Estimation of a sparse group of sparse vectors

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

Vadim Grinshtein
Department of Mathematics
The Open University of Israel
Raanana 43537, Israel
vadimg@openu.ac.il

Abstract

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

Keywords: Adaptive minimaxity; complexity penalty; maximum a posteriori rule; sparsity; thresholding.

1 Introduction

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

$$y_j = \mu_j + \epsilon_j, \quad \epsilon_j \text{i.i.d.} \sim \mathcal{N}_n(0, \sigma_n^2 I_n), \quad j = 1, \ldots, m$$
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 the model (1) is both within- and between-vectors sparsity (hereafter within- and between-sparsity for brevity). More specifically, we assume that part of $\mu_j$’s are identically zero vectors and the entire information in the noisy data is contained only in a small fraction of them (between-sparsity). Moreover, even within nonzero $\mu_j$’s, most of their components are still zeroes or at least “negligible” (within-sparsity). Formally, the within-sparsity can be quantified in terms of $l_0$, strong or weak $l_p$-balls introduced further. Neither the indices of non-zero $\mu_j$’s nor the locations of their “significant” components are known in advance.

Such a model appears in the variety of statistical applications as we illustrate by the following two examples.

**Example 1. Additive models.** Consider a nonparametric regression model $y_i = f(x_{1i}, \ldots, x_{mi}) + \epsilon_i$, $i = 1, \ldots, n$, where $f : \mathbb{R}^m \to \mathbb{R}$ is the unknown regression function assumed to belong to some class of functions (e.g., Hölder, Sobolev or Besov classes), and $\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_n^2)$. Estimating $f$ in such a general setup suffers from a severe “curse of dimensionality”, where typically the sample size $n$ should grow exponentially with the dimensionality $m$ to achieve consistent estimation. It is essential then to place some extra restrictions on the complexity of $f$. One of the most common approaches is to consider the additive models, where $f(x_1, \ldots, x_m) = f_1(x_1) + \ldots + f_m(x_m)$ 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 part of predictors among $x_1, \ldots, x_m$ are really “significant”, while the impact of others is negligible if at all. Such sparse additive models are especially relevant for $m \sim n$ and $m \gg n$ setups and have been considered in Lin & Zhang (2006), Meier, van de Geer & Bühlmann (2009), Ravikumar et al. (2009), Raskutti, Wainwright & Yu (2012).

Expand each $f_j$, $j = 1, \ldots, m$ into (univariate) orthonormal series $\{\psi_{ij}\}$ as $\sum_{i} \mu_{ij} \psi_{ij}(x_j)$, where $\mu_{ij} = \int f_j(x_j) \psi_{ij}(x_j)dx_j$. The original nonparametric additive model is then transformed into the equivalent problem of estimating vectors of corresponding coefficients $\mu_1, \ldots, \mu_m$ within Gaussian noise (1), where for sparse additive models, most of $\mu_j$ are zeroes (between-sparsity). Moreover, for a properly chosen bases $\{\psi_{ji}\}$ (e.g., Fourier series for Sobolev or wavelets for more general Besov classes), the nonzero $\mu_j$ will be also sparse (within-sparsity).

**Example 2. Time-course microarray experiments.** In time-course microarray experiments the data consists of measurements of differences in the expression levels between “treated” and “control” samples of $m$ genes recorded at different times. A record on $j$-th 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 expression of most genes are the same in both groups ($f_j \equiv 0$) and the goal is to
identify the differentially expressed genes and estimate the corresponding non-identically zero expression profile functions \(f_j\). Similar to the previous example, each \(f_j\) is commonly expanded into some “parsimonious” orthonormal basis (e.g., Legendre polynomials, Fourier or wavelets) as \(f_j(t) = \sum_i \mu_{ij} \psi_{ij}(t)\) and in the coefficients domain the original functional model becomes

\[ y_{ij} = \mu_{ij} + z_{ij}, \quad j = 1, ..., m; \quad i = 1, ..., n \]

where \(y_{ij}\) are empirical coefficients of the data on \(j\)-th gene and \(z_{ij}\) are Gaussian noise (see, e.g., Angelini et al., 2007). For most genes, \(\mu_j \equiv 0\) (between-sparsity), while due to the parsimony of the chosen basis, for differentially expressed genes, \(\mu_j\) will still have sparse representation (within-sparsity).

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

\[ \hat{\mu}_j = \arg\min_{\tilde{\mu}_j \in \mathbb{R}^n} ||y_j - \tilde{\mu}_j||^2_2 + \text{Pen}_j(\tilde{\mu}_j), \]

corresponding to different choices of penalties \(\text{Pen}_j(\tilde{\mu})\). In particular, the \(l_1\)-type penalty \(\text{Pen}_j(\tilde{\mu}_j) = \lambda ||\tilde{\mu}_j||_1\) leads to soft thresholding of components of \(\tilde{\mu}_j\) with a 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 the \(l_0\) or complexity type penalties \(\text{Pen}_j(||\tilde{\mu}_j||_0)\) on the number of nonzero components \(\tilde{\mu}_{ij}\), where \(||\tilde{\mu}_j||_0 = \#\{i : \tilde{\mu}_{ij} \neq 0\}\), that yield hard thresholding rules. In the simplest case, where \(\text{Pen}_j(||\tilde{\mu}_j||_0) = \lambda ||\tilde{\mu}_j||_0\), the resulting (constant) threshold is \(\sqrt{\lambda}\). More general complexity penalties were studied in Birgé & Massart (2001), Abramovich, Grinshtein & Pensky (2007), Abramovich et al. (2010) and Wu & Zhou (2012).

Penalizing each \(\tilde{\mu}_j\) separately, however, essentially ignores the between-sparsity, where it is assumed that most of \(\mu_j\) are identically zeroes and should be obviously estimated by \(\hat{\mu}_j = 0\). Thus, simultaneous estimation of all \(m\) mean vectors in (1) should involve an additional penalty \(\text{Pen}_0(\cdot)\) on the number of nonzero \(\hat{\mu}_j\)’s that are now defined as solutions of the following criterion:

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

(2)

where \(k = \#\{j : \hat{\mu}_j \neq 0\}\). In this paper we investigate the optimality of such an approach for estimating \(\mu_1, ..., \mu_m\) under various within- and between-sparsity setups. In particular, we specify the classes of complexity penalties \(\text{Pen}_j(||\tilde{\mu}_j||_0)\) and \(\text{Pen}_0(k)\) on respectively within- and
between sparsity for which the resulting estimators \( \hat{\mu}_1, ..., \hat{\mu}_m \) achieve asymptotically minimax rates simultaneously for the 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 Bayesian MAP testimation approach developed in Abramovich, Grinshtein & Pensky (2007) and Abramovich et al. (2010) for estimating a single normal mean vector to simultaneous estimation of a group of \( m \) vectors in the 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 \( \mu_j \) in (1) results in per-component (soft) thresholding of each \( y_j \) that handles within-sparsity but ignores between-sparsity. To address the latter, Yuan & Lin (2006) proposed a group lasso that for the particular model (1) at hand solves

\[
\min_{\tilde{\mu}_1, ..., \tilde{\mu}_m \in \mathbb{R}^n} \sum_{j=1}^m \left\{ \|y_j - \tilde{\mu}_j\|_2^2 + \lambda \|\tilde{\mu}_j\|_2 \right\}
\]

It can be easily shown that in such a setup, the group lasso estimator is available in the closed form, namely, \( \hat{\mu}_j = (1 - \lambda_1/2 \|\tilde{y}_j\|_2^2)_{+} \tilde{y}_j \), \( j = 1, ..., m \) which is the vector-level “shrink-or-kill” thresholding with a threshold \( \lambda/2 \). The \( \hat{\mu}_j \)'s 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 (2010) introduced the sparse group lasso that for the model (1) is defined as

\[
\min_{\tilde{\mu}_1, ..., \tilde{\mu}_m \in \mathbb{R}^n} \sum_{j=1}^m \left\{ \|y_j - \tilde{\mu}_j\|_2^2 + \lambda_1 \|\tilde{\mu}_j\|_2 + \lambda_2 \|\tilde{\mu}_j\|_1 \right\}
\]

yielding \( \hat{\mu}_j = (1 - \lambda_1/2 \|\tilde{y}_j\|_2^2)_{+} \tilde{y}_j \), \( j = 1, ..., m \), where \( \tilde{y}_{ij} = \text{sign}(y_{ij})(|y_{ij}| - \lambda_2/2)_{+} \), \( i = 1, ..., n \) is the result of component-level soft thresholding of each \( y_j \) with a threshold \( \lambda_2/2 \).

To the best of our knowledge, there are no theoretical results on optimality of sparse group lasso similar to those presented in this paper for the complexity penalized estimators (2). Moreover, we believe that, generally, \( l_0 \)-type penalties are more “natural” for representing sparsity and the main reason for other types of penalties (\( l_1 \) in particular) are mostly computational. For a general regression model, complexity penalties indeed imply combinatorial search over all possible models, while, for example, sparse group lasso estimator can be still efficiently computed by numerical iterative algorithms (see Friedman, Hastie & Tibshirani, 2010 and Simon et al., 2011 for details). However, for the model (1), that can be essentially viewed as a special case of a general regression setup, (2) can be also solved by fast algorithms (see Section 2) that makes such computational arguments irrelevant.

The paper is organized as follows. In Section 2 we develop a Bayesian formalism that gives raise to penalized estimators (2). The asymptotic (as both \( m \) and \( n \) increase) adaptive minimaxity of the resulting sparse group MAP estimators over various sparse and dense settings is investigated.
in Section 3. The short simulation study is presented in Section 4 and some concluding remarks are given in Section 5. All the proofs are placed in the Appendix.

2 Bayesian sparse group MAP estimation

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

We introduce first some notations. Let $J_0$ and $J_0^c$ be the sets of indices corresponding respectively to zero and nonzero mean vectors $\mu_j$’s, and $m_0 = |J_0^c| = \# \{ j : \mu_j \neq 0, \ j = 1, \ldots, m \}$. Denote by $h_j = \sum_{i=1}^{n} d_{ij} = \# \{ i : \mu_{ij} \neq 0, \ i = 1, \ldots, n \}$ the number of nonzero components in $\mu_j$, where evidently $h_j = 0$ for $j \in J_0$.

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

$$P(J_0^c \mid |J_0^c| = m_0) = \binom{m}{m_0}^{-1}$$

Obviously, $h_j \mid \{ j \in J_0 \} \sim \delta(0)$ and, thus, $d_j \mid \{ j \in J_0 \} \sim \delta(0)$ and $\mu_j \mid \{ j \in J_0 \} \sim \delta(0)$. For nonzero $\mu_j$ we place independent priors $\pi_j(\cdot)$ on the number of their nonzero components, that is, $h_j \mid \{ j \in J_0^c \} \sim \pi_j(h_j) > 0$, $h_j = 1, \ldots, n$. In this case, we again assume that for a given $h_j$, all possible $\binom{n}{h_j}$ indicator vectors $d_j$ with $h_j$ nonzero components have the same prior probabilities and, therefore,

$$P(d_j \mid \|d_j\|_0 = h_j, j \in J_0^c) = \binom{n}{h_j}^{-1}$$

Finally, to complete the prior for $11$, we have $\mu_{ij} \mid d_{ij} = 0 \sim \delta(0)$, while nonzero $\mu_{ij}$ are assumed to be i.i.d. $N(0, \gamma \sigma_i^2)$, where $\gamma > 0$.

A straightforward Bayesian calculus yields the posterior probability for a given indicator matrix $D$:

$$P(D \mid y) \propto \pi_0(m_0) \left( \frac{m}{m_0} \right)^{-1} \prod_{j \in J_0^c} \left\{ \pi_j(h_j) \left( \frac{n}{h_j} \right)^{-1} (1 + \gamma)^{-\frac{h_j}{2}} e^{\frac{\gamma}{2} \sum_{i=1}^{n} y_{ij}^2 d_{ij}} \right\}$$
Given the posterior distribution \( P(D|y) \) we apply the maximum a posteriori (MAP) rule to choose the most likely configuration of zero and nonzero \( \mu_{ij} \) that leads to the following MAP criterion:

\[
\sum_{j \in J_0} \sum_{i=1}^{n} y_{ij}^2 d_{ij} + 2\sigma_n^2 (1 + 1/\gamma) \ln \left( \pi_j(h_j) \left( \frac{n}{h_j} \right)^{-1} (1 + \gamma)^{-\frac{h_j}{\gamma}} \right) \rightarrow \max_D
\]

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

\[
\sum_{j \in J_0} \sum_{i=1}^{h_j} y_{ij}^2 (1 + 1/\gamma) \ln \left( \pi_j(h_j) \left( \frac{n}{h_j} \right)^{-1} (1 + \gamma)^{-\frac{h_j}{\gamma}} \right) \rightarrow \max_D
\]

where \( |y_{(1)j}| \geq \ldots \geq |y_{(n)j}| \). For every \( j = 1, \ldots, m \) define

\[
\hat{h}_j = \arg \min_{1 \leq h_j \leq n} \left\{ \sum_{i=h_j+1}^{n} y_{ij}^2 (1 + 1/\gamma) \ln \left( \pi_j^{-1}(h_j) \left( \frac{n}{h_j} \right)^{\frac{h_j}{\gamma}} \right) \right\}
\]

Then, (4) is equivalent to minimizing

\[
\sum_{j \in J_0} \left\{ -\sum_{i=1}^{\hat{h}_j} y_{ij}^2 (1 + 1/\gamma) \ln \left( \pi_j^{-1}(\hat{h}_j) \left( \frac{n}{\hat{h}_j} \right)^{\frac{\hat{h}_j}{\gamma}} \right) \right\} + 2\sigma_n^2 (1 + 1/\gamma) \ln \left( \pi_0^{-1}(m_0) \left( \frac{m}{m_0} \right) \right) \rightarrow \max_D
\]

over all subsets of indices \( J_0 \subseteq \{1, \ldots, m\} \). Define

\[
W_j = -\sum_{i=1}^{\hat{h}_j} y_{ij}^2 (1 + 1/\gamma) \ln \left( \pi_j^{-1}(\hat{h}_j) \left( \frac{n}{\hat{h}_j} \right)^{\frac{\hat{h}_j}{\gamma}} \right)
\]

Then, (7) is obviously reduced to

\[
\min_{0 \leq m_0 \leq m} \left\{ \sum_{j=1}^{m_0} W_{(j)} + 2\sigma_n^2 (1 + 1/\gamma) \ln \left( \frac{\pi_0^{-1}(m_0)}{m_0} \right) \right\},
\]

where \( W_{(1)} \leq \ldots \leq W_{(m)} \) and for \( m_0 = 0 \) the sum in the RHS of (7) evidently does not appear.

Summarizing, the efficient simple algorithm for finding the proposed sparse group MAP estimators of \( \mu_1, \ldots, \mu_m \) in (1) can be formulated as follows:

**Sparse group MAP estimation algorithm**

1. For every \( j = 1, \ldots, m \), find \( \hat{h}_j \) in (6) and calculate the corresponding \( W_j \) in (8).
2. Order $W_j$ in ascending order $W_{(1)} \leq \ldots \leq W_{(m)}$ and find

$$
\hat{m}_0 = \arg \min_{0 \leq m_0 \leq m} \left\{ \sum_{j=1}^{m_0} W_{(j)} + 2\sigma_n^2(1 + 1/\gamma) \ln \left( \pi_0^{-1}(m_0) \left( \frac{m}{m_0} \right) \right) \right\}
$$

3. Let $\hat{J}_0$ be the set of indices corresponding to the $m_0$ smallest $W_j$. Set $\hat{\mu}_j \equiv 0$ for all $j \in \hat{J}_0$, while for $j \in \hat{J}_0^c$, take the $\hat{h}_j$ largest $|y_{ij}|$ and threshold others, that is, $\hat{\mu}_{ij} = y_{ij} I(|y_{ij}| \geq |y_{(\hat{h}_j)j}|)$, $i = 1, \ldots, n$, $j \in \hat{J}_0^c$, where $|y_{(1)j}| \geq \ldots \geq |y_{(n)j}|$.

The resulting estimation procedure combines therefore vector-wise and component-wise thresholding. It is easily verified that the minimizer of \((7)\) is, in fact, the penalized likelihood thresholding estimator with a data-driven threshold

$$
Pen_j(0) = 0, \quad Pen_j(h_j) = 2\sigma_n^2(1 + 1/\gamma) \ln \left( \pi_j^{-1}(h_j) \left( \frac{n}{h_j} \right)^{h_j} \right), \quad h_j = 1, \ldots, m
$$

and

$$
Pen_0(m_0) = 2\sigma_n^2(1 + 1/\gamma) \ln \left( \pi_0^{-1}(m_0) \left( \frac{m}{m_0} \right) \right), \quad m_0 = 0, \ldots, m
$$

The specific types of penalties $Pen_j(\cdot)$’s and $Pen_0(\cdot)$ depend on the choices of priors $\pi_j(\cdot)$’s 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 $Pen(m_0) = 2\sigma_n^2 \lambda_0^2 m_0$ and $Pen_j(h_j) = 2\sigma_n^2 \lambda_j^2 h_j$ respectively, where $\lambda_0^2 = (1 + 1/\gamma) \ln \{ (1 - \xi_0)/\xi_0 \}$ and $\lambda_j^2 = (1 + 1/\gamma) \ln \{ \sqrt{\gamma + 1} - \gamma(1 - \xi_j)/\xi_j \}$. For such a choice of $\pi_j(\cdot)$, $W_j$ in \((8)\) is essentially obtained by hard thresholding of $y_j$ with a constant threshold $\sqrt{2}\sigma_n \lambda_j$. In particular, $\xi_j = \sqrt{\gamma + 1}/(\sqrt{\gamma + 1} + n^{\gamma/(\gamma+1)})$ leads to the universal thresholding of Donoho & Johnstone (1994a) with $\lambda_j = \sqrt{\ln n}$. 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 $2k\ln(n/k)$-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, Grinshtein & Pensky (2007), Abramovich et. al (2010), Wu & Zhou (2012).

### 3 Adaptive minimaxity of sparse group MAP estimators

In this section we investigate the goodness of the proposed sparse group MAP estimators \((2)\) with the penalties \((10)-(11)\), where the goodness-of-fit is measured by the global quadratic risk $\sum_{j=1}^{m} E \| \hat{\mu}_j - \mu_j \|^2$. 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 (P).** Assume that

$$
\pi_j(h) \leq \left( \frac{n}{h} \right) e^{-c(\gamma)h}, \quad h = 1, \ldots, n, \quad j = 1, \ldots, m,
$$

\((12)\)
where \( c(\gamma) = 8(\gamma + 3/4)^2 > 9/2 \). Assumption (P) is, in fact, not restrictive. Indeed, the obvious inequality \( \binom{n}{k} \geq (n/h)^k \) implies that for any \( \pi_j(\cdot) \), (12) holds for all \( h \leq ne^{-c(\gamma)} \). In particular, Assumption (P) is satisfied for binomial priors \( B(n, \xi_j) \) with \( \xi_j \leq e^{-c(\gamma)}/(1 + e^{-c(\gamma)}) \) and (truncated) geometric priors.

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

**Theorem 1** (general upper bound). Consider the sparse group MAP estimators \( \hat{\mu}_1, \ldots, \hat{\mu}_m \) of \( \mu_1, \ldots, \mu_m \) with the complexity penalties (10)-(11) in the model (7). Under Assumption (P) we have

\[
\begin{aligned}
\sum_{j=1}^m E\|\hat{\mu}_j - \mu_j\|^2 &\leq c_1(\gamma) \min_{J_0 \subseteq \{1, \ldots, m\}} \left\{ \sum_{j \in J_0} \min_{1 \leq h_j \leq n} \left( \sum_{i=h_j+1}^{n} \mu_{ij}^2 + \text{Pen}_j(h_j) \right) \right. \\
&\quad + \sum_{j \in J_0} \sum_{i=1}^{n} \mu_{ij}^2 + \text{Pen}_0(|J_0^c|) \right\} + c_2(\gamma) \sigma_n^2(1 - \pi_0(0)),
\end{aligned}
\]

where \( |\mu_{(1)}| \geq \ldots \geq |\mu_{(n)}| \) and \( c_1(\gamma), c_2(\gamma) \) depend only on \( \gamma \).

The results of Theorem 1 hold for any normal mean vectors \( \mu_1, \ldots, \mu_m \). Now we consider (11) under the extra within- and between-sparsity assumptions that will be defined more rigorously below.

The between-sparsity is naturally measured by the number \( m_0 \) of nonzero \( \mu_j \)'s. The within-sparsity can be introduced in several ways. The most intuitive measure of within-sparsity of 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 then an \( l_0 \)-ball \( l_0[\eta] \) of standardized radius \( \eta \) as a set of \( \mu \) with at most a proportion \( \eta \) of non-zero entries, that is

\[
l_0[\eta] = \{ \mu \in \mathbb{R}^n : ||\mu||_0 \leq \eta n \}
\]

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

\[
m_p[\eta] = \{ \mu \in \mathbb{R}^n : |\mu|_{(i)} \leq \sigma_n \eta (n/i)^{1/p}, \ i = 1, \ldots, n \},
\]

where \( \mu_{(1)} \geq \ldots \geq \mu_{(n)} \) are the ordered components of \( \mu \). For \( \mu \in m_p[\eta] \), the proportion of \( |\mu_i| \)'s larger than \( \sigma_n \delta \) for some \( \delta > 0 \) is at most \( (\eta/\delta)^p \).

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

\[
l_p[\eta] = \{ \mu \in \mathbb{R}^n : \frac{1}{n} \sum_{i=1}^{n} |\mu_i|^p \leq \sigma_n^p \eta^p \}
\]
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:

\[
l_p[\eta] \subset m_p[\eta] \not\subset l_{p'}[\eta], \; p' > p
\]

We recall first the known results on minimax rates for estimating a single normal mean vector \( \mu \) 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 \( \eta \) might depend on \( n \). The corresponding minimax quadratic risk for estimating a single \( \mu \) (\( m = 1 \)) over \( \Theta[\eta_n] \) in (1) is

\[
R(\Theta[\eta_n]) = \inf_{\hat{\mu}} \sup_{\mu \in \Theta[\eta_n]} \mathbb{E}[\|\hat{\mu} - \mu\|^2],
\]

where the infimum is taken over all estimates \( \hat{\mu} \) of \( \mu \). For \( p > 0 \) define

\[
\eta_0 = n^{-1/\min(p,2)} \sqrt{\ln n}.
\]

Depending on the behaviour of \( \eta_n \) as \( n \) increases, we distinguish between three cases for \( p > 0 \) and two cases for \( p = 0 \):

a) **dense**, where \( \eta_n \not\to 0 \)

b) **sparse**, where \( \eta_n \to 0 \) but \( \eta_n/\eta_0 \not\to 0 \) for \( p > 0 \) and, obviously, \( \eta_n \geq n^{-1} \) for \( p = 0 \)

c) **super-sparse** (for \( p > 0 \)), where \( \eta_n/\eta_0 \to 0 \)

The corresponding minimax convergence rates over \( R(\Theta[\eta_n]) \) for various cases and \( p \) are summarized in Table 1 below (see 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 \( p = 2 \), where there is an additional log-term. Table 1 defines dense and sparse zones for \( p = 0 \) and \( p \geq 2 \), and dense, sparse and super-sparse zones for \( 0 < p < 2 \) of different minimax rates.

| Case            | \( p = 0 \)                  | \( 0 < p < 2 \)                  | \( p \geq 2 \)                  |
|-----------------|-----------------------------|---------------------------------|---------------------------------|
| dense case      | \( \sigma_n^2 n \)          | \( \sigma_n^2 n \)              | \( \sigma_n^2 n \)              |
| sparse case     | \( \sigma_n^2 n \eta_n (\ln n)^{-1} \) | \( \sigma_n^2 n \eta_n^p (\ln n)^{-p+1/2} \) | \( \sigma_n^2 n \eta_n^2 \) |
| super-sparse case| \( - \)                    | \( \sigma_n^2 n \eta_n^2 \)     | \( \sigma_n^2 n \eta_n^2 \)     |

Table 1: Minimax rates (up to multiplying constants) 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 \( p = 2 \), where for \( m_p[\eta_n] \) there appears the additional log-term which is not presented in Table 1 for brevity.

Consider now the model (1) for \( m \geq 1 \). Recall that \( m_0 = \#\{j : \mu_j \neq 0\} \) and \( J_0^c \) is the set of indices for nonzero \( \mu_j \). In what follows we assume that \( \mu_j \in \Theta_j[\eta_n] \) for \( j \in J_0^c \), where the types \( (l_0, \text{weak } m_p \text{ or strong } l_p) \) 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 asymptotic upper bounds for the quadratic risks of the sparse group MAP estimator in (1) under within- and between sparsity assumptions:
Theorem 2 (upper bounds over sparse and dense settings). Consider the model \( \mathcal{J}_0^c \neq \emptyset \) (not pure noise). Assume that \( \mathbf{\mu}_j \in \Theta_j[\eta_{jn}] \) for all \( j \in \mathcal{J}_0^c \), where \( \eta_{jn} \geq n^{-1/\min(p_j,2)}\sqrt{\ln n} \) for all \( p_j > 0 \) (excluding, thus, super-sparse cases).

Let \( \hat{\mathbf{\mu}}_1, \ldots, \hat{\mathbf{\mu}}_m \) be the sparse group MAP estimators \(^2\) with the complexity penalties \(^{11}\), while to have

1. \( \pi_0(k) \geq (k/m)^c_k \), \( k = 1, \ldots, \lfloor m/e \rfloor \) and \( \pi_0(m) \geq e^{-c_0 m} \)
2. for all \( j = 1, \ldots, m \), \( \pi_j(\cdot) \) satisfy Assumption (P) and, in addition, \( \pi_j(h) \geq (h/n)^{c_j h}, \) \( h = 1, \ldots, \lfloor ne^{-c(\gamma)} \rfloor; \) \( \pi_j(n) \geq e^{-c_2 n} \)

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

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

for some constant \( C_1(\gamma) \) depending only on \( \gamma \), where the corresponding \( R(\Theta_j[\eta_{hn}]) \) are given in Table \( \mathcal{I} \) (up to multiplying constants).

Theorem 2 shows that as both \( m \) and \( n \) increase, the asymptotic convergence rates in (14) are either of order \( \sum_{j \in \mathcal{J}_0^c} R(\Theta_j[\eta_{jn}]) \) or \( \sigma_n^2 m_0 \ln(m/m_0) \). 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 \( m \) (see, e.g. Abramovich & Grinshtein, 2010; Raskutti, Wainwright & Yu, 2011; Rigollet & Tsybakov, 2011). From Table 1 it follows that for all within-dense and within-sparse cases, \( C_1 \sigma_n^2 \ln n \leq R(\Theta_j[\eta_{jn}]) \leq C_2 \sigma_n^2 n \), \( j \in \mathcal{J}_0^c \) for some \( C_1, C_2 > 0 \) and, therefore, the first term \( \sum_{j \in \mathcal{J}_0^c} R(\Theta_j[\eta_{hn}]) \) in the upper bound (14) is always dominating for \( m_0 > m/n \), while the second term \( \sigma_n^2 m_0 \ln(m/m_0) \) is necessarily the main one for \( m_0 < m/e^n \).

One can easily verify that the conditions on the priors \( \pi_0(\cdot) \) and \( \pi_j(\cdot) \) required in Theorem 2 are satisfied, for example, for the (truncated) geometric priors (see Section 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 \( \pi_j(n) = \xi_j^n \geq e^{-c_2 n} \) yields \( \xi_j \geq e^{-c_2} \), while to have \( \pi_j(1) = n \xi_j(1 - \xi_j)^{n-1} \geq n^{-c_1} \) one needs \( \xi_j \to 0 \) as \( n \) increases.

To establish the corresponding lower bound for the minimax risk, for simplicity of exposition we consider only the two cases, where \( p_j \) for \( j \in \mathcal{J}_0^c \) are either all zeroes or all positive. In fact, these are the two main scenarios appearing in various setups. Somewhat similar results for minimax lower bounds in the particular context of sparse nonparametric additive models (see Introduction) appear in Raskutti, Wainwright and Yu (2012).
Theorem 3 (minimax lower bounds for $l_0$-balls). Consider the model (1), where $\mu_j \in l_0[\eta_{jn}], j \in \mathcal{J}^c_0$. Assume that $|\mathcal{J}^c_0| = m_0 > 0$. Then, there exists a constant $C_2 > 0$ such that
\[
\inf_{\tilde{\mu}_1, \ldots, \tilde{\mu}_m} \sup_{\mu_j \in l_0[\eta_{jn}], j \in \mathcal{J}^c_0} \sum_{j=1}^m E||\tilde{\mu}_j - \mu_j||^2_2 \geq C_2 \max \left( \sum_{j \in \mathcal{J}^c_0} R(l_0[\eta_{jn}]), \sigma_n^2 m_0 \ln(m/m_0) \right),
\] (15)
where the infimum is taken over all estimates $\tilde{\mu}_1, \ldots, \tilde{\mu}_m$ of $\mu_1, \ldots, \mu_m$.

Theorem 3 shows that, as $m$ and $n$ increase, the rates in (14) cannot be improved for $l_0$-balls. The proposed sparse group MAP estimator in this case is, therefore, adaptive to the unknown degrees of within- and between-sparsity and is simultaneously rate-optimal (in the minimax sense) over entire range of dense and sparse $l_0$-balls 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 $p > 0$ (see above), it is sufficient to establish the minimax lower bounds for strong $l_p$-balls settings.

Theorem 4 (minimax lower bounds for $l_p$-balls). Consider the model (1), where $\mu_j \in l_{p_j}[\eta_{jn}], j \in \mathcal{J}^c_0$ and $|\mathcal{J}^c_0| = m_0 > 0$. In addition, assume that $\eta_{jn}^2 \geq n^{-2/\min(p_j,2)} \max(\ln n, \ln(m/m_0))$. Under this additional constraint, there exists a constant $C_2 > 0$ such that
\[
\inf_{\tilde{\mu}_1, \ldots, \tilde{\mu}_m} \sup_{\mu_j \in l_{p_j}[\eta_{jn}], j \in \mathcal{J}^c_0} \sum_{j=1}^m E||\tilde{\mu}_j - \mu_j||^2_2 \geq C_2 \max \left( \sum_{j \in \mathcal{J}^c_0} R(l_{p_j}[\eta_{jn}]), \sigma_n^2 m_0 \ln(m/m_0) \right),
\] (16)
where the infimum is taken over all estimates $\tilde{\mu}_1, \ldots, \tilde{\mu}_m$ of $\mu_1, \ldots, \mu_m$.

Similar to Theorem 3, Theorem 4 implies simultaneous optimality (in the minimax sense) of MAP sparse group estimator over strong and weak $l_p$-balls but with the restriction on $\eta_{jn}$ and $m_0$. In particular, it does not cover settings with within-super-sparsity but might also exclude part of the corresponding within-sparse zone (depending on $m_0$). Within- and between-sparsity cannot be “too strong” both. In fact, the condition $\eta_{jn}^2 < n^{-2/\min(p_j,2)} \max(\ln n, \ln(m/m_0))$, $j \in \mathcal{J}^c_0$ can be viewed as an extended definition of super-sparsity for $m > 1$. For such a super-sparse case, the minimax bound (16) does not hold and can be reduced. Indeed, consider the trivial zero estimator $\tilde{\mu} \equiv 0, j = 1, \ldots, m$, where, evidently,
\[
\sup_{\mu_j \in l_{p_j}[\eta_{jn}], j \in \mathcal{J}^c_0} \sum_{j=1}^m E||\tilde{\mu}_j - \mu_j||^2_2 = \sup_{\mu_j \in l_{p_j}[\eta_{jn}], j \in \mathcal{J}^c_0} \sum_{j \in \mathcal{J}^c_0} ||\mu_j||^2_2
\] (17)
The least favourable sequences that maximize $||\mu_j||^2_2$ over $l_{p_j}[\eta_{jn}]$ are $(\sigma_n \eta_{jn}, \ldots, \sigma_n \eta_{jn})$ and $(\sigma_n \eta_{jn} n^{1/p_j}, 0, \ldots, 0)$ for $p_j \geq 2$ 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 RHS of (17) is less than $\sigma_n^2 m_0 \ln(m/m_0)$ for $\eta_{jn}^2 < n^{-2/\min(p_j,2)} \ln(m/m_0)$, $j \in \mathcal{J}^c_0$. This goes along the lines with 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

A short simulation study was carried out to demonstrate the performance of the proposed approach.

The data was generated according to the model (1) with \( m = 10 \) vectors \( \mu_j \)'s of length \( n = 100 \). Five \( \mu_j \)'s were identically zeroes, while the other five had respectively 100, 70, 50, 20 and 5 nonzero components randomly sampled from \( N(0, \tau^2) \), \( \tau = 1, 3, 5 \) and zero others. Such a setup covers various types of within-sparsity. Finally, the independent standard Gaussian noise \( N(0, 1) \) was added to all components of each \( \mu_j \).

We tried binomial and truncated geometric priors for sparse group MAP estimators. For the binomial prior, we performed component-wise universal hard thresholding of Donoho & Johnstone (1994a) with a threshold \( \lambda = \sigma \sqrt{2 \log n} \) within each vector that essentially corresponds to \( \xi_j = \sqrt{\gamma + 1} / (\sqrt{\gamma + 1} + n^{\gamma/(\gamma + 1)}) \), where \( \gamma = \tau^2 / \sigma^2 \) (see Section 2), and used \( \xi_0 = 1/m \). For the geometric prior we set \( q_0 = q_j = 0.3 \). In addition, we compared the performances of sparse group MAP estimators with the sparse group lasso estimator (3) of Friedman, Hastie & Tibshirani (2010) described in Introduction. They do not discuss the optimal choices for \( \lambda_1 \) and \( \lambda_2 \) in (3). Some heuristical arguments are given in Simon et al. (2011). In our simulation study we considered instead two oracle-based choices for these tuning parameter giving thus a significant handicap to sparse group lasso estimators. Since in simulation examples the true mean vectors \( \mu_j \) are known, they can be used for optimal choosing \( \lambda_1 \) and \( \lambda_2 \). In particular, we considered a “semi-oracle” sparse group lasso estimator, where we set \( \lambda_2 = 2\sigma \sqrt{2 \log n} \) yielding universal soft thresholding within each vector (see Introduction) to compare the sparse group lasso with the binomial sparse group MAP. \( \lambda_1 \) was chosen by minimizing the mean squared error \( \sum_{j=1}^m E||\hat{\mu}_j(\lambda_1) - \mu_j||^2 \) estimated by averaging over a series of 1000 replications for each value of \( \lambda_1 \) by a grid search. In addition, we applied a “fully oracle” sparse group lasso estimator, where both \( \lambda_1 \) and \( \lambda_2 \) were chosen to minimize the mean squared error by the two-dimensional grid. It can be considered as a benchmark for the performance of sparse group lasso. Table 2 provides the resulting oracle choices for \( \lambda_1 \) and \( \lambda_2 \).

| \( \gamma \) | \( \lambda_1 \) | \( \lambda_2 \) |
|----------|----------|----------|
| 1        | 11.8     | 0.9      |
| 9        | 7.2      | 1.1      |
| 25       | 4.7      | 1.3      |

Table 2: The oracle choices for the parameters of the fully oracle sparse group lasso estimator \((\gamma = \tau^2 / \sigma^2)\).

Table 2 shows that for all \( \gamma \), the oracle choice for \( \lambda_2 \) in the sparse group lasso is much less than the conservative universal threshold \( 2\sigma \sqrt{2 \log n} \approx 6.06 \). The oracle thresholding within each vector is thus much less severe and keeps more coefficients. The oracle choices for \( \lambda_1 \) were also quite small.
and, as a result, for any $\gamma$, no single vector was thresholded by a fully oracle sparse group lasso, that is, all $\hat{\mu}_j \neq 0$. Thus it was really a non-sparse estimator for the considered setup.

In Table 3 we present the mean squared errors averaged over 1000 replications for the four sparse group estimators with the corresponding standard errors for various $\gamma$ (or, equivalently, $\tau$).

| $\gamma$ | Sparse Group MAP (binomial) | Sparse Group MAP (geometric) | Sparse Group Lasso (semi-oracle) | Sparse Group Lasso (fully oracle) |
|----------|-----------------------------|-------------------------------|---------------------------------|---------------------------------|
| 1        | 247.40 (0.71)               | 245.46 (0.70)                 | 236.85 (0.65)                   | 161.89 (0.43)                  |
| 9        | 608.02 (1.96)               | 378.87 (1.20)                 | 1120.99 (2.29)                  | 403.76 (0.91)                  |
| 25       | 549.77 (1.68)               | 351.52 (1.30)                 | 1595.91 (2.79)                  | 475.47 (1.07)                  |

Table 3: MSEs averaged over 1000 replications for four sparse group estimators and the corresponding standard errors (in brackets) for various $\gamma$.

For small $\gamma$ only few largest nonzero components can be distinguished from the noise that essentially corresponds to a sparse setting and explains good performance of binomial sparse group MAP and semi-oracle sparse group lasso estimators based on universal (respectively, hard and soft) thresholding within each vector. For larger $\gamma$, it becomes “over-conservative”. The negative effect of its conservativeness is much stronger for the soft than for hard thresholding (see comments below). The fully oracle sparse group lasso estimator strongly outperforms its semi-oracle counterpart especially for $\gamma = 9$ ($\tau = 3$) and $\gamma = 25$ ($\tau = 5$) also indicating that the universal thresholding is far from being optimal for sparse group lasso especially for moderate and large $\gamma$ (see also our previous comments on the optimal choice of $\lambda_2$).

On the other hand, geometric sparse group MAP estimator corresponding to a nonlinear $2k\ln(n/k)$-type penalty (see Section 2) provides good results for all $\gamma$ nicely following the theoretical results of Section 3. Moreover, for $\gamma = 9$ and $\gamma = 25$, it outperforms even the fully oracle sparse group lasso estimator that was essentially thought as a benchmark rather than a fair competitor. This indicates that that sparse group lasso faces general problems. In fact, it may be not so surprising since soft “shrink-or-kill” thresholding inherent for sparse group lasso is well-known to be superior to hard “keep-or-kill” thresholding in sparse group MAP estimation for small coefficients but worse for large ones due to the additional shrinkage. Moreover, sparse group lasso essentially involves a double amount of shrinkage - both within vectors and at each entire vector as a whole (see 3). It thus causes unnecessary extra bias growing with $\gamma$ that outweighs the benefits of variance reduction. Similar phenomenon appears also for naïve elastic set estimation (Zou & Hastie, 2005).
5 Concluding remarks

In this paper we considered estimation of a sparse group of sparse normal mean vectors. The proposed approach is based on penalized likelihood estimation with complexity penalties on both between- and within-sparsity and can be performed by a computationally fast algorithm. The resulting estimators naturally arise within Bayesian framework and can be viewed as MAP estimators corresponding to the priors on the number of nonzero mean vectors and the numbers of their nonzero components. Such a Bayesian perspective provides a natural tool for obtaining a wide class of penalized likelihood estimators with various complexity penalties.

We established the adaptive minimaxity of sparse group MAP estimators to the unknown degree of between- and within-sparsity over a wide range of sparse and dense settings. The short simulation study demonstrates the efficiency of the proposed approach that outperforms the recently presented sparse group lasso estimator.

Acknowledgments. Both authors were supported by the Israel Science Foundation grant ISF-248/08. We are grateful to Ofir Harari for his assistance in running simulation examples and Saharon Rosset for fruitful discussions.

Appendix

Throughout the proofs 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 $\gamma$.

Proof of Theorem 1

As we have mentioned in Section 2, the sparse group MAP estimator can be viewed as a penalized likelihood estimator (2) with the complexity penalties (10) and (11). We first re-write it in a somewhat different form that will allow us then to apply the general results of Birgé & Massart (2001) for complexity penalized estimators.

Let $y = (y_{11}, ..., y_{n1}, ..., y_{1m}, ..., y_{nm})'$ be an amalgamated $n \times m$ vector of data. Similarly, $\mu = (\mu_{11}, ..., \mu_{n1}, ..., \mu_{1m}, ..., \mu_{nm})'$, $\epsilon = (\epsilon_{11}, ..., \epsilon_{n1}, ..., \epsilon_{1m}, ..., \epsilon_{nm})'$ and the original model (1) can be re-written now as

$$y_i = \mu_i + \epsilon_i, \quad \epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2_n), \quad i = 1, ..., nm$$

(18)

Define an indicator vector $d$, where $d_i = I\{\mu_i \neq 0\}, \quad i = 1, ..., nm$. In terms of the model (18), $h_j = \sum_{i=n(j-1)+1}^{nj} d_i, \quad j = 1, ..., m$ and $m_0 = \#\{j : h_j > 0\}$. For a given $d$, define
\[ D_d = \sum_{j=1}^{m} h_j = \#\{i : d_i = 1, \ i = 1, \ldots, nm\} \]

\[ L_d = \frac{1}{D_d} \left( \sum_{j=1}^{m} \ln \left( \pi_{j}^{-1}(h_j) \left( \frac{n}{h_j} \right) \right) + \ln \left( \pi_{0}^{-1}(m_0) \left( \frac{m}{m_0} \right) \right) \right) \]

for \( d \neq 0 \) and \( L_0 = 2 \ln \pi_{0}^{-1}(0) \), where we formally set \( \pi_j(0) = 1 \). Then, the sparse group MAP estimator \( \hat{\mu} = (\hat{\mu}_{11}, \ldots, \hat{\mu}_{nm})' \) is the penalized likelihood estimator of \( \mu \) with the complexity penalty

\[ Pen(d) = 2\sigma_n^2 (1 + 1/\gamma) \left( \sum_{j=1}^{m} \ln \left( \pi_{j}^{-1}(h_j) \left( \frac{n}{h_j} \right) \right) (1 + \gamma) \frac{h_j}{n} \right) + \ln \left( \pi_{0}^{-1}(m_0) \left( \frac{m}{m_0} \right) \right) \]

for \( d \neq 0 \) and \( Pen(0) = \sigma_n^2 (1 + 1/\gamma) L_0 \).

One can verify that

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

A straightforward calculus (see the proof of Theorem 1 of Abramovich, Grinshtein & Pensky, 2007 for more details) implies also that for any \( d \) under Assumption (P),

\[ (1 + 1/\gamma)(2L_d + \ln(1 + \gamma)) \geq C(\gamma)(1 + \sqrt{2L_d})^2, \]

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

\[ \sum_{j=1}^{m} E||\hat{\mu}_j - \mu_j||^2 \leq c_1(\gamma) \min_{J_0 \subseteq \{1, \ldots, m\}} \left\{ \sum_{j \in J_0} \min_{1 \leq h_j \leq n} \left( \sum_{i=h_j+1}^{n} \mu_{ij}^2 + Pen_j(h_j) \right) + \sum_{j \in J_0} \sum_{i=1}^{n} \mu_{ij}^2 + Pen_0(m_0) \right\} + c_2(\gamma)\sigma_n^2 (1 - \pi_0(0)) \]

(19)

\[ \square \]

**Proof of Theorem 2**

One can easily check from Table 1 that for \( \eta_{jn} \geq n^{-1/\min(p_j, 2)} \sqrt{\ln n} \) for \( p_j > 0 \), the last term \( c_2(\gamma)\sigma_n^2 (1 - \pi_0(0)) \) in the RHS of (13) is of order \( O(\sigma_n^2) = o(R(\Theta_j [\eta_{jn}])) \) for all nonzero \( \mu_j \) and all \( p_j \geq 0 \).

Let \( J_0^* \) be the true (unknown) subset of nonzero \( \mu \)’s and \( m_0^* = |J_0^*| \).
I. \( m_0^* \leq \lfloor m/e \rfloor \).

Apply Theorem 1 for \( J_0 = J_0^* \):

\[
\sum_{j=1}^{m} E[||\hat{\mu}_j - \mu_j||_2^2] \leq c_1(\gamma) \left\{ \sum_{j \in J_0^*} \min_{1 \leq h_j \leq n} \left( \sum_{i=h_j+1}^{n} \frac{1}{h_j} \right) + 2\sigma_n^2(1 + 1/\gamma) \ln \left( \pi_j^{-1}(h_j) \left( \frac{n}{h_j} \right)^{(1 + \gamma) \frac{h_j}{2}} \right) \right\} + c_2(\gamma)\sigma_n^2(1 - \pi_0(0))
\]

Since for \( m_0 = 1, \ldots, \lfloor m/e \rfloor \), \( (\frac{m}{m_0}) \leq (m/m_0)^{2m_0} \) (see Lemma A1 of Abramovich et al. 2010), the required conditions on \( \pi_0(\cdot) \) ensure that

\[
2\sigma_n^2(1 + 1/\gamma) \ln \left( \pi_0^{-1}(m_0) \left( \frac{m}{m_0} \right) \right) \leq C(\gamma)\sigma_n^2m_0 \ln(m/m_0)
\]

To complete the proof for this case we consider now separately

\[
\min_{1 \leq h_j \leq n} \left( \sum_{i=h_j+1}^{n} \mu_{(i)j}^2 + 2\sigma_n^2(1 + 1/\gamma) \ln \left( \pi_j^{-1}(h_j) \left( \frac{n}{h_j} \right)^{(1 + \gamma) \frac{h_j}{2}} \right) \right)
\]

for each \( j \in J_0^* \) and show that it is \( O(R(\Theta_j[\eta_{jn}])) \) (see 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 Section 3.

Case 1: \( \mu_j \in \Theta_j[\eta_{jn}], \ \eta_{jn} > e^{-c(\gamma)} \) 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 (20) is \( O(\sigma_n^2n) = O(R(\Theta_j[\eta_{jn}])) \).

Case 2: \( \mu_j \in l_0[\eta_{jn}], \ \eta_{jn} \leq e^{-c(\gamma)} \). Note that since \( \mu_j \neq 0 \), \( \eta_{jn} \geq n^{-1} \). Choose \( h_j^* = m\eta_{jn} \) and repeat the arguments of the proof of Theorem 3 of Abramovich, Grinshtein & Pensky (2007) using a slightly more general Lemma A1 of Abramovich et al. (2010) for approximating the binomial coefficient in (20) instead of their original Lemma A1.

Case 3: \( \mu_j \in m_{p_j}[\eta_{jn}], \ 0 < p_j < 2, \ n^{-1}(\ln n)^{p_j/2} \leq \eta_{jn}^{p_j} \leq e^{-c(\gamma)} \). Take \( 1 \leq h_j^* = m\eta_{jn}(\ln \eta_{jn}^{-1/p_j})^{-p_j/2} \leq ne^{-c(\gamma)} \) and follow the proof of Theorem 4 of Abramovich, Grinshtein & Pensky (2007) with a more general version of Lemma A1 (see Case 2).

Case 4: \( \mu_j \in m_{p_j}[\eta_{jn}], \ \eta_{jn} \geq n^{p_j/2}(\ln n)^{p_j/2} \leq \eta_{jn}^{p_j} \leq e^{-c(\gamma)} \). Take \( h_j^* = 1 \). Then, for \( p_j > 2 \)

\[
\sum_{i=h_j^*+1}^{n} \mu_{(i)j}^2 < \sigma_n^2n^{2/p_j}\eta_{jn}^{2} \int_{1}^{n} x^{-2/p_j}dx < \frac{p_j}{p_j - 2} \sigma_n^2n^{2/p_j}\eta_{jn}^{2}n^{1-2/p_j} = O(\sigma_n^2n^{2}\eta_{jn}^{2})
\]

and, similarly, for \( p_j = 2 \)

\[
\sum_{i=h_j^*+1}^{n} \mu_{(i)j}^2 < \sigma_n^2n\eta_{jn}^{2} \int_{1}^{n} x^{-1}dx = \sigma_n^2n\eta_{jn}^{2} \ln n
\]
On the other hand, under the conditions on \( \pi_j(\cdot) \), \( \pi_j(1) \geq n^{-c_1} \) that yields
\[
2\sigma_n^2(1 + 1/\gamma) \ln \left( \pi_j^{-1}(1)n\sqrt{1 + \gamma} \right) = O(\sigma_n^2 \ln n) = O(\sigma_n^2 m\eta_n^2)
\]
for \( \eta_n \geq \sqrt{n^{-1} \ln n} \).

II. \(|m/e| < m_0^* \leq m\).

Apply Theorem 2 for \( J_0^c = \{1, ..., m\} \) (or, equivalently, \( J_0 = \emptyset \)) and \( h_j = 1 \) for \( j \in J_0^c \):
\[
\sum_{j=1}^{m} E||\hat{\mu}_j - \mu_j||_2^2 \leq c_1(\gamma) \left\{ \sum_{j \in J_0^c} \min_{1 \leq h_j \leq n} \left( \sum_{i=h_j+1}^{n} \mu_{\Theta_j}^2(ij) + Pen_j(h_j) \right) + \sum_{j \in J_0^c} Pen_j(1) + Pen_0(m) \right\} + c_2(\gamma)\sigma_n^2(1 - \pi_0(0)),
\]
where the conditions on \( \pi_j(1) \) and \( \pi_0(m) \) imply \( \sum_{j \in J_0^c} Pen_j(1) = O(\sigma_n^2 m \ln n) \) and \( Pen_0(m) = O(R(\Theta_j[\eta_j n])) \), \( j \in J_0^c \) and, therefore, the first term \( \sum_{j \in J_0^c} \) in the RHS of (21) is dominating for \( m_0^* \sim m \).

\[\Box\]

**Proof of Theorems 3 and 4**

The ideas of the proofs of both theorems on the minimax lower bounds are similar and can be combined.

Note first that any estimator cannot perform better than an oracle who knows the true \( J_0 \). In this (ideal) case one would obviously set \( \hat{\mu}_j \equiv 0 \) for all \( j \in J_0 \) with zero risk and, therefore, due to the additivity of the risk function,
\[
\inf_{\mu_1, ..., \mu_m} \sup_{\Theta_j[\eta_j n], J_0^c} \sum_{j=1}^{m} E||\hat{\mu}_j - \mu_j||_2^2 \geq C \sum_{j \in J_0^c} R(\Theta_j[\eta_j n])
\]
for any \( \Theta_j[\eta_j n] \) (see, e.g., Johnstone, 2011, Proposition 4.14).

Furthermore, following Case II in the proof of Theorem 2, \( \sum_{j \in J_0^c} R(\Theta_j[\eta_j n]) \) dominates over \( \sigma_n^2 m_0 \ln(m/m_0) \) in (15) and (16) for \( m_0 > m/2 \). To complete the proof we need to show, therefore, that for \( m_0 \leq m/2 \), the minimal unavoidable price for not being an oracle for selecting nonzero \( \mu_j \)’s is of order \( \sigma_n^2 m_0 \ln(m/m_0) \).

The main idea of the proof is to find a subset \( M_{m_0} \) of \( n \times m \) vectors \( \mu = (\mu_{11}, ..., \mu_{m_1}, ..., \mu_{1m}, ..., \mu_{nm})' \) with \( m_0 \) nonzero \( \mu_j = (\mu_{1j}, ..., \mu_{nj})' \in \Theta_j[\eta_j n] \) such that for any pair \( \mu_1, \mu_2 \in M_{m_0} \) and some \( C > 0 \), \( ||\mu_1 - \mu_2||_2^2 \geq C\sigma_n^2 m_0 \ln(m/m_0) \), while the Kullback-Leibler divergence \( K(\mathbb{P}_\mu_1, \mathbb{P}_\mu_2) = ||\mu_1 - \mu_2||_2^2/(2\sigma_n^2) \leq (1/16) \ln \text{card}(M_{m_0}) \). The result will then follow immediately from Lemma A.1 of Bunea, Tsybakov & Wegkamp (2007).

Define the subset \( \hat{D}_{m_0} \) of all \( m \)-dimensional indicator vectors with \( m_0 \) entries of ones, that is \( \hat{D}_{m_0} = \{d : d \in \{0, 1\}^m, ||d||_0 = m_0\} \). By Lemma A.3 of Rigollet & Tsybakov (2011), for
There exists a subset \( \mathcal{D}_{m_0} \subset \tilde{\mathcal{D}}_{m_0} \) such that for some constant \( \tilde{c} > 0 \), \( \ln \text{card}(\mathcal{D}_{m_0}) \geq \tilde{c} m_0 \ln(m/m_0) \), and for any pair \( \mathbf{d}_1, \mathbf{d}_2 \in \mathcal{D}_{m_0} \), the Hamming distance \( \rho(\mathbf{d}_1, \mathbf{d}_2) = \sum_{j=1}^{m} I\{d_{1j} \neq d_{2j}\} \geq \tilde{c} m_0 \).

To any indicator vector \( \mathbf{d} \in \mathcal{D}_{m_0} \) assign the corresponding mean vector \( \mu \in \mathcal{M}_{m_0} \) as follows. Let \( \tilde{C}^2 = (1/16)\sigma_n^2 \tilde{c} \ln(m/m_0) \). Define \( \mu_j = (\tilde{C}, 0, ..., 0)' I\{d_j = 1\} \) for \( 0 \leq p_j < 2 \) and \( \mu_j = (\tilde{C} n^{-1/2}, \tilde{C} n^{-1/2}, ..., \tilde{C} n^{-1/2})' I\{d_j = 1\} \) for \( p_j \geq 2, j = 1, ..., m \). Hence, \( \text{card}(\mathcal{M}_{m_0}) = \text{card}(\mathcal{D}_{m_0}) \).

Obviously, the resulting \( \mu_j \in l_0[\eta_{jn}] \) and a straightforward calculus shows that under the additional constraint on \( \eta_{jn} \) and \( m_0 \) in Theorem 4, \( \mu_j \in l_p[\eta_{jn}] \).

For any \( \mu^1, \mu^2 \in \mathcal{M}_{m_0} \) and the corresponding \( \mathbf{d}_1, \mathbf{d}_2 \in \mathcal{D}_{m_0} \), we then have

\[
||\mu^1 - \mu^2||_2^2 = \tilde{C}^2 \sum_{j=1}^{m} I\{\mathbf{d}_1j \neq \mathbf{d}_2j\} \geq \tilde{C}^2 \tilde{c} m_0 = (1/16)\sigma_n^2 \tilde{c}^2 m_0 \ln(m/m_0)
\]

and

\[
K(\mathbb{P}_{\mu^1}, \mathbb{P}_{\mu^2}) = \frac{\tilde{C}^2}{2\sigma_n^2} \sum_{j=1}^{m} I\{\mathbf{d}_1j \neq \mathbf{d}_2j\} \leq \frac{\tilde{C}^2 m_0}{\sigma_n^2} \leq (1/16) \ln \text{card}(\mathcal{M}_{m_0})
\]

\qed

References

[1] Abramovich, F. and Grinshtein, V. (2010). MAP model selection in Gaussian regression. \textit{Electron. J. Stat.} \textbf{4}, 932–949.

[2] Abramovich, F., Grinshtein, V. and Pensky, M. (2007). On optimality of Bayesian estimation in the normal means problem. \textit{Ann. Statist.} \textbf{35}, 2261–2286.

[3] Abramovich, F., Grinshtein, V., Petsa, A. and Sapatinas, T. (2010). On Bayesian estimation and its application to wavelet thresholding. \textit{Biometrika} \textbf{97}, 181–198.

[4] Angelini, C., De Canditiis, D., Mutarelli, M. and Pensky, M. (2007). A Bayesian approach to estimation and testing in time-course microarray experiments. \textit{Statist. Appl. Genetics & Molecul. Biology} \textbf{6}, 1–30.

[5] Antoniadis, A. and Fan, J. (2001). Regularization of wavelet approximations. \textit{J. Amer. Statist. Assoc.} \textbf{96}, 939–955.

[6] Birgé, L. and Massart, P. (2001). Gaussian model selection. \textit{J. Eur. Math. Soc.} \textbf{3}, 203–268.

[7] Bunea, F., Tsybakov, A. and Wegkamp, M.H. (2007). Aggregation for Gaussian regression. \textit{Ann. Statist.} \textbf{35}, 1674–1697.

[8] Donoho, D. L. and Johnstone, I. M. (1994a). Ideal spatial adaptation via wavelet shrinkage. \textit{Biometrika} \textbf{81}, 425-455.
[9] Donoho, D. L. and Johnstone, I. M. (1994b). Minimax risk over \( \ell_p \)-balls for \( \ell_q \)-error. *Probab. Theory Relat. Flds* **99**, 277-303.

[10] Donoho, D. L., Johnstone, I. M., Hoch, C. and Stern, A. (1992). Maximum entropy and the nearly black object. *J. R. Statist. Soc. B* **54**, 41-81 (with discussion).

[11] Friedman, J., Hastie, T. and Tibshirani, R. (2010). A note on the group lasso and a sparse group lasso. *Technical Report*, [http://www-stat.stanford.edu/~tibs/research.html](http://www-stat.stanford.edu/~tibs/research.html).

[12] Johnstone, I.M. (1994). Minimax Bayes, asymptotic minimax and sparse wavelet priors. *in* S. Gupta & J. Berge, ed., *Statistical Decision Theory and Related Topics*, V, Springer Verlag, 5-14.

[13] Johnstone, I.M. (2011). *Gaussian Estimation: Sequence and Multiresolution Models*. [http://www-stat.stanford.edu/~imj/Book100611.pdf](http://www-stat.stanford.edu/~imj/Book100611.pdf)

[14] Lin, Y. and Zhang, H.H. (2006). Component selection and smoothing in multivariate nonparametric regression. *Ann. Statist.* **34**, 2272–2297.

[15] Meier, L., van de Geer, S. and Buhlmann, P. (2009). High-dimensional additive modelling. *Ann. Statist.* **37**, 3779–3821.

[16] Raskutti, G., Wainwright, M.J. and Yu, B. (2011). Minimax rates of estimations for high-dimensional regression over \( l_q \) balls. *IEEE Trans. Inform. Theory* **57**, 6976–6694.

[17] Raskutti, G., Wainwright, M.J. and Yu, B. (2012). Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *J. Mach. Learn. Research*, to appear.

[18] Ravikumar, P., Lafferty, J., Liu, H. and Wasserman, L. (2009). Sparse additive models. *J.R. Statist. Soc. B* **71**, 1009-1030.

[19] Rigollet, P. and Tsybakov, A. (2011). Exponential screening and optimal rates of sparse estimation. *Ann. Statist.* **39**, 731-771.

[20] Simon, N., Friedman, J., Hastie, T. and Tibshirani, R. (2011). A sparse-group lasso. *Technical Report, Department of Statistics, Stanford University*, [http://www-stat.stanford.edu/~n Simon/SGLpaper.pdf](http://www-stat.stanford.edu/~n Simon/SGLpaper.pdf).

[21] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. R. Statist. Soc. B* **58**, 267–288.

[22] Wu, Z. and Zhou, H. (2012). Model selection and sharp asymptotic minimaxity. *Prob. Theory Relat. Fields*, to appear.
[23] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J.R. Statist. Soc. B* 68, 49–67.

[24] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J.R. Statist. Soc. B* 67, 301–320.