0 = \mathcal{J}_0^*$ to get

$$\sum_{j=1}^{m} E\left(\|\hat{\mu}_j - \mu_j\|_2^2\right) \leqslant c_1(\gamma)\left(\sum_{j\in\mathcal{J}_0^{c*}} \min_{1\leqslant h_j\leqslant n}\left[\sum_{i=h_j+1}^{n} \mu_{(i)j}^2 + 2\sigma_n^2(1+1/\gamma)\log\left\{\pi_j^{-1}(h_j)\binom{n}{h_j}(1+\gamma)^{h_j/2}\right\}\right]\right.$$
$$\left. + 2\sigma_n^2(1+1/\gamma)\log\left\{\pi_0^{-1}(m_0)\binom{m}{m_0}\right\}\right) + c_2(\gamma)\sigma_n^2\{1-\pi_0(0)\}.$$

Since for $m_0 = 1, \ldots, \lfloor m/e \rfloor$,

$$\binom{m}{m_0} \leqslant \left(\frac{m}{m_0}\right)^{2m_0}$$

(Lemma A1 of Abramovich et al., 2010), the required conditions on $\pi_0(\cdot)$ ensure that

$$2\sigma_n^2(1+1/\gamma)\log\left\{\pi_0^{-1}(m_0)\binom{m}{m_0}\right\} \leqslant C(\gamma)\sigma_n^2 m_0 \log(m/m_0).$$

To complete the proof for this case we now consider separately

$$\min_{1\leqslant h_j\leqslant n}\left[\sum_{i=h_j+1}^{n} \mu_{(i)j}^2 + 2\sigma_n^2(1+1/\gamma)\log\left\{\pi_j^{-1}(h_j)\binom{n}{h_j}(1+\gamma)^{h_j/2}\right\}\right] \tag{A3}$$

for each $j \in \mathcal{J}_0^{c*}$ and show that it is $O[R\{\Theta_j(\eta_{jn})\}]$, where $R\{\Theta_j(\eta_{jn})\}$ are given in Table 1. We distinguish between several cases, where the proofs for strong l_p-balls will follow immediately from the proofs for the corresponding weak l_p-balls due to the embedding properties mentioned in § 3.

Case 1. Let $\mu_j \in \Theta_j(\eta_{jn})$, $\eta_{jn} > e^{-c(\gamma)}$ for $p_j = 0$, and $\eta_{jn}^{p_j} > e^{-c(\gamma)}$ for $p_j > 0$. Taking $h_j^* = n$, under the condition on $\pi_j(n)$, implies that (A3) is $O(\sigma_n^2 n) = O[R\{\Theta_j(\eta_{jn})\}]$.

Case 2. Let $\mu_j \in l_0(\eta_{jn})$, $\eta_{jn} \leqslant e^{-c(\gamma)}$. Since $\mu_j \neq 0$, $\eta_{jn} \geqslant n^{-1}$. Choose $h_j^* = n\eta_{jn}$ and repeat the arguments in the proof of Theorem 3 of Abramovich et al. (2007) using a slightly more general version of Lemma A1 of Abramovich et al. (2010) for approximating the binomial coefficient in (A3) instead of their original Lemma A1.

Case 3. Let $\mu_j \in m_{p_j}(\eta_{jn})$, $0 < p_j < 2$, and $n^{-1}(\log n)^{p_j/2} \leqslant \eta_{jn}^{p_j} \leqslant e^{-c(\gamma)}$. Take $1 \leqslant h_j^* = n\eta_{jn}^{p_j}(\log \eta_{jn}^{-p_j})^{-p_j/2} \leqslant ne^{-c(\gamma)}$ and follow the proof of Theorem 4 of Abramovich et al. (2007) with a more general version of Lemma A1, see Case 2.

Case 4. Let $\mu_j \in m_{p_j}(\eta_{jn})$, $p_j \geqslant 2$ and $n^{-p_j/2}(\log n)^{p_j/2} \leqslant \eta_{jn}^{p_j} \leqslant e^{-c(\gamma)}$. Take $h_j^* = 1$. Then, for $p_j > 2$,

$$\sum_{i=h_j^*+1}^{n} \mu_{(i)j}^2 < \sigma_n^2 n^{2/p_j}\eta_{jn}^2 \int_1^n x^{-2/p_j}\,\mathrm{d}x < \frac{p_j}{p_j-2}\sigma_n^2 n^{2/p_j}\eta_{jn}^2 n^{1-2/p_j} = O(\sigma_n^2 n\eta_{jn}^2)$$

and, similarly, for $p_j = 2$,

$$\sum_{i=h_j^*+1}^{n} \mu_{(i)j}^2 < \sigma_n^2 n\eta_{jn}^2 \int_1^n x^{-1}\,\mathrm{d}x = \sigma_n^2 n\eta_{jn}^2 \log n.$$

On the other hand, under the conditions on $\pi_j(\cdot)$, $\pi_j(1) \geqslant n^{-c_1}$, which yields

$$2\sigma_n^2(1+1/\gamma)\log\left\{\pi_j^{-1}(1)n(1+\gamma)^{\frac{1}{2}}\right\} = O(\sigma_n^2\log n) = O(\sigma_n^2 n\eta_{jn}^2)$$

for $\eta_{jn} \geqslant (n^{-1}\log n)^{1/2}$.

II. Consider $\lfloor m/e \rfloor < m_0^* \leqslant m$. Applying Theorem 1 for $\mathcal{J}_0^c = \{1, \ldots, m\}$ or, equivalently, for $\mathcal{J}_0 = \emptyset$ and $h_j = 1$ for $j \in \mathcal{J}_0^*$, yields

$$\sum_{j=1}^{m} E\left(\|\hat{\mu}_j - \mu_j\|_2^2\right) \leqslant c_1(\gamma)\left[\sum_{j \in \mathcal{J}_0^{c*}} \min_{1 \leqslant h_j \leqslant n}\left\{\sum_{i=h_j+1}^{n} \mu_{(i)j}^2 + \mathrm{P}_j(h_j)\right\} + \sum_{j \in \mathcal{J}_0^*} \mathrm{P}_j(1) + \mathrm{P}_0(m)\right] + c_2(\gamma)\sigma_n^2\{1 - \pi_0(0)\}, \tag{A4}$$

where the conditions on $\pi_j(1)$ and $\pi_0(m)$ imply that $\sum_{j \in \mathcal{J}_0^*} \mathrm{P}_j(1) = O(\sigma_n^2 m \log n)$ and $\mathrm{P}_0(m) = O(\sigma_n^2 m)$. From Table 1 one can verify that for all dense and sparse cases, $\sigma_n^2 \log n = O[R\{\Theta_j(\eta_{jn})\}]$ $(j \in \mathcal{J}_0^{c*})$ and, therefore, the first term $\sum_{j \in \mathcal{J}_0^{c*}}$ in the right-hand side of (A4) is dominating for $m_0^* \sim m$. □

Proof of Theorems 3–4. The ideas of the proofs of Theorems 3–4 are similar and can be combined.

No estimator can perform better than an oracle that knows the true $\mathcal{J}_0$. In this ideal case one would obviously set $\hat{\mu}_j = 0$ for all $j \in \mathcal{J}_0$ with zero risk and, therefore, due to the additivity of the risk function,

$$\inf_{\tilde{\mu}_1, \ldots, \tilde{\mu}_m} \sup_{\mu_j \in \Theta_j(\eta_{jn}), j \in \mathcal{J}_0^c} \sum_{j=1}^{m} E\left(\|\tilde{\mu}_j - \mu_j\|_2^2\right) \geqslant C \sum_{j \in \mathcal{J}_0^c} R\{\Theta_j(\eta_{jn})\}$$

for any $\Theta_{jn}(\eta_{jn})$. See, e.g., Proposition 4.14 of an unpublished 2011 Stanford University, Department of Statistics manuscript of Johnstone entitled Gaussian estimation: Sequence and wavelet methods.

Furthermore, following II in the proof of Theorem 2, $\sum_{j \in \mathcal{J}_0^c} R\{\Theta_j(\eta_{jn})\}$ dominates over $\sigma_n^2 m_0 \log(m/m_0)$ in (16) and (17) for $m_0 > m/2$. To complete the proof we need to show, therefore, that for $m_0 \leqslant m/2$, the minimal unavoidable price for not being an oracle for selecting nonzero μ_j is of order $\sigma_n^2 m_0 \log(m/m_0)$.

The main idea of the proof is to find a subset $\mathcal{M}_{m_0}$ of vectors $\mu = (\mu_{11}, \ldots, \mu_{n1}, \ldots, \mu_{1m}, \ldots, \mu_{nm})^{\mathrm{T}}$ with m_0 nonzero $\mu_j = (\mu_{1j}, \ldots, \mu_{nj})^{\mathrm{T}} \in \Theta_j[\eta_{jn}]$ such that for any pair $\mu^1, \mu^2 \in \mathcal{M}_{m_0}$ and some $C > 0$, $\|\mu^1 - \mu^2\|_2^2 \geqslant C\sigma_n^2 m_0 \log(m/m_0)$, while the Kullback–Leibler divergence $K(\mathbb{P}_{\mu^1}, \mathbb{P}_{\mu^2}) = \|\mu^1 - \mu^2\|_2^2/(2\sigma_n^2) \leqslant (1/16)\log \mathrm{card}(\mathcal{M}_{m_0})$. The result then follows immediately from Lemma A.1 of Bunea et al. (2007).

Define the subset $\tilde{\mathcal{D}}_{m_0}$ of all m-dimensional indicator vectors with m_0 entries of ones, that is $\tilde{\mathcal{D}}_{m_0} = \{d : d \in \{0, 1\}^m,\ \|d\|_0 = m_0\}$. By Lemma A.3 of Rigollet & Tsybakov (2011), for $m_0 \leqslant m/2$ there exists a subset $\mathcal{D}_{m_0} \subset \tilde{\mathcal{D}}_{m_0}$ such that for some constant $\tilde{c} > 0$, $\log \mathrm{card}(\mathcal{D}_{m_0}) \geqslant \tilde{c} m_0 \log(m/m_0)$, and for any pair $d_1, d_2 \in \mathcal{D}_{m_0}$, the Hamming distance $\rho(d_1, d_2) = \sum_{j=1}^{m} \mathbb{I}\{d_{1j} \neq d_{2j}\} \geqslant \tilde{c} m_0$.

To any indicator vector $d \in \mathcal{D}_{m_0}$ assign the corresponding mean vector $\mu \in \mathcal{M}_{m_0}$ as follows. Let $\tilde{C}^2 = (1/16)\sigma_n^2 \tilde{c} \log(m/m_0)$. Define $\mu_j = (\tilde{C}, 0, \ldots, 0)^{\mathrm{T}} \mathbb{I}\{d_j = 1\}$ for $0 \leqslant p_j < 2$ and $\mu_j = (\tilde{C} n^{-1/2}, \tilde{C} n^{-1/2}, \ldots, \tilde{C} n^{-1/2})^{\mathrm{T}} \mathbb{I}\{d_j = 1\}$ for $p_j \geqslant 2$ $(j = 1, \ldots, m)$. Hence, $\mathrm{card}(\mathcal{M}_{m_0}) = \mathrm{card}(\mathcal{D}_{m_0})$. Obviously, the resulting $\mu_j \in l_0[\eta_{jn}]$ and straightforward calculus show that under the additional constraint on η_{jn} and m_0 in Theorem 4, $\mu_j \in l_{p_j}(\eta_{jn})$.

For any $\mu^1, \mu^2 \in \mathcal{M}_{m_0}$ and the corresponding $d_1, d_2 \in \mathcal{D}_{m_0}$, we then have

$$\|\mu^1 - \mu^2\|_2^2 = \tilde{C}^2 \sum_{j=1}^{m} \mathbb{I}\{d_{1j} \neq d_{2j}\} \geqslant \tilde{C}^2\, \tilde{c}\, m_0 = \frac{1}{16}\sigma_n^2 \tilde{c}^2 m_0 \log(m/m_0),$$

$$K(\mathbb{P}_{\mu^1}, \mathbb{P}_{\mu^2}) = \frac{\tilde{C}^2}{2\sigma_n^2} \sum_{j=1}^{m} \mathbb{I}\{d_{1j} \neq d_{2j}\} \leqslant \frac{\tilde{C}^2 m_0}{\sigma_n^2} \leqslant \frac{1}{16} \log \mathrm{card}(\mathcal{M}_{m_0}),$$

which completes the proof. □

REFERENCES

ABRAMOVICH, F. & GRINSHTEIN, V. (2010). MAP model selection in Gaussian regression. *Electron. J. Statist.* **4**, 932–49.

ABRAMOVICH, F., GRINSHTEIN, V. & PENSKY, M. (2007). On optimality of Bayesian testimation in the normal means problem. *Ann. Statist.* **35**, 2261–86.

ABRAMOVICH, F., GRINSHTEIN, V., PETSA, A. & SAPATINAS, T. (2010). On Bayesian testimation and its application to wavelet thresholding. *Biometrika* **97**, 181–98.

ANGELINI, C., DE CANDITIIS, D., MUTARELLI, M. & PENSKY, M. (2007). A Bayesian approach to estimation and testing in time-course microarray experiments. *Statist. Appl. Genet. Molec. Biol.* **6**, 1–30.

ANTONIADIS, A. & FAN, J. (2001). Regularization of wavelet approximations. *J. Am. Statist. Assoc.* **96**, 939–55.

BIRGÉ, L. & MASSART, P. (2001). Gaussian model selection. *J. Eur. Math. Soc.* **3**, 203–68.

BUNEA, F., TSYBAKOV, A. & WEGKAMP, M. H. (2007). Aggregation for Gaussian regression. *Ann. Statist.* **35**, 1674–97.

DONOHO, D. L. & JOHNSTONE, I. M. (1994a). Ideal spatial adaptation via wavelet shrinkage. *Biometrika* **81**, 425–55.

DONOHO, D. L. & JOHNSTONE, I. M. (1994b). Minimax risk over ℓ_p-balls for ℓ_q-error. *Prob. Theory Rel. Fields* **99**, 277–303.

DONOHO, D. L., JOHNSTONE, I. M., HOCH, C. & STERN, A. (1992). Maximum entropy and the nearly black object (with Discussion). *J. R. Statist. Soc.* B **54**, 41–81.

JOHNSTONE, I. M. (1994). Minimax Bayes, asymptotic minimax and sparse wavelet priors. In *Statistical Decision Theory and Related Topic V*, Ed. S. Gupta & J. Berge, pp. 5–14. Springer Verlag.

LIN, Y. & ZHANG, H. H. (2006). Component selection and smoothing in multivariate nonparametric regression. *Ann. Statist.* **34**, 2272–97.

MEIER, L., VAN DE GEER, S. & BUHLMANN, P. (2009). High-dimensional additive modelling. *Ann. Statist.* **37**, 3779–821.

RASKUTTI, G., WAINWRIGHT, M. J. & YU, B. (2011). Minimax rates of estimations for high-dimensional regression over l_q balls. *IEEE Trans. Info. Theory* **57**, 6976–94.

RASKUTTI, G., WAINWRIGHT, M. J. & YU, B. (2012). Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *J. Mach. Learn. Res.* **13**, 389–427.

RAVIKUMAR, P., LAFFERTY, J., LIU, H. & WASSERMAN, L. (2009). Sparse additive models. *J. R. Statist. Soc.* B **71**, 1009–30.

RIGOLLET, P. & TSYBAKOV, A. (2011). Exponential screening and optimal rates of sparse estimation. *Ann. Statist.* **39**, 731–71.

SIMON, N., FRIEDMAN, J., HASTIE, T. & TIBSHIRANI, R. (2013). The sparse-group lasso. *J. Comp. Graph. Statist.*, doi: 10.1080/10618600.2012.681250

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B **58**, 267–88.

WU, Z. & ZHOU, H. (2013). Model selection and sharp asymptotic minimaxity. *Prob. Theory Rel. Fields*, doi: 10.1007/s00440-012-0424-5

YUAN, M. & LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc.* B **68**, 49–67.

ZOU, H. & HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Statist. Soc.* B **67**, 301–20.

[*Received October* 2011. *Revised November* 2012]