Noname manuscript No.
(will be inserted by the editor)

ℓ_1-Penalization for Mixture Regression Models

Nicolas Städler, Peter Bühlmann & Sara van de Geer

Received: date / Accepted: date

Abstract We consider a finite mixture of regressions (FMR) model for high-dimensional inhomogeneous data where the number of covariates may be much larger than sample size. We propose an ℓ_1-penalized maximum likelihood estimator in an appropriate parameterization. This kind of estimation belongs to a class of problems where optimization and theory for non-convex functions is needed. This distinguishes itself very clearly from high-dimensional estimation with convex loss- or objective functions, as for example with the Lasso in linear or generalized linear models. Mixture models represent a prime and important example where non-convexity arises.

For FMR models, we develop an efficient EM-algorithm for numerical optimization with provable convergence properties. Our penalized estimator is numerically better posed (e.g. boundedness of the criterion function) than unpenalized maximum likelihood estimation, and it allows for effective statistical regularization including variable selection. We also present some asymptotic theory and oracle inequalities: due to non-convexity of the negative log-likelihood function, different mathematical arguments are needed than for problems with convex losses. Finally, we apply the new method to both simulated and real data.

Keywords Adaptive Lasso · Finite mixture models · Generalized EM algorithm · High-dimensional estimation · Lasso · Oracle inequality

1 Introduction

In applied statistics, a tremendous amount of applications deal with relating a random response variable Y to a set of explanatory variables or covariates $X = (X^{(1)}, \ldots, X^{(p)})$ through a regression-type model. The homogeneity assumption that the regression coefficients are the same for different observations $(Y_1, X_1), \ldots, (Y_n, X_n)$ is often inadequate. Parameters may change for different subgroups of observations. Such heterogeneity can be modeled with a Finite Mixture of Regressions (FMR) model. Especially with high-dimensional data, where the number of covariates p is much larger than sample size n, the homogeneity assumption seems rather restrictive: at least a fraction of covariates may exhibit a different influence on the response among various observations (i.e. sub-populations). Hence, addressing the issue of heterogeneity in high-dimensional data is important in many practical applications. We will empirically demonstrate with real data in Section 7.2 that substantial prediction improvements are possible by incorporating a heterogeneity structure to the model.

Nicolas Städler · Peter Bühlmann · Sara van de Geer
Seminar für Statistik, ETH Zürich, CH-8092 Zürich, Switzerland.
E-mail: staedler@stat.math.ethz.ch, buehlmann@stat.math.ethz.ch, geer@stat.math.ethz.ch.

We propose here an ℓ_1-penalized method, i.e. a Lasso-type estimator (Tibshirani, 1996), for estimating a high-dimensional Finite Mixture of Regressions (FMR) model where $p \gg n$. Our procedure is related to the proposal in Khalili and Chen (2007). However, we argue that a different parameterization leads to more efficient computation in high-dimensional optimization for which we prove numerical convergence properties. Our algorithm can easily handle problems where p is in the thousands. Furthermore, regarding statistical performance, we present an oracle inequality which includes the setting where $p \gg n$: this is very different from Khalili and Chen (2007) who use fixed p asymptotics in the low-dimensional framework. Our theory for deriving oracle inequalities in the presence of non-convex loss functions, as the negative log-likelihood in a mixture model is non-convex, is rather non-standard but sufficiently general to cover other cases than FMR models.

From a more general point of view, we show in this paper that high-dimensional estimation problems with non-convex loss functions can be addressed with high computational efficiency and good statistical accuracy. Regarding the computation, we develop a rather generic block coordinate descent generalized EM-algorithm which is surprisingly fast even for large p. Progress in efficient gradient descent methods build on various developments by Tseng (2001) and Tseng and Yun (2008), and their use for solving Lasso-type convex problems has been worked out by e.g. Fu (1998), Friedman et al (2007), Meier et al (2008) and Friedman et al (2008). We present in Section 7.3 some computation times for the more involved case with non-convex objective function using a block coordinate descent generalized EM-algorithm. Regarding statistical theory, almost all results for high-dimensional Lasso-type problems have been developed for convex loss functions, e.g. the squared error in a Gaussian regression (Greenshtein and Ritov, 2004; Meinshausen and Bühlmann, 2006; Zhao and Yu, 2006; Bunea et al, 2007; Zhang and Huang, 2008; Meinshausen and Yu, 2009; Wainwright, 2009; Bickel et al, 2009; Cai et al, 2009b; Candès and Plan, 2009; Zhang, 2009b) or the negative log-likelihood in a generalized linear model (van de Geer, 2008). We present a non-trivial modification of the mathematical analysis of ℓ_1-penalized estimation with convex loss to non-convex but smooth likelihood problems.

When estimation is defined via optimization of a non-convex objective function, there is a major gap between the actual computation and the procedure studied in theory. The computation is typically guaranteed to find a local optimum of the objective function only, whereas the theory is usually about the estimator defined by a global optimum. Particularly in high-dimensional problems, it is difficult to compute a global optimum and it would be desirable to have some theoretical properties of estimators arising from local optima. We do not provide an answer to this difficult issue in this paper. The beauty of e.g. the Lasso or the Dantzig selector (Candès and Tao, 2007) in high-dimensional problems is the provable correctness or optimality of the estimator which is actually computed. A challenge for future research is to establish such provable correctness of estimators involving non-convex objective functions. A noticeable exception is presented in Zhang (2009a) for linear models, where some theory is derived for an estimator based on a local optimum of a non-convex optimization criterion.

The rest of this article is mainly focusing on Finite Mixture of Regressions (FMR) models. Some theory for high-dimensional estimation with non-convex loss functions is presented in Section 6 for more general settings than FMR models. The further organization of the paper is as follows: Section 2 describes the FMR model with an appropriate parameterization, Section 3 introduces ℓ_1-penalized maximum-likelihood estimation, Sections 4 and 5 present mathematical theory for the low- and high-dimensional case, Section 6 develops some efficient generalized EM algorithm and describes its numerical convergence properties and Section 7 reports on simulations, real data analysis and computational timings.

2 Finite mixture of Gaussian regressions model

Our primary focus is on the following mixture model involving Gaussian components:

$$\begin{aligned}
&Y_i|X_i \text{ independent for } i=1,\dots,n,\\
&Y_i|X_i = x \sim f_\xi(y|x)dy \text{ for } i=1,\dots,n,\\
&f_\xi(y|x) = \sum_{r=1}^{k} \pi_r \frac{1}{\sqrt{2\pi}\sigma_r} \exp(-\frac{(y-x^T\beta_r)^2}{2\sigma_r^2}),\\
&\xi = (\beta_1,\dots,\beta_k,\sigma_1,\dots,\sigma_k,\pi_1,\dots,\pi_{k-1}) \in \mathbb{R}^{kp}\times\mathbb{R}^k_{>0}\times\Pi,\\
&\Pi = \{\pi;\pi_r>0 \text{ for } r=1,\dots,k-1 \text{ and } \sum_{r=1}^{k-1}\pi_r<1\}.
\end{aligned} \tag{2.1}$$

Thereby, $X_i \in \mathbb{R}^p$ are fixed or random covariates, $Y_i \in \mathbb{R}$ is a univariate response variable and $\xi = (\beta_1,\dots,\beta_k,\sigma_1,\dots,\sigma_k,\pi_1,\dots,\pi_{k-1})$ denotes the $(p+2)\cdot k-1$ free parameters and π_k is given by $\pi_k = 1-\sum_{r=1}^{k-1}\pi_r$. The model in (2.1) is a mixture of Gaussian regressions, where every component r has its individual vector of regressions coefficients β_r and error variances σ_r^2. We are particularly interested in the case where $p \gg n$.

2.1 Reparameterized mixture of regressions model

We prefer to work with a reparameterized version of model (2.1) whose penalized maximum likelihood estimator is scale-invariant and easier to compute. The computational aspect will be discussed in greater detail in Sections 3.1 and 6. Define new parameters

$$\phi_r = \beta_r/\sigma_r, \quad \rho_r = \sigma_r^{-1}, \quad r=1,\dots,k.$$

This yields a one-to-one mapping from ξ in (2.1) to a new parameter vector
$\theta = (\phi_1,\dots,\phi_k,\rho_1,\dots,\rho_k,\pi_1,\dots,\pi_{k-1})$ and the model (2.1) in reparameterized form then equals:

$$\begin{aligned}
&Y_i|X_i \text{ independent for } i=1,\dots,n,\\
&Y_i|X_i = x \sim h_\theta(y|x)dy \text{ for } i=1,\dots,n,\\
&h_\theta(y|x) = \sum_{r=1}^{k} \pi_r \frac{\rho_r}{\sqrt{2\pi}} \exp(-\frac{1}{2}(\rho_r y - x^T\phi_r)^2)\\
&\theta = (\phi_1,\dots,\phi_k,\rho_1,\dots,\rho_k,\pi_1,\dots,\pi_{k-1}) \in \mathbb{R}^{kp}\times\mathbb{R}^k_{>0}\times\Pi\\
&\Pi = \{\pi;\pi_r>0 \text{ for } r=1,\dots,k-1 \text{ and } \sum_{r=1}^{k-1}\pi_r<1\},
\end{aligned} \tag{2.2}$$

with $\pi_k = 1-\sum_{r=1}^{k-1}\pi_r$. This is the main model we are analyzing and working with.

The log-likelihood function of this model equals:

$$\ell(\theta;Y) = \sum_{i=1}^{n} \log\left(\sum_{r=1}^{k} \pi_r \frac{\rho_r}{\sqrt{2\pi}} \exp(-\frac{1}{2}(\rho_r Y_i - X_i^T\phi_r)^2)\right). \tag{2.3}$$

Since we want to deal with the $p \gg n$ case, we have to regularize the maximum likelihood estimator (MLE) in order to obtain reasonably accurate estimates. We propose below some ℓ_1-norm penalized MLE which is different from a naive ℓ_1-norm penalty for the MLE in the non-transformed model (2.1). Furthermore, it is well known that the (log-) likelihood function is generally unbounded. We will see in Section 3.2 that our penalization will mitigate this problem.

3 ℓ_1-norm penalized maximum likelihood estimator

We argue first for the case of a (non-mixture) linear model why the reparameterization above in Section 2.1 is useful and quite natural.

3.1 ℓ_1-norm penalization for reparameterized linear models

Consider a Gaussian linear model

$$\begin{aligned} Y_i &= \sum_{j=1}^{p} \beta_j X_i^{(j)} + \varepsilon_i, \ i = 1, \ldots, n, \\ \varepsilon_1, \ldots, \varepsilon_n &\text{ i.i.d. } \sim \mathcal{N}(0, \sigma^2), \end{aligned} \tag{3.4}$$

where X_i are either fixed or random covariates. In short, we often write

$$Y = \mathbf{X}\beta + \varepsilon,$$

with $n \times 1$ vectors Y and ε, $p \times 1$ vector β and $n \times p$ matrix $\mathbf{X}$. In the sequel, $\|.\|$ denotes the Euclidean norm. The ℓ_1-norm penalized estimator, called the Lasso (Tibshirani (1996)), is defined as:

$$\hat{\beta}_\lambda = \text{argmin}_\beta n^{-1} \|Y - \mathbf{X}\beta\|^2 + \lambda \sum_{j=1}^{p} |\beta_j|. \tag{3.5}$$

Here λ is a non-negative regularization parameter. The Gaussian assumption is not crucial in model (3.4) but it is useful to make connections to the likelihood framework. The Lasso estimator in (3.5) is equivalent to minimizing the penalized negative log-likelihood $n^{-1}\ell(\beta; Y_1, \ldots, Y_n)$ as a function of the regression coefficients β and using the ℓ_1-penalty $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$: equivalence means here that we obtain the same estimator for a potentially different tuning parameter. But the Lasso estimator in (3.5) does not provide an estimate of the nuisance parameter σ^2.

In mixture models, it will be crucial to have a good estimator of σ^2 and the role of the scaling of the variance parameter is much more important than in homogeneous regression models. Hence, it is important to take σ^2 into the definition and optimization of the penalized maximum likelihood estimator: we could proceed with the following estimator,

$$\begin{aligned} \hat{\beta}_\lambda, \hat{\sigma}_\lambda^2 &= \text{argmin}_{\beta,\sigma^2}(-n^{-1}\ell(\beta, \sigma^2; Y_1, \ldots, Y_n) + \lambda\|\beta\|_1) \\ &= \text{argmin}_{\beta,\sigma^2}(\log(\sigma) + \|Y - \mathbf{X}\beta\|^2/(2n\sigma^2) + \lambda\|\beta\|_1). \end{aligned} \tag{3.6}$$

Note that we are penalizing only the β-parameter. However, the scale parameter estimate $\hat{\sigma}_\lambda^2$ is influenced indirectly by the amount of shrinkage λ.

There are two main drawbacks of the estimator in (3.6). First, it is not equivariant (Lehmann, 1983) under scaling of the response. More precisely, consider the transformation

$$Y' = bY, \quad \beta' = b\beta, \quad \sigma' = b\sigma \quad (b > 0) \tag{3.7}$$

which which leaves model (3.4) invariant. A reasonable estimator based on transformed data Y' should lead to estimators $\hat{\beta}', \hat{\sigma}'$ which are related to $\hat{\beta}, \hat{\sigma}$ through $\hat{\beta}' = b\hat{\beta}$ and $\hat{\sigma}' = b\hat{\sigma}$. This is not the case for the estimator in (3.6). Secondly, and as important as the first issue, the optimization in (3.6) is non-convex and hence, some of the major computational advantages of Lasso for high-dimensional problems is lost. We address these drawbacks by using the penalty term $\lambda\frac{\|\beta\|_1}{\sigma}$ leading to the following estimator:

$$\hat{\beta}_\lambda, \hat{\sigma}_\lambda^2 = \text{argmin}_{\beta,\sigma^2}(\log(\sigma) + \|Y - \mathbf{X}\beta\|^2/(2n\sigma^2) + \lambda\frac{\|\beta\|_1}{\sigma}).$$

This estimator is equivariant under the scaling transformation (3.7), i.e. the estimators $\hat{\beta}', \hat{\sigma}'$ based on Y' transform as $\hat{\beta}' = b\hat{\beta}$ and $\hat{\sigma}' = b\hat{\sigma}$. Furthermore, it penalizes the ℓ_1-norm of the coefficients and small variances σ^2 simultaneously which has some close relations to the Bayesian Lasso (Park and Casella, 2008). For the latter, a Bayesian approach is used with a conditional Laplace prior specification of the form

$$p(\beta|\sigma^2) = \prod_{j=1}^{p} \frac{\lambda}{2\sqrt{\sigma^2}} \exp(-\lambda \frac{|\beta_j|}{\sqrt{\sigma^2}})$$

and a noninformative scale-invariant marginal prior $p(\sigma^2) = 1/\sigma^2$ for σ^2. Park and Casella (2008) argue that conditioning on σ^2 is important, because it guarantees a unimodal full posterior.

Most importantly, we can re-parameterize to achieve convexity of the optimization problem:

$$\phi_j = \beta_j/\sigma, \quad \rho = \sigma^{-1}.$$

This then yields the following estimator which is equivariant under scaling and whose computation involves convex optimization:

$$\hat{\phi}_\lambda, \hat{\rho}_\lambda = \arg\min_{\phi,\rho}(-\log(\rho) + \frac{1}{2n}||\rho Y - X\phi||^2 + \lambda||\phi||_1). \tag{3.8}$$

From an algorithmic point of view, fast algorithms are available to solve the optimization in (3.8). Shooting algorithms (Fu, 1998) with coordinate-wise descent are especially suitable, as demonstrated by e.g. Friedman et al (2007), Meier et al (2008) or Friedman et al (2008). We describe in Section 6.1 an algorithm for estimation in a mixture of regressions model, a more complex task than the optimization for (3.8). As we will see in Section 6.1, we will make use of the Karush-Kuhn-Tucker (KKT) conditions in the M-step of a generalized EM-algorithm. For the simpler criterion in (3.8) for a non-mixture model, the KKT conditions imply the following which we state without a proof. Denote by $\langle \cdot, \cdot \rangle$ the inner product in n-dimensional Euclidean space.

Proposition 1 *Every solution $(\hat{\phi}, \hat{\rho})$ of (3.8) satisfies:*

$$\begin{aligned} -\hat{\rho}X_j^T Y + X_j^T X\hat{\phi} + n\lambda sign(\hat{\phi}_j) &= 0 \quad \text{if} \quad \hat{\phi}_j \neq 0, \\ |-\hat{\rho}X_j^T Y + X_j^T X\hat{\phi}| &\leq n\lambda \quad \text{if} \quad \hat{\phi}_j = 0, \end{aligned}$$

and

$$\hat{\rho} = \frac{\langle Y, X\hat{\phi}\rangle + \sqrt{\langle Y, X\hat{\phi}\rangle^2 + 4||Y||^2 n}}{2||Y||^2}.$$

3.2 ℓ_1-norm penalized MLE for mixture of Gaussian regressions

Consider the mixture of Gaussian regressions model in (2.2). Assuming that p is large, we want to regularize the MLE. In the spirit of the approach in (3.8), we propose for the unknown parameter $\theta = (\phi_1, \ldots, \phi_k, \rho_1, \ldots, \rho_k, \pi_1, \ldots, \pi_{k-1})$ the estimator:

$$\hat{\theta}_\lambda^{(\gamma)} = \arg\min_{\theta \in \Theta} -n^{-1}\ell_{pen,\lambda}^{(\gamma)}(\theta), \quad \theta = (\phi_1, \ldots, \phi_k, \rho_1, \ldots, \rho_k, \pi_1, \ldots, \pi_{k-1}), \tag{3.9}$$

$$\begin{aligned} -n^{-1}\ell_{pen,\lambda}^{(\gamma)}(\theta) = -n^{-1}\sum_{i=1}^{n} \log\left(\sum_{r=1}^{k} \pi_r \frac{\rho_r}{\sqrt{2\pi}} \exp(-\frac{1}{2}(\rho_r Y_i - X_i^T\phi_r)^2)\right) \\ + \lambda \sum_{r=1}^{k} \pi_r^\gamma ||\phi_r||_1, \end{aligned} \tag{3.10}$$

$$\Theta = \mathbb{R}^{kp} \times \mathbb{R}_{>0}^k \times \Pi, \tag{3.11}$$

where $\Pi = \{\pi; \pi_r > 0 \text{ for } r = 1, \ldots, k-1 \text{ and } \sum_{r=1}^{k-1} \pi_r < 1\}$ with $\pi_k = 1 - \sum_{r=1}^{k-1} \pi_r$. The value of $\gamma \in \{0, 1/2, 1\}$ parameterizes three different penalties.

The first penalty function with $\gamma = 0$ is independent of the component probabilities π_r. As we will see in Sections 6.1 and 6.4, the optimization for computing $\hat{\theta}_\lambda^{(0)}$ is easiest and we establish a rigorous result about numerical convergence of a generalized EM algorithm. The penalty with $\gamma = 0$ works fine if the components are not very unbalanced, i.e. the true π_r's aren't too different. In case of strongly unbalanced components, the penalties with values $\gamma \in \{1/2, 1\}$ are to be preferred, at the price of having to pursue a more difficult optimization problem. The value of $\gamma = 1$ has been proposed by Khalili and Chen (2007) for the naively parameterized likelihood from model (2.1). We will report in Section 7.1 about empirical comparisons with the three different penalties involving $\gamma \in \{0, 1/2, 1\}$.

All three penalty functions involve the ℓ_1-norm of the component specific ratio's $\phi_r = \frac{\beta_r}{\sigma_r}$ and hence small variances are penalized. The penalized criteria therefore stay finite whenever $\sigma_r \to 0$: this is in sharp contrast to the unpenalized MLE where the likelihood tends to infinity if $\sigma_r \to 0$, see for example (McLachlan and Peel, 2000).

Proposition 2 *Assume that $Y_i \neq 0$ for all $i = 1, \ldots, n$. Then the penalized negative likelihood $-n^{-1}\ell_{pen,\lambda}^{(0)}(\theta)$ is bounded from below for all values $\theta \in \Theta$ from (3.11).*

A proof is given in Appendix C. Even though Proposition 2 is only stated and proved for the penalized negative likelihood with $\gamma = 0$ we expect that the statement is also true for $\gamma = 1/2$ or 1.

Due to the ℓ_1-norm penalty, the estimator is shrinking some of the components of $\phi_1, \ldots, \phi_k$ exactly to zero, depending on the magnitude of the regularization parameter λ. Thus, we can do variable selection as follows. Denote by

$$\widehat{S} = \Big\{(r, j);\ 1 \le r \le k,\ 1 \le j \le p, \hat{\phi}_{r,j} \neq 0\Big\}. \tag{3.12}$$

The set $\widehat{S}$ denotes the collection of non-zero estimated, i.e. selected, regression coefficients in the k mixture components. Note that no significance testing is involved, but of course, $\widehat{S} = \widehat{S}_\lambda^{(\gamma)}$ depends on the specification of the regularization parameter λ and the type of penalty described by γ.

3.3 Adaptive ℓ_1-norm penalization

A two-stage adaptive ℓ_1-norm penalization for linear models has been proposed by Zou (2006), called the adaptive Lasso. It is an effective way to address some bias problems of the (one-stage) Lasso which may employ strong shrinkage of coefficients corresponding to important variables.

The two-stage adaptive ℓ_1-norm penalized estimator for a mixture of Gaussian regressions is defined as follows. Consider an initial estimate θ^{ini}, for example from the estimator in (3.9). The adaptive criterion to be minimized involves a re-weighted ℓ_1-norm penalty term:

$$\begin{aligned} -n^{-1}\ell_{adapt}^{(\gamma)}(\theta) &= -n^{-1}\sum_{i=1}^{n}\log\left(\sum_{r=1}^{k}\pi_r\frac{\rho_r}{\sqrt{2\pi}}\exp(-\frac{1}{2}(\rho_r Y_i - X_i^T\phi_r)^2)\right) \\ &\quad + \lambda\sum_{r=1}^{k}\pi_r^\gamma\sum_{j=1}^{p}w_{r,j}|\phi_{r,j}|, \\ w_{r,j} &= \frac{1}{|\phi_{r,j}^{ini}|}, \quad \theta = (\rho_1, \ldots, \rho_k, \phi_1, \ldots, \phi_k, \pi_1, \ldots, \pi_{k-1}), \end{aligned} \tag{3.13}$$

where $\gamma \in \{0, 1/2, 1\}$. The estimator is then defined as

$$\hat{\theta}_{adapt;\lambda}^{(\gamma)} = \arg\min_{\theta\in\Theta} -n^{-1}\ell_{adapt}^{(\gamma)}(\theta), \tag{3.14}$$

where Θ is as in (3.11).

The adaptive Lasso in linear models has better variable selection properties than the Lasso, see Zou (2006), Huang et al (2008), Zhou et al (2009). We present some theory for the adaptive estimator in FMR models in Section 4. Furthermore, we report some empirical results in Section 7.1 indicating that the two-stage adaptive method often outperforms the one-stage ℓ_1-penalized estimator.

3.4 Selection of the tuning parameters

The regularization parameters to be selected are the number of components k, the penalty parameter λ and we may also want to select the type of the penalty function, i.e. selection of γ.

One possibility is to use a modified BIC criterion which minimizes

$$\text{BIC} = -2\ell(\hat{\theta}^{(\gamma)}_{\lambda,k}) + log(n)\text{d}_\text{e}, \tag{3.15}$$

over a grid of candidate values for k, λ and maybe also γ. Here, $\hat{\theta}^{(\gamma)}_{\lambda,k}$ denotes the estimator in (3.9) using the parameters λ, k, γ in (3.10), and $-\ell(\cdot)$ is the negative log-likelihood. Furthermore, $\text{d}_\text{e} = k + (k-1) + \sum_{j=1\ldots p, r=1\ldots k} 1_{\{\hat{\phi}_{r,j}\neq 0\}}$ is the effective number of parameters (Pan and Shen, 2007).

Alternatively, we may use a cross-validation scheme for tuning parameter selection minimizing some cross-validated negative log-likelihood.

Regarding the grid of candidate values for λ, we consider $0 \leq \lambda_1 < \ldots < \lambda_M \leq \lambda_{max}$, where λ_{max} is given by

$$\lambda_{max} = \max_{j=1,\ldots,p} \left| \frac{\langle Y, X^{(j)} \rangle}{\sqrt{n}||Y||} \right|. \tag{3.16}$$

At λ_{max}, all coefficients $\hat{\phi}_j,\ (j = 1, \ldots, p)$ of the one-component model are exactly zero. Equation (3.16) easily follows from Proposition 1.

For the adaptive ℓ_1-norm penalized estimator minimizing the criterion in (3.13) we proceed analogously but replacing $\hat{\theta}^{(\gamma)}_{\lambda,k}$ in (3.15) by $\hat{\theta}^{(\gamma)}_{adapt;\lambda}$ in (3.14). As initial estimator in the adaptive criterion, we propose to use the estimate in (3.9) which is optimally tuned using the modified BIC or some cross-validation scheme.

4 Asymptotic properties for fixed p and k

Following the penalized likelihood theory of Fan and Li (2001), we establish first some asymptotic properties of the estimator in (3.10). We assume in this section that the number of covariates p and the number of mixture components k are fixed as sample size $n \to \infty$. Of course, this does not reflect a truly high-dimensional scenario, but the theory and methodology is much easier for this case. An extended theory for p potentially very large in relation to n is presented in Section 5.

Denote by θ_0 the true parameter.

Theorem 1 *(Consistency) Consider model (2.2) with fixed design and fixed p and k. If $\lambda = O(n^{-1/2})$ $(n \to \infty)$, then there exists a local minimizer $\hat{\theta}^{(\gamma)}_\lambda$ of $-n^{-1}\ell_{pen,\lambda}(\theta)$ in (3.10) ($\gamma \in \{0, 1/2, 1\}$) such that*

$$\sqrt{n}\left(\hat{\theta}^{(\gamma)}_\lambda - \theta_0\right) = O_P(1).$$

A proof is given in Appendix A. Theorem 1 can be easily misunderstood. It does not guarantee the existence of an asymptotically consistent sequence of estimates. The only claim is that a clairvoyant statistician (with pre-knowledge of θ_0) can choose a consistent sequence of roots in the neighborhood of θ_0 (van der Vaart, 2007). In the case where $-n^{-1}\ell_{pen,\lambda}(\theta)$ has a unique minimizer, which is the case for a FMR model with one component, the resulting estimator would be root-n consistent. But for a general FMR model with more than one component and typically several local minimizers this does not hold anymore. In this sense the preceding theorem might look better than it is.

Next, we present an asymptotic oracle result in the spirit of Fan and Li (2001) for the two-stage adaptive procedure described in Section 3.3. Denote by S the population analogue of (3.12), i.e. the set of non-zero regression coefficients. Furthermore, let $\theta_S = (\{\phi_{r,j};\ (r,j) \in S\}, \rho_1, \ldots, \rho_k, \pi_1, \ldots, \pi_{k-1})$ the sub-vector of parameters corresponding to the true non-zero regression coefficients (denoted by S) and analogously for $\hat{\theta}_S$.

Theorem 2 *(Asymptotic oracle result for adaptive procedure) Consider model (2.2) with fixed design and fixed p and k. If $\lambda = o(n^{-1/2})$, $n\lambda \to \infty$ and if θ^{ini} satisfies $\theta^{ini} - \theta_0 = O_P(n^{-1/2})$, then there exists a local minimizer $\hat{\theta}^{(\gamma)}_{adapt;\lambda}$ of $-n^{-1}\ell^{(\gamma)}_{adapt}(\theta)$ in (3.13) ($\gamma \in \{0, 1/2, 1\}$) which satisfies:*

1. Consistency in variable selection: $\mathbb{P}[\widehat{S}^{(\gamma)}_{adapt;\lambda} = S] \to 1$ $(n \to \infty)$.
2. Oracle Property: $\sqrt{n}\left(\hat{\theta}^{(\gamma)}_{adapt;\lambda,S} - \theta_{0,S}\right) \rightsquigarrow^d \mathcal{N}(0, I_S(\theta_0))$, *where $I_S(\theta_0)$ is the Fisher information knowing that $\theta_{S^c} = 0$ (i.e. the submatrix of the Fisher information at θ_0 corresponding to the variables in S).*

A proof is given in Appendix A. As in Theorem 1, the assertion of the Theorem is only making a statement about *some* local optimum. Furthermore this result only holds for the adaptive criterion with weights $w_{r,j} = \frac{1}{|\phi^{ini}_{r,j}|}$ coming from a root-n consistent initial estimator θ^{ini}: this is a rather strong assumption given the fact that Theorem 1 only ensures existence of such an estimator. The non-adaptive estimator wit the ℓ_1-norm penalty as in (3.10) cannot achieve sparsity and maintain root-n consistency due to the bias problem mentioned in Section 3.3 (see also Khalili and Chen (2007)).

5 General theory for high-dimensional setting with non-convex smooth loss

We present here some theory, entirely different from Theorems 1 and 2, which reflects some consistency and optimality behavior of the ℓ_1-norm penalized maximum likelihood estimator for the potentially high-dimensional framework with $p \gg n$. In particular, we derive some oracle inequality which is non-asymptotic. We intentionally present this theory for ℓ_1-penalized smooth likelihood problems which are generally non-convex: ℓ_1-penalized likelihood estimation in FMR models is then a special case discussed in Section 5.3. The following Sections 5.1 - 5.2 introduce some mathematical conditions and derive auxiliary results and an general oracle inequality (Theorem 3): the interpretation of these conditions and of the oracle result is discussed for the case of FMR models at the end of Section 5.3.1.

5.1 The setting and notation

Let $\{f_\psi;\ \psi \in \Psi\}$ be a collection of densities with respect to the Lebesgue measure μ on $\mathbb{R}$ (i.e. the range for the response variable). The parameter space Ψ is assumed to be a bounded subset of some finite-dimensional space, say

$$\Psi \subset \{\psi \in \mathbb{R}^d;\ \|\psi\|_\infty \leq K\},$$

where we have equipped (quite arbitrarily) the space $\mathbb{R}^d$ with the sup-norm $\|\psi\|_\infty = \max_{1\leq j\leq d} |\psi_j|$. In our setup, the dimension d will be regarded as a fixed constant (which still covers high-dimensionality of the covariates, as we will see). Then, equivalent metrics are e.g. the ones induced by the ℓ_q-norm $\|\psi\|_q = (\sum_{j=1}^d |\psi_j|_q)^{1/q}$ $(q \geq 1)$.

We observe a covariate $X \in \mathbb{R}^p$ and a response variable $Y \in \mathbb{R}$. The true conditional density of Y given $X = x$ is assumed to be equal to

$$f_{\psi_0}(\cdot|x) = f_{\psi_0(x)},$$

where

$$\psi_0(x) \in \Psi,\ \forall\ x \in \mathbb{R}^p.$$

That is, we assume that the true conditional density of Y given x is depending on x only through some parameter function $\psi_0(x)$. Of course, the introduced notation also applies to fixed instead of random covariates.

The parameter $\{\psi_0(x);\ x \in \mathbb{R}^p\}$ is assumed to have a nonparametric part of interest $\{g_0(x);\ x \in \mathbb{R}^p\}$ and a low-dimensional nuisance part η_0, i.e.,

$$\psi_0(\cdot)^T = (g_0(\cdot)^T, \eta_0^T),$$

with

$$g_0(x) \in \mathbb{R}^k,\ \forall\ x \in \mathbb{R}^p,\ \eta_0 \in \mathbb{R}^m,\ k + m = d.$$

In case of FMR models, $g(x)^T = (\phi_1^T x, \phi_2^T x, \ldots, \phi_k^T x)$ and η involves the parameters $\rho_1, \ldots, \rho_k, \pi_1, \ldots, \pi_{k-1}$. More details are given in Section 5.3.

With minus the log-likelihood as loss function, the so-called excess risk

$$\mathcal{E}(\psi|\psi_0) = -\int \log\left[\frac{f_\psi}{f_{\psi_0}}\right] f_{\psi_0} d\mu$$

is the Kullback-Leibler information. For fixed covariates $x_1, \ldots, x_n$, we define the average excess risk

$$\bar{\mathcal{E}}(\psi|\psi_0) = \frac{1}{n}\sum_{i=1}^n \mathcal{E}\Big(\psi(x_i)\Big|\psi_0(x_i)\Big),$$

and for random design, we take the expectation $\mathbb{E}(\mathcal{E}(\psi(X)|\psi_0(X))$.

5.1.1 The margin

Following Tsybakov (2004) and van de Geer (2008) we call the behavior of the excess risk $\mathcal{E}(\psi|\psi_0)$ near ψ_0 the margin. We will show in Lemma 1 that the margin is quadratic.

Denote by

$$l_\psi = \log f_\psi$$

the log-density. Assuming the derivatives exist, we define the score function

$$s_\psi = \frac{\partial l_\psi}{\partial \psi},$$

and the Fisher information

$$I(\psi) = \int s_\psi s_\psi^T f_\psi d\mu = -\int \frac{\partial^2 l_\psi}{\partial \psi \partial \psi^T} f_\psi d\mu.$$

Of course, we can then also look at $I(\psi(x))$ using the parameter function $\psi(x)$.

In the sequel, we introduce some conditions (Conditions 1 - 6). Their interpretation for the case of FMR models is given at the end of Section 5.3.1. First, we will assume boundedness of third derivatives.

Condition 1 *It holds that*

$$\sup_{\psi\in\Psi}\max_{(j_1,j_2,j_3)\in\{1,\ldots,d\}^3}\left|\frac{\partial^3}{\partial\psi_{j_1}\partial\psi_{j_2}\partial\psi_{j_3}}l_\psi(\cdot)\right|\leq G_3(\cdot),$$

where

$$\sup_x\int G_3(y)f_{\psi_0}(y|x)d\mu(y)\leq C_3<\infty.$$

For a symmetric, positive semi-definite matrix A, we let $\Lambda^2_{\min}(A)$ be its smallest eigenvalue.

Condition 2 *For all x, the Fisher information matrix $I(g_0(x),\eta_0)$ is positive definite, and in fact*

$$\Lambda_{\min}=\inf_x\Lambda_{\min}(I(\psi_0(x)))>0.$$

Further we will need the following identifiability condition.

Condition 3 *For all $\varepsilon>0$, there exists an $\alpha_\varepsilon>0$, such that*

$$\inf_x\inf_{\substack{\psi\in\Psi\\ \|\psi-\psi_0(x)\|_2>\varepsilon}}\mathcal{E}(\psi|\psi_0(x))\geq\alpha_\varepsilon.$$

Based on these three conditions we have the following result:

Lemma 1 *Assume Conditions 1, 2, and 3. Then*

$$\inf_x\frac{\mathcal{E}(\psi|\psi_0(x))}{\|\psi-\psi_0(x)\|_2^2}\geq\frac{1}{c_0^2},$$

where

$$c_0^2=\max\left[\frac{1}{\varepsilon_0},\frac{dK^2}{\alpha_{\varepsilon_0}}\right],\ \varepsilon_0=\frac{3\Lambda^2_{\min}}{2d^{3/2}}.$$

A proof is given in Appendix B.

5.1.2 The empirical process

We now specialize to the case where

$$\psi(x)^T=(g(x)^T,\eta^T),$$

where (with some abuse of notation)

$$\begin{aligned}&g(x)^T=g_\phi(x)^T=(g_1(x),\ldots,g_k(x)),\\&g_r(x)=g_{\phi_r}(x)=x^T\phi_r,\ x\in\mathbb{R}^p,\ \phi_r\in\mathbb{R}^p,\ r=1,\ldots,k.\end{aligned}$$

We also write

$$\psi_\vartheta(x)^T=(g_\phi(x)^T,\eta^T),\qquad\vartheta^T=(\phi_1^T,\ldots,\phi_k^T,\eta^T)$$

to make the dependence of the parameter function $\psi(x)$ on ϑ more explicit.

We will assume that

$$\sup_x\|\phi^Tx\|_\infty=\sup_x\max_{1\leq r\leq k}|\phi_r^Tx|\leq K.$$

Our parameter space is now

$$\tilde\Theta\subset\{\vartheta^T=(\phi_1^T,\ldots,\phi_k^T,\eta^T);\ \sup_x\|\phi^Tx\|_\infty\leq K,\ \|\eta\|_\infty\leq K\}.\tag{5.17}$$

Note that $\tilde{\Theta}$ is in principle $(pk+m)$-dimensional. The true parameter ϑ_0 is assumed to be an element of $\tilde{\Theta}$.

Let us define

$$L_\vartheta(x,\cdot) = \log f_{\psi(x)}(\cdot), \quad \psi(x)^T = \psi_\vartheta(x)^T = (g_\phi(x)^T, \eta^T), \ \vartheta^T = (\phi_1^T, \ldots, \phi_k^T, \eta^T),$$

and the empirical process for fixed covariates $x_1, \ldots, x_n$

$$V_n(\vartheta) = \frac{1}{n}\sum_{i=1}^n \left[L_\vartheta(x_i, Y_i) - \mathbb{E}\Big(L_\vartheta(x_i, Y) \Big| X = x_i \Big) \right].$$

We now fix some $T \geq 1$ and $\lambda_0 \geq 0$ and define the set

$$\mathcal{T} = \left\{ \sup_{\vartheta^T = (\phi^T, \eta^T) \in \tilde{\Theta}} \frac{\Big| V_n(\vartheta) - V_n(\vartheta_0) \Big|}{(\|\phi - \phi_0\|_1 + \|\eta - \eta_0\|_2) \vee \lambda_0} \leq T\lambda_0 \right\}. \tag{5.18}$$

5.2 Oracle inequality for the Lasso for non-convex loss functions

For an optimality result, we need some condition on the design. Denote the active set, i.e. the set of non-zero coefficients, by

$$S = \{(r,j);\ \phi_{r,j} \neq 0\}, \ s = |S|,$$

and let

$$\phi_J = \{\phi_{(r,j)};\ (r,j) \in J\}, \ J \subset \{1, \ldots, p\}^k.$$

Condition 4 *(Restricted eigenvalue condition). There exists a constant $\kappa \geq 1$, such that for all $\phi \in \mathbb{R}^{pk}$ satisfying*

$$\|\phi_{S^c}\|_1 \leq 6\|\phi_S\|_1,$$

it holds that

$$\|\phi_S\|_2^2 \leq \kappa^2 \sum_{r=1}^k \phi_r^T \Sigma_n \phi_r.$$

For $\psi(\cdot)^T = (g(\cdot)^T, \eta^T)$, we use the notation

$$\|\psi\|_{Q_n}^2 = \frac{1}{n}\sum_{i=1}^n \sum_{r=1}^k g_r^2(x_i) + \sum_{j=1}^m \eta_j^2.$$

We also write for $g(\cdot) = (g_1(\cdot), \ldots, g_k(\cdot))^T$,

$$\|g\|_{Q_n}^2 = \frac{1}{n}\sum_{i=1}^n \sum_{r=1}^k g_r^2(x_i).$$

Thus

$$\|g_\phi\|_{Q_n}^2 = \sum_{r=1}^k \phi_r^T \Sigma_n \phi_r,$$

and the bound in the restricted eigenvalue condition then reads

$$\|\phi_S\|_2^2 \leq \kappa^2 \|g_\phi\|_{Q_n}^2.$$

Bounding $\|g_\phi\|_{Q_n}^2$ in terms of $\|\phi_S\|_2^2$ can be done directly using e.g. the Cauchy-Schwarz inequality. The restricted eigenvalue condition ensures a bound in the other direction which itself is needed

for an oracle inequality. Some references about the restricted eigenvalue condition are provided at the end of Section 5.3.1.

We employ the Lasso-type estimator

$$\hat{\vartheta}^T = (\hat{\phi}^T, \hat{\eta}^T) = \underset{\vartheta^T=(\phi^T,\eta^T)\in\tilde{\Theta}}{\arg\min} \left\{ -\frac{1}{n}\sum_{i=1}^{n} L_\vartheta(x_i, Y_i) + \lambda \sum_{r=1}^{k} \|\phi_r\|_1 \right\}. \tag{5.19}$$

We omit in the sequel the dependence of $\hat{\vartheta}$ on λ. Note that we consider here a global minimizer: it may be difficult to compute if the empirical risk $n^{-1}\sum_{i=1}^n L_\vartheta(x_i, Y_i)$ is non-convex in ϑ. We then write $\|\phi\|_1 = \sum_{r=1}^k \|\phi_r\|_1$. We let

$$\hat{\psi}(x)^T = (g_{\hat{\phi}}(x)^T, \hat{\eta}^T),$$

which depends only on the estimate $\hat{\vartheta}$, and we denote by

$$\psi_0(x)^T = (g_{\phi_0}(x)^T, \eta_0^T).$$

Theorem 3 *(Oracle result for fixed design). Assume fixed covariates $x_1, \ldots, x_n$, Conditions 1-3 and 4, and that $\lambda \geq 2T\lambda_0$ for the estimator in (5.19) with T and λ_0 as in (5.18). Then on $\mathcal{T}$, defined in (5.18), for the average excess risk (average Kullback-Leibler loss),*

$$\bar{\mathcal{E}}(\hat{\psi}|\psi_0) + 2(\lambda - T\lambda_0)\|\hat{\phi}_{S^c}\|_1 \leq 8(\lambda + T\lambda_0)^2 c_0^2 \kappa^2 s,$$

where c_0 and κ are defined in Lemma 1 and Condition 4, respectively.

A proof is given in Appendix B. We will give an interpretation of this result in Section 5.3.1, where we specialize to FMR models. In the case of FMR models the probability of the set $\mathcal{T}$ is large as shown in detail by Lemma 3 below.

Before specializing to FMR models, we present more general results for lower bounding the probability of the set $\mathcal{T}$. We make the following assumptions.

Condition 5 *For the score function $s_\vartheta(\cdot) = s_{\psi_\vartheta}(\cdot)$ we have:*

$$\sup_{\vartheta\in\tilde{\Theta}} \|s_\vartheta(\cdot)\|_\infty \leq G_1(\cdot).$$

Condition 5 primarily has notational character. Later, in Lemma 2 and particular in Lemma 3, the function $G_1(\cdot)$ needs to be sufficiently regular to ensure small corresponding probabilities.

Let

$$\Sigma_n = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T,$$

and let $\Lambda^2_{\max}(\Sigma_n)$ be the largest eigenvalue of Σ_n.

Condition 6 *For a constant $\Lambda_{\max} < \infty$, it holds that $\Lambda_{\max}(\Sigma_n) \leq \Lambda_{\max}$.*

Condition 6 has again primarily notational character, and to avoid digressions, in what follows we shall not explicitly give the dependency on $\Lambda_{\max}(\Sigma_n)$. This is appropriate when there is a bound $\Lambda_{\max}$ that does not depend on p or n.

Define

$$\lambda_0 = M_n \sqrt{\frac{\log^3 n}{n}}. \tag{5.20}$$

As we will see, we usually choose $M_n \asymp \sqrt{\log(n)}$. Let $\mathbb{P}_{\mathbf{x}}$ denote the conditional probability given $(X_1, \ldots, X_n) = (x_1, \ldots, x_n) = \mathbf{x}$ and with the expression $1\{\cdot\}$ we denote the indicator function.

Lemma 2 *Assume Conditions 4 and 5. We have for constants c_1, c_2 and c_3 depending on $\Lambda_{\max}$, k, and K, and for all $T \geq 1$,*

$$\sup_{\vartheta^T=(\phi^T,\eta^T)\in\tilde{\Theta}} \frac{\Big|V_n(\vartheta) - V_n(\vartheta_0)\Big|}{(\|\phi-\phi_0\|_1 + \|\eta-\eta_0\|_2) \vee \lambda_0} \leq c_1 T \lambda_0,$$

with $\mathbb{P}_{\mathbf{x}}$ probability at least

$$1 - c_2 \exp\left[-\frac{T^2 \log^3 n}{c_3^2}\right] - \mathbb{P}_{\mathbf{x}}\left(\frac{1}{n}\sum_{i=1}^n F(Y_i) > T\lambda_0^2/(dK)\right).$$

where (for $i = 1, \ldots, n$)

$$F(Y_i) = G(Y_i)\mathrm{l}\{G_1(Y_i) > M_n\} + \mathbb{E}\Big(G_1(Y)\mathrm{l}\{G_1(Y) > M_n\}\Big|X = x_i\Big).$$

Regarding the constants λ_0 and K, see (5.20) and (5.17), respectively.

A proof is given in Appendix B.

5.3 FMR models

In the finite mixture of regressions model from (2.2) with k components, the parameter is $\vartheta^T = (\phi^T, \eta^T) = (\phi_1^T, \ldots, \phi_k^T, \log\rho_1, \ldots, \log\rho_k, \log\pi_1, \ldots, \log\pi_{k-1})$, where the $\rho_r = \sigma_r^{-1}$ are the inverse standard deviations in mixture component r and the π_r are the mixture coefficients. For mathematical convenience and simpler notation, we consider here the log-transformed ρ and π parameters in order to have lower and upper bounds for ρ and π. Obviously, there is a one-to-one correspondence between ϑ and θ from Section 2.1.

Let the parameter space be

$$\tilde{\Theta} \subset \{\vartheta^T;\ \sup_x \|\phi^T x\|_\infty \leq K, \|\log\rho\|_\infty \leq K, -K \leq \log\pi_1 \leq 0, \ldots, -K \leq \log\pi_{k-1} \leq 0, \sum_{r=1}^{k-1}\pi_r < 1\}, \tag{5.21}$$

and $\pi_k = 1 - \sum_{r=1}^{k-1}\pi_r$.

We consider the estimator

$$\hat{\vartheta}_\lambda = \arg\min_{\vartheta\in\tilde{\Theta}} -n^{-1}\sum_{i=1}^n \log\left(\sum_{r=1}^k \pi_r \frac{\rho_r}{\sqrt{2\pi}}\exp(-\frac{1}{2}(\rho_r Y_i - X_i^T\phi_r)^2)\right) + \lambda\sum_{r=1}^k ||\phi_r||_1. \tag{5.22}$$

This is the estimator from Section 3.2 with $\gamma = 0$. We emphasize the boundedness of the parameter space by using the notation $\tilde{\Theta}$. In contrast to Section 4, we focus here on any global minimizer of the penalized negative log-likelihood which is arguably difficult to compute.

In the following we transform the estimator $\hat{\vartheta}_\lambda$ to $\hat{\theta}_\lambda$ in the parameterization θ from Section 2.1. Using some abuse of notation we denote the average excess risk by $\bar{\mathcal{E}}(\hat{\theta}_\lambda|\theta_0)$.

5.3.1 Oracle result for FMR models

We specialize now our results from Section 5.2 to FMR models.

Proposition 3 *For fixed design FMR models as in (2.2) with $\tilde{\Theta}$ in (5.21), Conditions 1,2 and 3 are met, for appropriate C_3, $\Lambda_{\min}$ and $\{\alpha_\varepsilon\}$, depending on k and K. Also Condition 5 holds, with*

$$G_1(y) = \mathrm{e}^K|y| + K.$$

Proof This follows from straightforward calculations.

In order to show that the probability for the set $\mathcal{T}$ is large, we invoke Lemma 2 and the following result.

Lemma 3 *For fixed design FMR models as in (2.2) with $\tilde{\Theta}$ in (5.21): for some constants c_4, c_5 and c_6, depending on k, and K, and for $M_n = c_4\sqrt{\log n}$ and $n \geq c_6$, the following holds:*

$$\mathbb{P}_{\mathbf{x}}\left(\frac{1}{n}\sum_{i=1}^{n} F(Y_i) > c_5\frac{\log n}{n}\right) \leq \frac{1}{n},$$

where (for $i = 1, \ldots, n$)

$$F(Y_i) = G_1(Y_i)\mathrm{l}\{G_1(Y_i) > M_n\} + \mathbb{E}\Big(G_1(Y_i)\mathrm{l}\{G_1(Y_i) > M_n\}\Big|X = x_i\Big),$$

and $G_1(\cdot)$ is as in Proposition 3.

A proof is given in Appendix B.

Hence, the oracle result in Theorem 3 for our ℓ_1-norm penalized estimator in the FMR model holds on a set $\mathcal{T}$, summarized in Theorem 4, and this set $\mathcal{T}$ has large probability due to Lemma 2 and Lemma 3 as described in the following corollary.

Corollary 1 *For fixed design FMR models as in (2.2) with $\tilde{\Theta}$ in (5.21), we have for constants c_2, c_4, c_7, c_8 depending on $\Lambda_{\max}$, k, and K,*

$$\mathbb{P}[\mathcal{T}] \geq 1 - c_2\exp\left[-\frac{T^2\log^3 n}{c_7^2}\right] - n^{-1} \text{ for all } n \geq c_8,$$

where $\mathcal{T}$ is defined with $\lambda_0 = M_n\sqrt{\log^3(n)/n}$ and $M_n = c_4\sqrt{\log n}$.

Theorem 4 *(Oracle result for FMR models). Consider a fixed design FMR model as in (2.2) with $\tilde{\Theta}$ in (5.21). Assume Condition 4 (restricted eigenvalue condition) and that $\lambda \geq 2T\lambda_0$ for the estimator in (5.22). Then on $\mathcal{T}$, which has large probability as stated in Corollary 1, for the average excess risk (average Kullback-Leibler loss),*

$$\bar{\mathcal{E}}(\hat{\theta}_\lambda|\theta_0) + 2(\lambda - T\lambda_0)\|\hat{\phi}_{S^c}\|_1 \leq 8(\lambda + T\lambda_0)^2 c_0^2\kappa^2 s,$$

where c_0 and κ are defined in Lemma 1 and Condition 4, respectively.

The oracle inequality of Theorem 4 has the following well-known interpretation. First, we obtain

$$\bar{\mathcal{E}}(\hat{\theta}_\lambda|\theta_0) \leq 8(\lambda + T\lambda_0)^2 c_0^2\kappa^2 s.$$

That is, the average Kullback-Leibler risk is of the order $O(s\lambda_0^2) = O(s\log(n)^4/n)$ (take $\lambda = 2T\lambda_0$, use definition (5.20) and the assumption on M_n in Lemma 3 above) which is up to the factor $\log(n)^4$ the optimal convergence rate if one would know the s non-zero coefficients. Interestingly, instead of a factor $\log(p)$ we have here the factor $\log(n)^4$. Therefore, the dimensionality p dies not enter

explicitly in the oracle bound but it is implicitly present in the bound for the maximal eigenvalue $\Lambda_{\max}$ in Condition 6 and by considering a compact parameter space with $\sup_x \|\phi^T x\|_\infty \leq K < \infty$. As a second implication we obtain

$$\|\hat{\phi}_{S^c}\|_1 \leq 4(\lambda + T\lambda_0)c_0^2\kappa^2 s.$$

saying that the noise components in S^c have small estimated values (e.g. its ℓ_1-norm converges to zero at rate $O(s\lambda_0)$).

Note that the Conditions 1, 2, 3 and 5 hold automatically for FMR models, as described in Proposition 3. Condition 6 about the maximal eigenvalue of the design is required to ensure that the set $\mathcal{T}$ has large probability, see Lemma 2. Finally, we also require a restricted eigenvalue condition on the design, here Condition 4. In fact, for the Lasso or Dantzig selector in linear models, restricted eigenvalue conditions (Koltchinskii, 2009; Bickel et al, 2009) are considerably weaker than coherence conditions (Bunea et al, 2007; Cai et al, 2009a) or assuming the restricted isometry property (Candès and Tao, 2005; Cai et al, 2009b); for an overview among the relations see van de Geer and Bühlmann (2009).

5.3.2 High-dimensional consistency of FMR models

We finally give a consistency result for FMR models under weaker conditions than the oracle result from Section 5.3.1. Denote by θ_0 the true parameter vector in a FMR model. In contrast to Section 4, the number of covariates p can grow with the number of observations n. Therefore also the true parameter θ_0 depends on n. To guarantee consistency we have to assume some sparsity condition, i.e. the ℓ_1-norm of the true parameter can only grow with $o(\sqrt{n/\log^4(n)})$.

Theorem 5 *(Consistency). Consider a fixed design FMR model (2.2) with $\tilde{\Theta}$ in (5.21) and fixed k, and assume that Condition 6 holds. Moreover, assume that $\|\phi_0\|_1 = \sum_{r=1}^k \|\phi_{0,r}\|_1 = o(\sqrt{n/\log^4(n)})$ $(n \to \infty)$. If $\lambda = C\sqrt{\log^4(n)/n}$ for some $C > 0$ sufficiently large, then any (global) minimizer $\hat{\theta}_\lambda$ as in (5.22) satisfies*

$$\bar{\mathcal{E}}(\hat{\theta}_\lambda|\theta_0) = o_P(1) \ (n \to \infty).$$

A proof is given in Appendix B. The (restricted eigenvalue) Condition 4 on the design is not required: this is typical for a high-dimensional consistency result, see Greenshtein and Ritov (2004) for the Lasso in linear models. In our asymptotic framework, Condition 6 is a "real" condition: $\Lambda_{\max}(\Sigma_n) \leq \Lambda_{\max}$, for a constant $\Lambda_{\max} < \infty$ which does not depend on n or p. Finally, Conditions 1, 2, 3 and 5 hold automatically for FMR models as shown in Proposition 3.

6 Numerical optimization

We present a generalized EM (GEM) algorithm for optimizing the criterion in (3.10) in Section 6.1. In Section 6.2 and 6.3 we give further details on speeding-up and on initializing the algorithm. Finally, we discuss numerical convergence properties in Section 6.4. For the convex penalty ($\gamma = 0$) function we prove convergence to a stationary point.

6.1 GEM algorithm for optimization

Maximization of the log-likelihood of a mixture density is often done using the traditional EM algorithm of Dempster et al (1977). Consider the complete log-likelihood:

$$\ell_c(\theta; Y, \Delta) = \sum_{i=1}^n \sum_{r=1}^k \Delta_{i,r} \log\left(\frac{\rho_r}{\sqrt{2\pi}} e^{-\frac{1}{2}(\rho_r Y_i - X_i^T \phi_r)^2}\right) + \Delta_{i,r} \log(\pi_r).$$

Here $(\Delta_{i,1}, \ldots, \Delta_{i,k})$, $i = 1, \ldots, n$, are i.i.d unobserved multinomial variables showing the component-membership of the ith observation in the FMR model: $\Delta_{i,r} = 1$ if observation i belongs to component r and $\Delta_{i,r} = 0$ otherwise. The expected complete (scaled) negative log-likelihood is then:

$$Q(\theta|\theta') = -n^{-1}\mathbb{E}_{\theta'}[\ell_c(\theta; Y, \Delta)|Y],$$

and the expected complete (scaled) penalized negative log-likelihood is

$$Q_{pen}(\theta|\theta') = Q(\theta|\theta') + \lambda \sum_{r=1}^{k} \pi_r^{\gamma} ||\phi_r||_1.$$

The EM-algorithm works by alternating between the E- and M-step. Denote the parameter value at EM-iteration m by $\theta^{(m)}$ $(m = 0, 1, 2, \ldots)$, where $\theta^{(0)}$ is a vector of starting values.

E-Step: Compute $Q(\theta|\theta^{(m)})$ or equivalently

$$\hat{\gamma}_{i,r} = \mathbb{E}_{\theta^{(m)}}[\Delta_{i,r}|Y] = \frac{\pi_r^{(m)} \rho_r^{(m)} e^{-\frac{1}{2}(\rho_r^{(m)} Y_i - X_i^T \phi_r^{(m)})^2}}{\sum_{r=1}^{k} \pi_r^{(m)} \rho_r^{(m)} e^{-\frac{1}{2}(\rho_r^{(m)} Y_i - X_i^T \phi_r^{(m)})^2}} \quad r = 1, \ldots, k, \quad i = 1, \ldots, n.$$

Generalized M-Step: Improve $Q_{pen}(\theta|\theta^{(m)})$ w.r.t $\theta \in \Theta$.

a) *Improvement with respect to π:*
fix ϕ at the present value $\phi^{(m)}$ and improve

$$-n^{-1} \sum_{i=1}^{n} \sum_{r=1}^{k} \hat{\gamma}_{i,r} \log(\pi_r) + \lambda \sum_{r=1}^{k} \pi_r^{\gamma} ||\phi_r^{(m)}||_1 \tag{6.23}$$

with respect to the probability simplex

$$\{\pi; \pi_r > 0 \text{ for } r = 1, \ldots, k \text{ and } \sum_{r=1}^{k} \pi_r = 1\}$$

by a feasible descent step. Denote by $\bar{\pi}^{(m+1)} = \frac{\sum_{i=1}^{n} \hat{\gamma}_i}{n}$ which is a feasible point. As the simplex is convex, $\bar{\pi}^{(m+1)} - \pi^{(m)}$ is a feasible descent direction (Bertsekas (1995)). Therefore we update π as

$$\pi^{(m+1)} = \pi^{(m)} + t^{(m)}(\bar{\pi}^{(m+1)} - \pi^{(m)})$$

where $t^{(m)} \in (0, 1]$. In practice $t^{(m)}$ is chosen to be the largest value in the grid $\{\delta^k;\ k = 0, 1, 2, \ldots\}$ $(0 < \delta < 1)$ such that (6.23) is decreased. In our examples $\delta = 0.1$ worked well. The Limited Minimization Rule or the Armijo Rule (Bertsekas (1995)) for choosing $t^{(m)}$ are also possible.

b) *Coordinate descent improvement with respect to ϕ and ρ:*
A simple calculation shows, that the M-Step decouples for each component into k distinct optimization problems of the form

$$-\log(\rho_r) + \frac{1}{2n_r}||\rho_r \tilde{Y} - \tilde{X}\phi_r||^2 + \frac{n\lambda}{n_r}\left(\pi_r^{(m+1)}\right)^{\gamma} ||\phi_r||_1, \qquad r = 1, \ldots, k \tag{6.24}$$

with

$$n_r = \sum_{i=1}^{n} \hat{\gamma}_{i,r}, \quad (\tilde{Y}_i, \tilde{X}_i) = \sqrt{\hat{\gamma}_{i,r}}(Y_i, X_i), \qquad r = 1, \ldots, k.$$

Problem (6.24) has the same form as (3.8): in particular, it is convex in $(\rho_r, \phi_{r,1}, \ldots, \phi_{r,p})$. Instead of fully optimizing (6.24) we only minimize with respect to each of the coordinates,

holding the other coordinates at their current value. Closed-form coordinate updates can easily be computed for each component r ($r = 1, \ldots, k$) using Proposition 1:

$$\rho_r^{(m+1)} = \frac{\langle \tilde{Y}, \tilde{X}\phi_r^{(m)} \rangle + \sqrt{\langle \tilde{Y}, \tilde{X}\phi_r^{(m)} \rangle^2 + 4||\tilde{Y}||^2 n_r}}{2||\tilde{Y}||^2},$$

$$\phi_{r,j}^{(m+1)} = \begin{cases} 0 & \text{if } |S_j| \leq n\lambda \left(\pi_r^{(m+1)}\right)^\gamma, \\ \left(n\lambda \left(\pi_r^{(m+1)}\right)^\gamma - S_j\right) / ||\tilde{X}_j||^2 & \text{if } S_j > n\lambda \left(\pi_r^{(m+1)}\right)^\gamma, \\ -\left(n\lambda \left(\pi_r^{(m+1)}\right)^\gamma + S_j\right) / ||\tilde{X}_j||^2 & \text{if } S_j < -n\lambda \left(\pi_r^{(m+1)}\right)^\gamma, \end{cases}$$

where S_j is defined as

$$S_j = -\rho_r^{(m+1)} \langle \tilde{X}_j, \tilde{Y} \rangle + \sum_{s<j} \phi_{r,s}^{(m+1)} \langle \tilde{X}_j, \tilde{X}_s \rangle + \sum_{s>j} \phi_{r,s}^{(m)} \langle \tilde{X}_j, \tilde{X}_s \rangle$$

and $j = 1, \ldots, p$.

Because we only improve $Q_{pen}(\theta|\theta^{(m)})$ instead of a full minimization, see M-step a) and b), this is a generalized EM (GEM) algorithm. We call it the block coordinate descent generalized EM algorithm (BCD-GEM); the word block refers to the fact that we are up-dating all components of π at once. Its numerical properties are discussed in Section 6.4.

Remark 1 For the convex penalty function with $\gamma = 0$, a minimization with respect to π in M-step a) is achieved with $\pi^{(m+1)} = \frac{\sum_{i=1}^n \hat{\gamma}_i}{n}$, i.e. using $t^{(m)} = 1$. Then, our M-Step corresponds to exact coordinate-wise minimization of $Q_{pen}(\theta|\theta^{(m)})$.

6.2 Active set algorithm

There is a simple way to speed-up the algorithm described above. When updating the coordinates $\phi_{r,j}$ in the M-step b), we restrict ourselves during every 10 EM-iterations to the current active set (the non-zero coordinates) and visit the remaining coordinates every 11th EM-iteration to update the active set. In very high-dimensional and sparse settings this leads to a remarkable decrease in computational times. A similar active set strategy is also used in Friedman et al (2007) and Meier et al (2008). We illustrate in Section 7.3 the gain of speed when staying during every 10 EM-iterations within the active set.

6.3 Initialization

The algorithm of Section 6.1 requires the specification of starting values $\theta^{(0)}$. We found empirically that the following initialization works well. For each observation i, $i = 1, \ldots, n$, draw randomly a class $\kappa \in \{1, \ldots, k\}$. Assign for observation i and component κ the weight $\tilde{\gamma}_{i,\kappa} = 0.9$ and weights $\tilde{\gamma}_{i,r} = 0.1$ for all other components. Finally, normalize $\tilde{\gamma}_{i,r}$, $r = 1, \ldots, k$, to achieve that summing over the indices k yields the value one, to get the normalized values $\hat{\gamma}_{i,r}$. Note that this can be viewed as an initialization of the E-step. In the M-step which follows afterward, we update all coordinates from the initial values $\phi_{r,j}^{(0)} = 0$, $\rho_r^{(0)} = 2$, $\pi_r^{(0)} = 1/k$, $r = 1, \ldots, k$, $j = 1, \ldots, p$.

6.4 Numerical Convergence of the BCD-GEM algorithm

We are addressing here convergence properties of BCD-GEM algorithm described in Section 6.1. A detailed account of the convergence properties of the EM algorithm in a general setting has been given by Wu (1983). Under regularity conditions including differentiability and continuity, convergence to stationary points is proved for the EM algorithm. For the GEM algorithm similar statements are true under conditions which are often hard to verify.

As a GEM algorithm, our BCD-GEM algorithm has the descent property which means, that the criterion function is reduced in each iteration,

$$-n^{-1}\ell^{(\gamma)}_{pen,\lambda}(\theta^{(m+1)}) \leq -n^{-1}\ell^{(\gamma)}_{pen,\lambda}(\theta^{(m)}). \tag{6.25}$$

Since $-n^{-1}\ell^{(0)}_{pen,\lambda}(\theta)$ is bounded from below (Proposition 2), the following result holds.

Proposition 4 *For the BCD-GEM algorithm, $-n^{-1}\ell^{(0)}_{pen,\lambda}(\theta^{(m)})$ decreases monotonically to some value $\bar{\ell} > -\infty$.*

In Remark 1 we noted, that for the convex penalty function with $\gamma = 0$, the M-Step of the algorithm corresponds to exact coordinate-wise minimization of $Q_{pen}(\theta|\theta^{(m)})$. In this case convergence to a stationary point can be shown.

Theorem 6 *Consider the BCD-GEM algorithm for the criterion function in (3.10) with $\gamma = 0$. Then, every cluster point $\bar{\theta} \in \Theta$ of the sequence $\{\theta^{(m)}; m = 0, 1, 2, \ldots\}$, generated by the BCD-GEM algorithm, is a stationary point of the criterion function in (3.10).*

A proof is given in Appendix C. It uses the crucial facts that $Q_{pen}(\theta|\theta')$ is a convex function in θ and that it is strictly convex in each coordinate of θ.

7 Simulations, real data example and computational timings

7.1 Simulations

We consider four different simulation setups. Simulation scenario 1 compares the performance of the unpenalized MLE with our estimator from Section 3.2 (FMRLasso) and Section 3.3 (FMRAdapt) in a situation where the total number of noise covariates grows successively. For computing the unpenalized MLE we used the `R`-package Flexmix (Leisch, 2004; Grün and Leisch, 2007, 2008); Simulation 2 explores sparsity; Simulation 3 compares cross-validation and BIC; and Simulation 4 compares the different penalty functions with the parameters $\gamma = 0, 1/2, 1$. For every setting, the results are based on 100 independent simulation runs.

All simulations are based on Gaussian FMR models as in (2.2): the coefficients π_r, β_r, σ_r and the sample size n are specified below. The covariate X is generated from a multivariate normal distribution with mean 0 and covariance structure as specified below.

Unless otherwise specified, the penalty with $\gamma = 1$ is used in all simulations. As explored empirically in Simulation 4, in case of balanced problems (approximately equal π_r), the FMRLasso performs similarly for all three penalties. In unbalanced situations the best results are typically achieved with $\gamma = 1$. In addition, unless otherwise specified the true number of components k is assumed to be known.

For all models, training-, validation- and test data are generated of equal size n. The estimators are computed on the training data, with the tuning parameter (e.g. λ) selected by minimizing twice the negative log-likelihood (log-likelihood loss) on the validation data. As performance measure, the predictive log-likelihood loss (twice the negative log-likelihood) of the selected model is computed on the test data.

Regarding variable selection, we count a covariable $X^{(j)}$ as selected if $\hat{\beta}_{r,j} \neq 0$ for at least one $r \in \{1, \ldots, k\}$. To assess the performance of FMRLasso on recovering the sparsity structure, we report the number of truly selected covariates (True Positives) and falsely selected covariates (False Positives).

Obviously, the performances depend on the signal to noise ratio (SNR) which we define for an FMR model as:

$$\text{SNR} = \frac{\text{Var}(Y)}{\text{Var}(Y|\beta_r = 0; r = 1, \ldots, k)} = \frac{\sum_{r=1}^{k} \pi_r(\beta_r^T \text{Cov}(X)\beta_r + \sigma_r^2)}{\sum_{r=1}^{k} \pi_r \sigma_r^2},$$

where the last inequality follows since $\mathbb{E}[X] = 0$.

7.1.1 Simulation 1

We consider five different FMR models: M1, M2, M3, M4 and M5. The parameters $(\pi_k, \beta_k, \sigma_k)$, the sample size n of the training-, validation- and test-data, the correlation structure of covariates $corr_{l,m} = corr(X_l, X_m)$ and the signal to noise ratio (SNR) are specified in Table 1. Models M1, M2, M3 and M5 have two components and five active covariates, whereas model M4 has three components and six active covariates. M1, M2 and M3 differ only in their variances σ_1^2, σ_2^2 and hence have different signal to noise ratios. Model M5 has a non-diagonal covariance structure. Furthermore in model M5, the variances σ_1^2, σ_2^2 are tuned to achieve the same signal to noise ratio as in model M1.

We compare the performances of the maximum likelihood estimator (MLE), the FMRLasso and the FMRAdapt in a situation where the number of noise covariates grows successively. For the models M1, M2, M3, M5 with two components, we start with $p_{tot} = 5$ (no noise covariates) and go up to $p_{tot} = 125$ (120 noise covariates). For the three component model M4 we start with $p_{tot} = 6$ (no noise covariates) and go up to $p_{tot} = 155$ (149 noise covariates).

The boxplots in Figures 1 - 5 of the predictive log-likelihood loss, denoted by *Error*, the True Positives (*TP*) and the False Positives (*FP*) over 100 simulation runs summarize the results for the different models. We read off from these boxplots that the MLE performs very badly when we adding noise covariates. On the other hand, our penalized estimators remain stable. For example, for M1 the MLE with $p_{tot} = 20$ performs worse than the FMRLasso with $p_{tot} = 125$, or for M4 the MLE with $p_{tot} = 10$ performs worse than the FMRLasso with $p_{tot} = 75$. Impressive is also the huge gain of the FMRAdapt method over FMRLasso in terms of log-likelihood loss and false positives.

	M1	M2	M3	M4	M5
n	100	100	100	150	100
β_1	(3,3,3,3,3)	(3,3,3,3,3)	(3,3,3,3,3)	(3,3,0,0,0,0)	(3,3,3,3,3)
β_2	(-1,-1,-1,-1,-1)	(-1,-1,-1,-1,-1)	(-1,-1,-1,-1,-1)	(0,0,-2,-2,0,0)	(-1,-1,-1,-1,-1)
β_3	-	-	-	(0,0,0,0,-3,2)	-
σ	0.5, 0.5	1, 1	1.5, 1.5	0.5, 0.5, 0.5	0.95, 0.95
π	0.5, 0.5	0.5, 0.5	0.5, 0.5	1/3, 1/3, 1/3	0.5, 0.5
$corr_{l,m}$	$\delta_{l,m}$	$\delta_{l,m}$	$\delta_{l,m}$	$\delta_{l,m}$	$0.8^{\lvert l-m \rvert}$
SNR	101	26	12.1	53	101

Table 1 Models for simulation 1.

7.1.2 Simulation 2

In this Section we explore the sparsity properties of the FMRLasso. The model specifications are given in Table 2. Consider the ratio of $p_{act} : n : p_{tot}$. The total number of covariates p_{tot} grows faster than the number of observations n and the number of active covariates p_{act}: when p_{tot} is doubled, p_{act} is raised by one and n is raised by fifty from model to model. In particular, we obtain

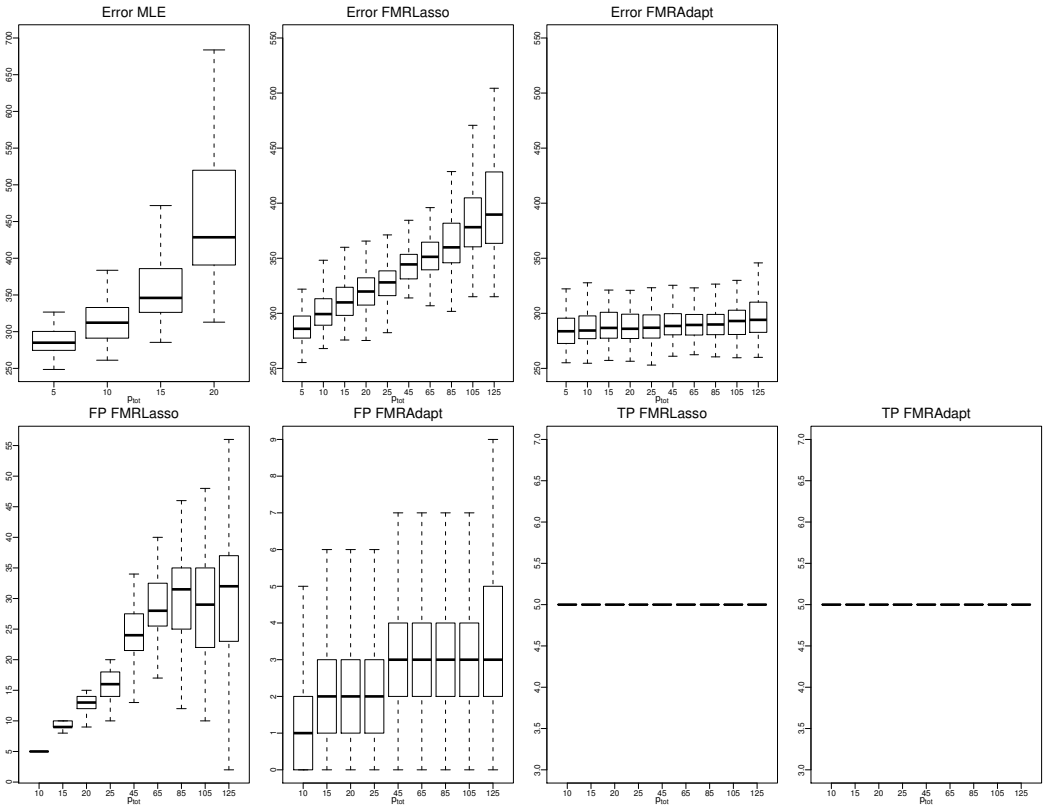

Fig. 1 Simulation 1, Model M1. Top: predictive log-likelihood loss (*Error*) for MLE, FMRLasso, FMRAdapt. Bottom: False Positives (*FP*) and True Positives (*TP*) for FMRLasso and FMRAdapt.

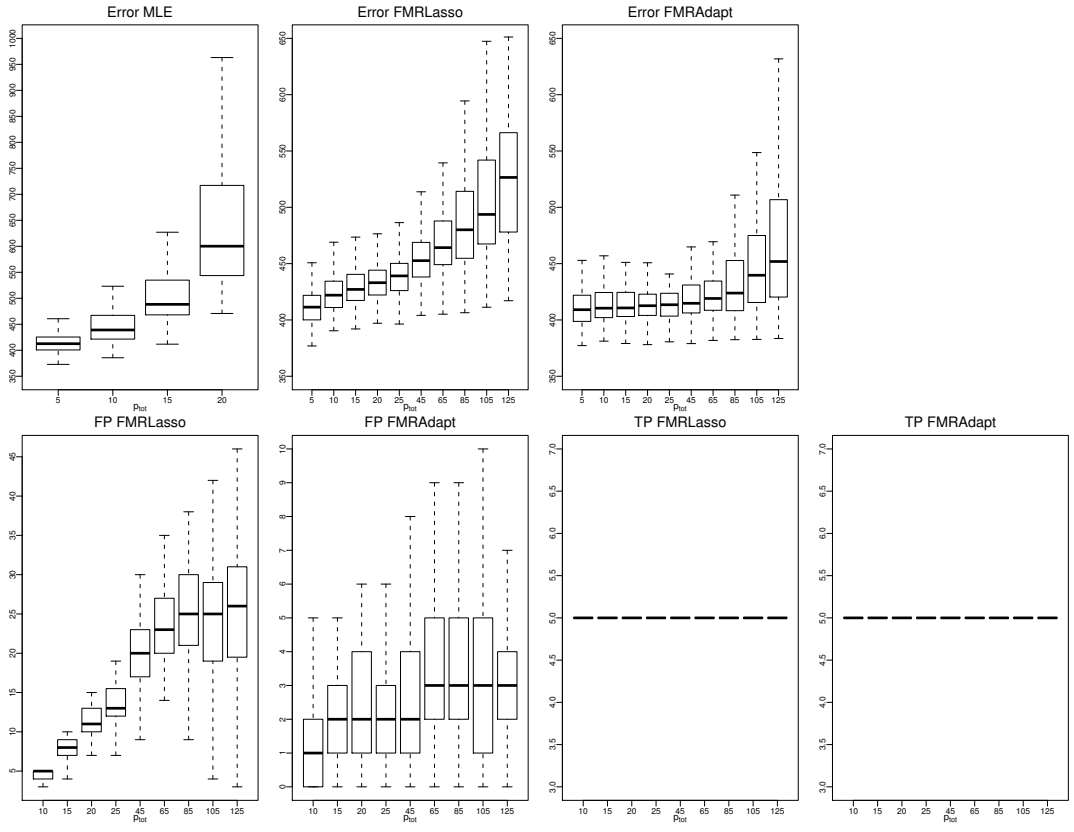

Fig. 2 Simulation 1, Model M2. Same notation as in Figure 1.

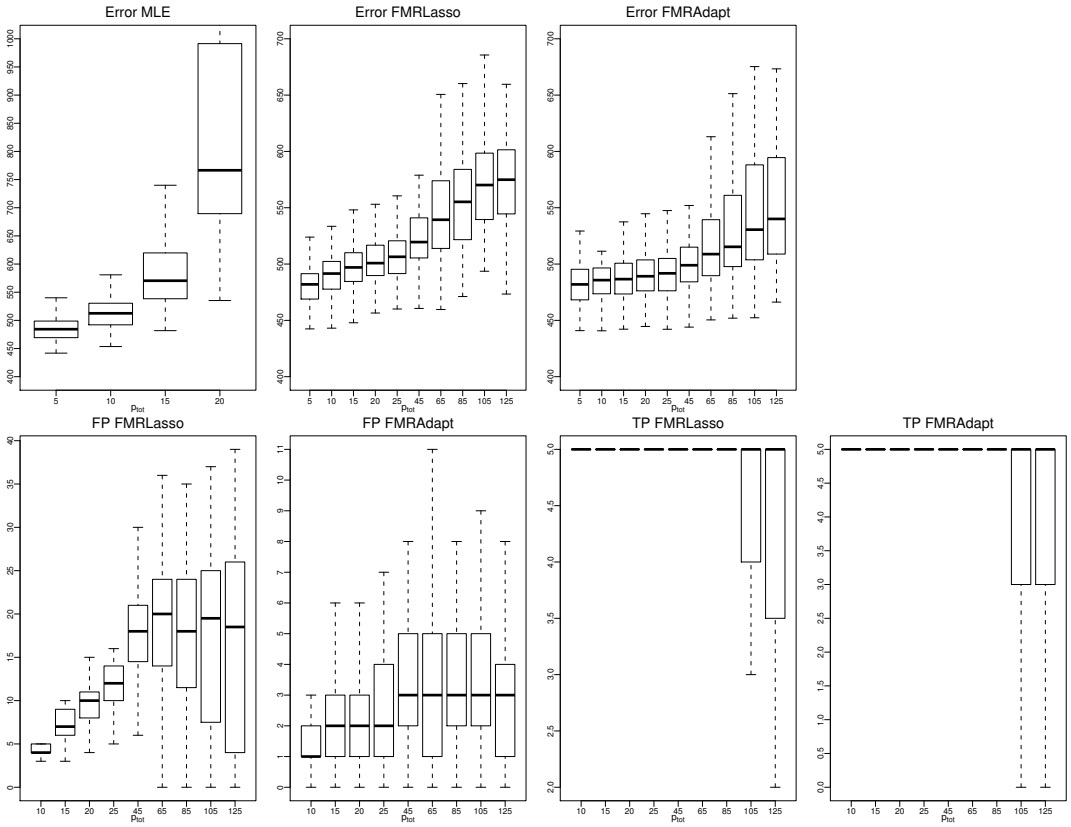

Fig. 3 Simulation 1, Model M3. Same notation as in Figure 1.

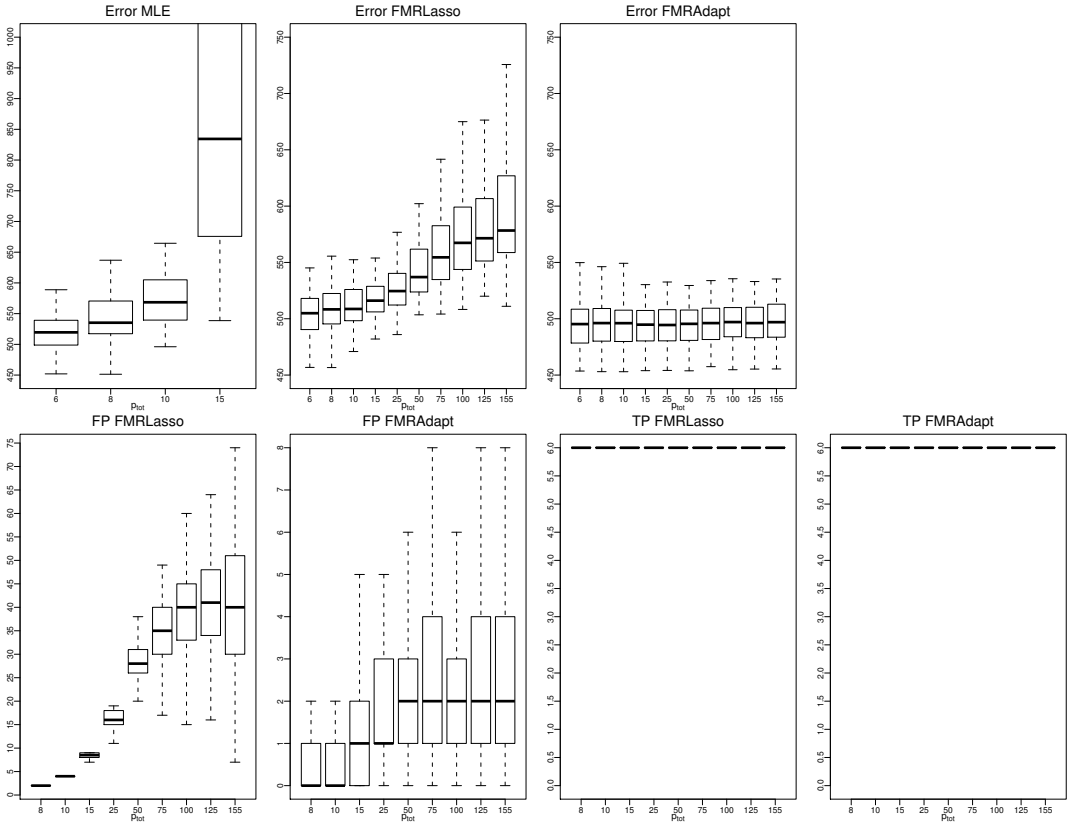

Fig. 4 Simulation 1, Model M4. Same notation as in Figure 1.

a series of models which gets "sparser" as n grows (larger ratio n/p_{act}). In order to compare the performance of the FMRLasso we report the True Positive Rate (TPR) and the False Positive Rate (FPR) defined as:

$$\text{TPR} = \frac{\#\text{truly selected covariates}}{\#\text{active covariates}},$$
$$\text{FPR} = \frac{\#\text{falsely selected covariates}}{\#\text{inactive covariates}}.$$

These numbers are reported in Figure 6. We see that the False Positive Rate approaches zero for sparser models indicating that the FMRLasso recovers the true model better in sparser settings regardless of the large number of noise covariates.

p_{act}	3	4	5	6	7	8	9
n	50	100	150	200	250	300	350
p_{tot}	10	20	40	80	160	320	640
β_1				(3,3,3,0,0,...)			
β_2				(-1,-1,-1,0,0,...)			
σ				0.5, 0.5			
π				0.5, 0.5			

Table 2 Series of models for simulation 2 which gets "sparser" as n grows: when p_{tot} is doubled, p_{act} is raised by one and n is raised by fifty from model to model.

7.1.3 Simulation 3

So far, we regarded the number k of components as given, while we have chosen an optimal λ_{opt} by minimizing the negative log-likelihood loss on validation data. In this section we compare the performance of 10-fold cross-validation and the BIC criterion presented in Section 3.4 for selecting the tuning parameters k and λ. We use model M1 of Section 7.1.1 with $p_{tot} = 25, 50, 75$. For each of these models we tune the FMRLasso estimator according to the following strategies:

(1) Assume the number of components is given ($k = 2$). Choose the optimal tuning parameter λ_{opt} using 10-fold cross-validation.
(2) Assume the number of components is given ($k = 2$). Choose λ_{opt} by minimizing the BIC criterion (3.15).
(3) Choose the number of components $k \in \{1, 2, 3\}$ and λ_{opt} by minimizing the BIC criterion (3.15).

The results of this simulation are presented in Figure 7, where boxplots of the log-likelihood loss (*Error*) are shown. All three strategies perform equally well. With $p_{tot} = 25$ the BIC criterion in strategy (3) chooses always $k = 2$. For the model with $p_{tot} = 50$ strategy (3) chooses in ninety-eight simulation runs $k = 2$ and in two runs $k = 3$. Finally with $p_{tot} = 75$ the third strategy chooses ninety-two times $k = 2$ and eight times $k = 3$.

7.1.4 Simulation 4

In the preceding simulations we always used the value $\gamma = 1$ in the penalty term of the FMRLasso estimator (3.10). In this Section we compare the FMRLasso for different values $\gamma = 0, 1/2, 1$. First we compute the FMRLasso for $\gamma = 0, 1/2, 1$ on model M1 of Section 7.1.1 with $p_{tot} = 50$. Then we do the same calculations for an "unbalanced" version of this model with $\pi_1 = 0.3$ and $\pi_2 = 0.7$.

In Figure 8, the boxplots of the log-likelihood loss (*Error*), the False Positives (FP) and the True Positives (TP) over 100 simulation runs are shown. We see that the FMRLasso performs similarly for $\gamma = 0, 1/2, 1$. Nevertheless the value $\gamma = 1$ is slightly preferable in the "unbalanced" setup.

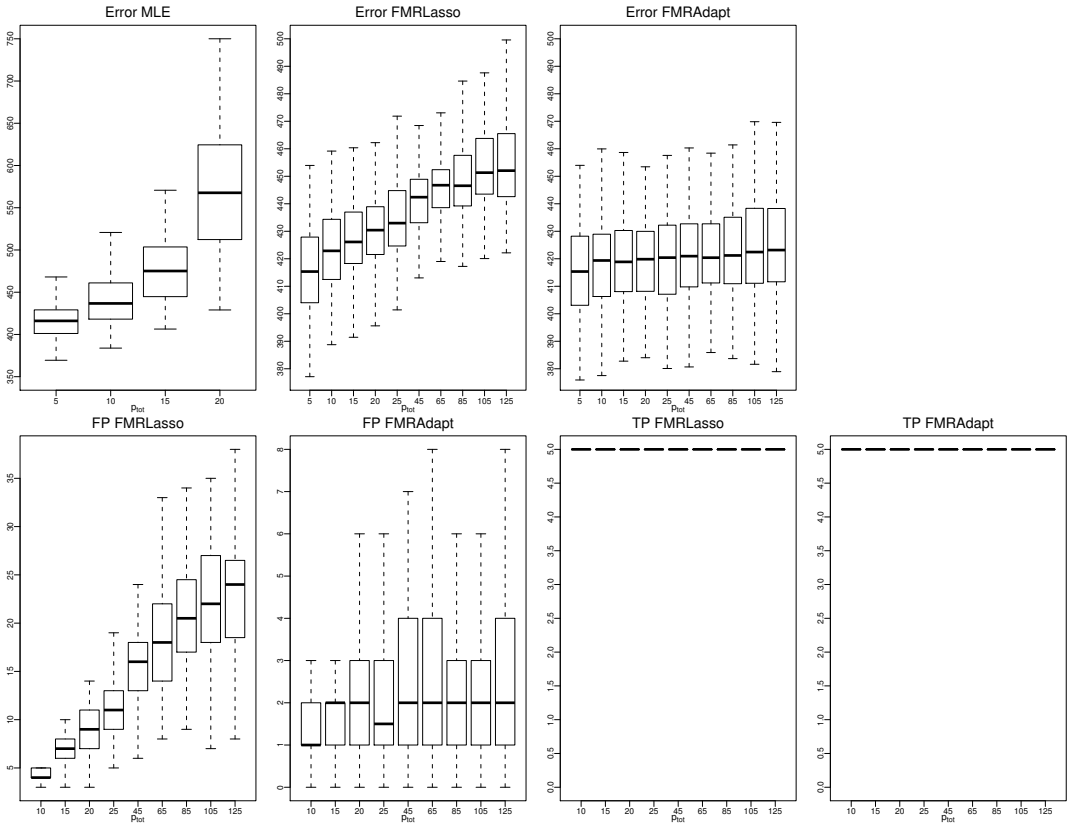

Fig. 5 Simulation 1, Model M5. Same notation as in Figure 1.

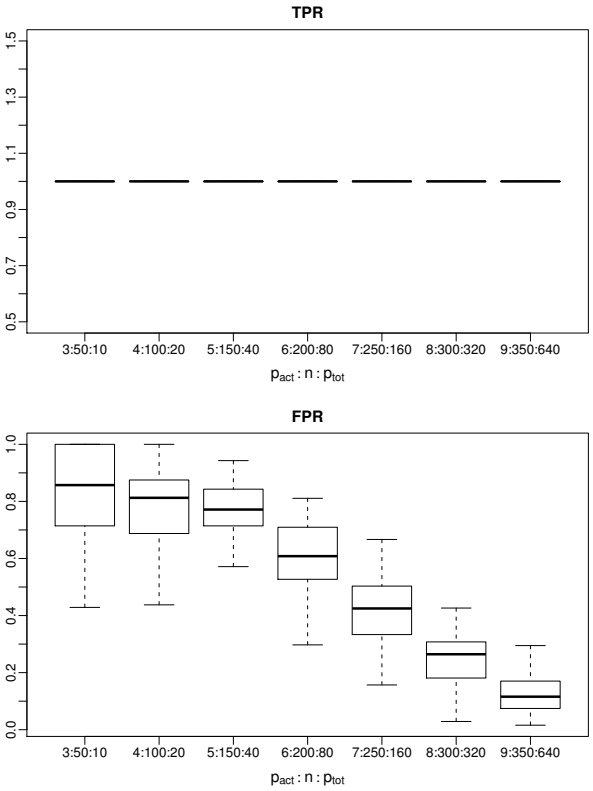

Fig. 6 Simulation 2 compares the performance of the FMRLasso for a series of models which gets "sparser" as the sample size grows. Top: True Positive Rate (TPR). Bottom: False Positive Rate (FPR) over 100 simulation runs.

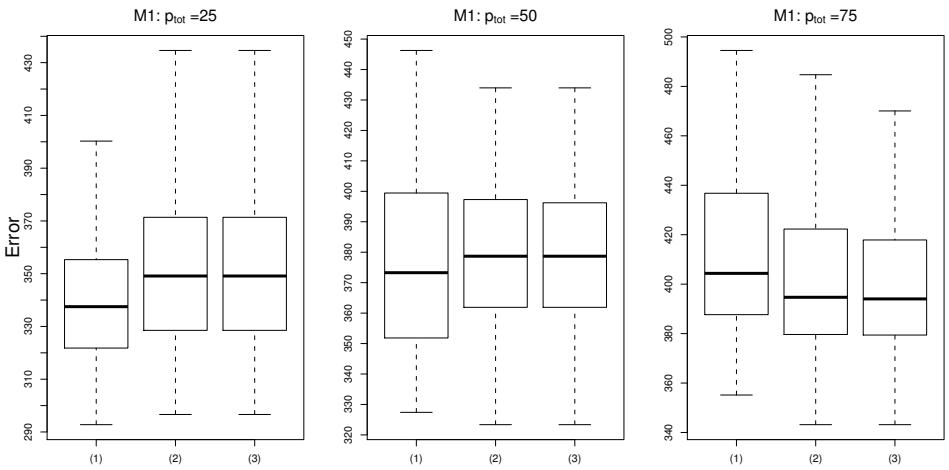

Fig. 7 Simulation 3 compares different strategies for choosing the tuning parameters k and λ. The boxplots show the predictive log-likelihood loss (*Error*) of the FMRLasso, tuned by strategies (1), (2) and (3), for model M1 with $p_{tot} = 25, 50, 75$.

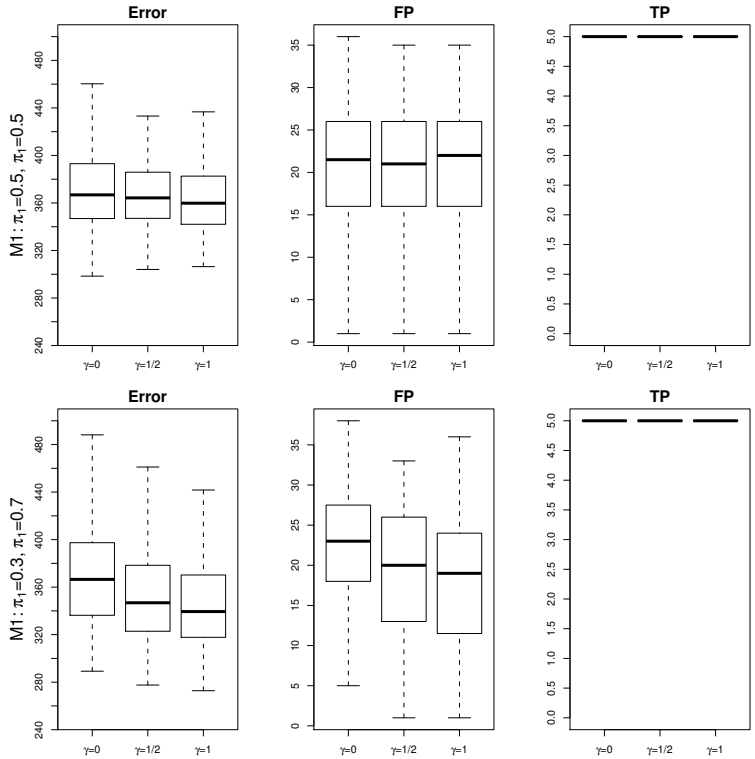

Fig. 8 Simulation 4 compares the FMRLasso for different values $\gamma = 0, 1/2, 1$. The upper row of the panels shows the boxplots of the log-likelihood loss (*Error*), the False Positives (*FP*) and the True Positives (*TP*) for model M1 with $p_{tot} = 50$ and $\pi_1 = \pi_2 = 0.5$. The lower row of the panels shows the same boxplots for an "unbalanced" version of model M1 with $\pi_1 = 0.3$ and $\pi_2 = 0.7$.

7.2 Real data example

We now apply the FMRLasso to a data set about riboflavin (vitamin B_2) production by *Bacillus Subtilis*. The real-valued response variable is the logarithm of the riboflavin production rate. The data has been kindly provided by DSM (Switzerland). There are $p = 4088$ covariates (genes) measuring the logarithm of the expression level of 4088 genes and measurements of $n = 146$ genetically engineered mutants of *Bacillus Subtilis*. The population seems to be rather heterogeneous as there are different strains of *Bacillus Subtilis* which are cultured under different fermentation conditions. We do not know the different homogeneity subgroups. For this reason, a FMR model with more than one component might be more appropriate than a simple linear regression model.

We compute the FMRLasso estimator for $k = 1, \ldots, 5$ components. To keep the computational effort reasonable we use only the 100 covariates (genes) exhibiting the highest empirical variances. We choose the optimal tuning parameter λ_{opt} by 10-fold cross-validation (using the log-likelihood loss). As a result we get six different estimators which we compare according to their cross-validated log-likelihood loss (*CV Error*). These numbers are plotted in Figure 9. The estimator with three components performs clearly best, resulting in a 17% improvement in prediction over a (non-mixture) linear model, and it selects 51 genes. In Figure 10 the coefficients of the twenty most important genes, ordered according to $\sum_{r=1}^{3} |\hat{\beta}_{r,j}|$, are shown. From the important variables only gen 83 shows opposite sign of the estimated regression coefficients among the three different mixture components. However, it happens that some covariates (genes) exhibit a strong effect in one or two mixture components but none in the remaining other components. Finally, for comparison, the one-component (non-mixture) model selects 26 genes where 24 selected genes from the one-component model are also selected in the the three-component model.

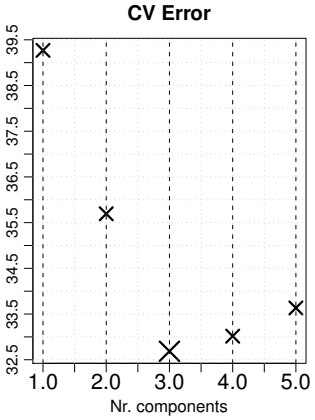

Fig. 9 Riboflavin production data. Cross-validated negative log-likelihood loss (*CV Error*) for the FMRLasso estimator when varying over different numbers of components.

7.3 Computational timings

In this section we report on the run times of the BCD-GEM algorithm on two high-dimensional examples. In particular, we focus on the substantial gain of speed achieved by using the active set version of the algorithm described in Section 6.2. All computations were carried out with the statistical computing language and environment `R`. Timings depend on the stopping criterion used in the algorithm. We stop the algorithm if the relative function improvement and the relative change of the parameter vector are small enough, i.e.

$$\frac{|\ell^{(\gamma)}_{pen,\lambda}(\theta^{(m+1)}) - \ell^{(\gamma)}_{pen,\lambda}(\theta^{(m)})|}{1 + |\ell^{(\gamma)}_{pen,\lambda}(\theta^{(m+1)})|} \leq \tau, \qquad \max_j \left\{ \frac{|\theta_j^{(m+1)} - \theta_j^{(m)}|}{1 + |\theta_j^{(m+1)}|} \right\} \leq \sqrt{\tau}, \qquad \tau = 10^{-6}.$$

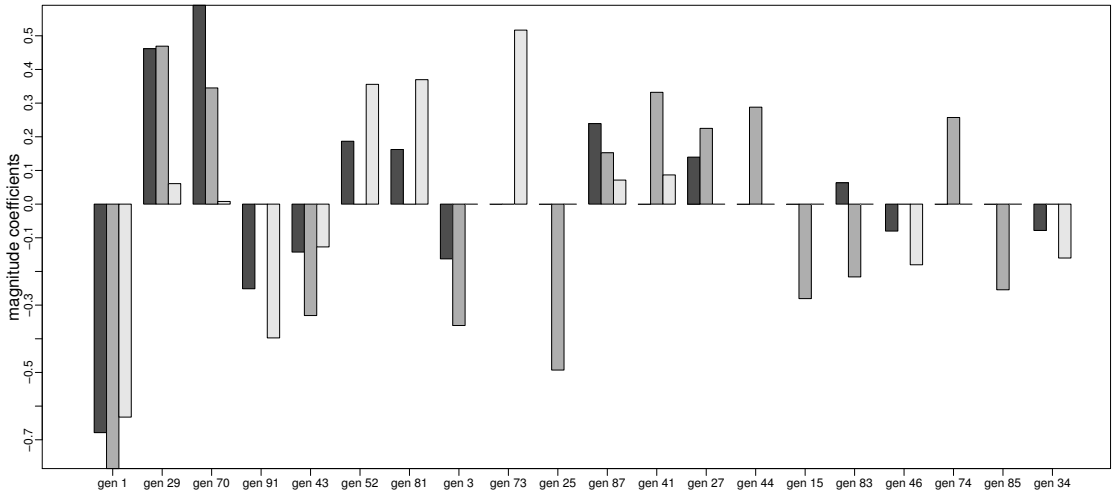

Fig. 10 Riboflavin production data. Coefficients of the twenty most important genes, ordered according to $\sum_{r=1}^{3} |\beta_{r,j}|$, for the prediction optimal model with three components.

We consider a high-dimensional version of the two component model M1 from Section 7.1.1 with $n = 200$, $p_{tot} = 1000$ and the riboflavin data set from Section 7.2 with three components, $n = 146$ and $p_{tot} = 100$. We use the BCD-GEM algorithm with and without active set strategy to fit the FMRLasso on a small grid of eight values for λ. The corresponding BIC, CPU times (in seconds) and number of EM-iterations are reported in Tables 3 and 4. The values for the BCD-GEM without active set strategy are written in brackets. For model M1 and an appropriate λ with minimal BIC score, the active set algorithm converges in 5.96 seconds whereas the standard BCD-GEM needs 53.15 seconds. There is also a considerable gain of speed for the real data: 0.89 seconds versus 3.57 seconds for λ with optimal BIC. Note that in Table 3, the BIC scores sometimes differ substantially for inappropriate values of λ. For such regularization parameters, the solutions are unstable and different local optima are attained depending on the algorithm used. However, if the regularization parameter is in a reasonable range with low BIC score, the results stabilize.

λ	10.0	15.6	21.1	26.7	32.2	37.8	43.3	48.9
BIC	2033 (2022)	1606 (1748)	951 (959)	941 (940)	989 (983)	1236 (1073)	1214 (1216)	1206 (1203)
CPU [s]	26.78 (269.91)	17.05 (165.78)	8.63 (82.78)	5.96 (53.15)	5.08 (44.23)	4.23 (37.27)	3.35 (18.99)	3.30 (15.62)
# EM-iter.	277.0 (341.5)	196.0 (205.0)	96.0 (100.5)	63.5 (64.5)	56.0 (53.5)	41.5 (46.0)	31.5 (23.0)	25.0 (19.0)

Table 3 Model M1 with $n = 200$ and $p_{tot} = 1000$. Median over 10 simulation runs of BIC, CPU times and number of EM-iterations for the BCD-GEM with and without active set strategy (the latter in brackets).

λ	3.0	13.8	24.6	35.4	46.2	57.0	67.8	78.6
BIC	560 (628)	536 (530)	516 (522)	532 (525)	541 (540)	561 (580)	592 (591)	611 (613)
CPU [s]	22.40 (29.98)	1.35 (3.28)	0.89 (3.57)	0.86 (3.34)	0.78 (3.87)	0.69 (2.42)	0.37 (2.56)	0.85 (4.05)
# EM-iter.	3389 (2078)	345 (239)	287 (266)	298 (247)	296 (290)	248 (184)	129 (192)	313 (302)

Table 4 Riboflavin data with $k = 3$, $n = 146$ and $p_{tot} = 100$. BIC, CPU times and number of EM-iterations for the BCD-GEM with and without active set strategy (the latter in brackets).

8 Discussion

We have presented an ℓ_1-penalized estimator for a finite mixture of high-dimensional Gaussian regressions where the number of covariates may greatly exceed sample size. Such a model and the corresponding Lasso-type estimator are useful to blindly account for often encountered inhomogeneity of high-dimensional data. On a high-dimensional real data example, we demonstrate a 17% gain in prediction accuracy over a (non-mixture) linear model.

The computation and mathematical analysis in such a high-dimensional mixture model is challenging due to the non-convex behavior of the negative log-likelihood. Moreover, with high-dimensional estimation defined via optimization of a non-convex objective function, there is a major gap between the actual computation and the procedure analyzed in theory. We do not provide an answer to this issue in this paper. Regarding the computation in FMR models, a simple reparameterization is very beneficial and the ℓ_1-penalty term makes the optimization problem numerically much better behaved. We develop an efficient generalized EM-algorithm and we prove its numerical convergence to a stationary point. Regarding the statistical properties, besides standard low-dimensional asymptotics, we present a non-asymptotic oracle inequality for the Lasso-type estimator in a high-dimensional setting with general, non-convex but smooth loss functions. The mathematical arguments are different than what is typically used for convex losses.

A Proofs for Section 4

A.1 Proof of Theorem 1

The regularity assumptions (A)-(C) of Fan and Li (2001) are fulfilled for finite mixtures of Gaussians (Lehmann (1983), page 442). Therefore, the Theorem follows from Theorem 1 of Fan and Li (2001). □

A.2 Proof of Theorem 2

In order to keep the notation simple we give the proof for a two class mixture with $k = 2$. All arguments in the proof do also hold for a general mixture with more than two components. Remember that $-n^{-1}\ell_{adapt}(\theta)$ is given by

$$-n^{-1}\ell_{adapt}(\theta) = -n^{-1}\ell(\theta) + \lambda\left(\pi_1^{\gamma}\sum_{j=1}^{p} w_{1,j}|\phi_{1,j}| + (1-\pi_1)^{\gamma}\sum_{j=1}^{p} w_{2,j}|\phi_{2,j}|\right),$$

where $\ell(\theta) \equiv \ell(\theta; Y) = \sum_{i=1}^{n}\log(h_\theta(y_i|x_i))$ is the log-likelihood function of a FMR model (see also equation (2.3) in Section 2.1). The weights $w_{r,j}$ are given by $w_{r,j} = \frac{1}{|\phi_{r,j}^{ini}|}$, $r = 1, 2$ and $j = 1, \ldots, p$.

Assertion 1.
Let $\hat{\theta}$ be a root-n consistent local minimizer of $-n^{-1}\ell_{adapt}(\theta)$ (construction as in Fan and Li (2001)).

For all $(r, j) \in S$ from consistency of $\hat{\theta}$ we easily see that $\mathbb{P}[(r, j) \in \hat{S}] \to 1$. It then remains to show that for all $(r, j) \in S^c$, $\mathbb{P}[(r, j) \in \hat{S}^c] \to 1$. Assume the contrary, i.e. w.l.o.g there is a $s \in \{1, \ldots, p\}$ with $\phi_{1,s} = 0$ such that $\hat{\phi}_{1,s} \neq 0$ with non-vanishing probability.

By using Taylor's theorem there exists a (random) vector $\tilde{\theta}$ on the line segment between θ_0 and $\hat{\theta}$ such that

$$\begin{aligned}
&\frac{1}{n}\frac{\partial \ell_{adapt}}{\partial \phi_{1,s}}|_{\hat{\theta}} \\
&= \underbrace{\frac{1}{n}\frac{\partial \ell}{\partial \phi_{1,s}}|_{\theta_0}}_{(1)} + \underbrace{\frac{1}{n}\frac{\partial \ell'}{\partial \phi_{1,s}}|_{\theta_0}}_{(2)}\left(\hat{\theta} - \theta_0\right) + \frac{1}{2}\left(\hat{\theta} - \theta_0\right)^T \underbrace{\frac{1}{n}\frac{\partial \ell''}{\partial \phi_{1,s}}|_{\tilde{\theta}}}_{(3)}\left(\hat{\theta} - \theta_0\right) - \lambda\hat{\pi}^{\gamma} w_{1,s} sgn(\hat{\phi}_{1,s}).
\end{aligned}$$

Now, term (1) is an average of the i.i.d. random vectors $\frac{\partial \log(h_\theta(y_i|x_i))}{\partial \phi_{1,s}}|_{\theta_0}$, which have mean 0. By the central limit theorem, term (1) is therefore of order $O_P(\frac{1}{\sqrt{n}})$. Similar, term (2) is of order $O_P(1)$ by the law of large numbers.

Term (3) is of order $O_P(1)$ by the law of large numbers and the regularity condition on 3rd derivatives (condition (C) of Fan and Li (2001)).

Therefore we have

$$\frac{1}{n}\frac{\partial \ell_{adapt}}{\partial \phi_{1,s}}|_{\hat{\theta}} = O_P(\frac{1}{\sqrt{n}}) + \left(O_P(1) + \left(\hat{\theta}-\theta_0\right)^T O_P(1)\right)\left(\hat{\theta}-\theta_0\right) - \lambda\hat{\pi}^{\gamma} w_{1,s} sgn(\hat{\phi}_{1,s}).$$

As $\hat{\theta}$ is root-n consistent we get

$$\begin{aligned}\frac{1}{n}\frac{\partial \ell_{adapt}}{\partial \phi_{1,s}}|_{\hat{\theta}} &= O_P(\frac{1}{\sqrt{n}}) + \left(O_P(1) + o_P(1)O_P(1)\right) O_P(\frac{1}{\sqrt{n}}) - \lambda\hat{\pi}^{\gamma} w_{1,s} sgn(\hat{\phi}_{1,s}) \\ &= \frac{1}{\sqrt{n}}\left(O_P(1) - \frac{n\lambda}{\sqrt{n}}\hat{\pi}^{\gamma} w_{1,s} sgn(\hat{\phi}_{1,s})\right). \end{aligned} \tag{A.26}$$

From the assumption on the initial estimator we have:

$$\frac{n\lambda}{\sqrt{n}} w_{1,s} = \frac{n\lambda}{\sqrt{n}|\phi_{1,s}^{ini}|} = \frac{n\lambda}{O_P(1)} \to \infty \qquad \text{as} \qquad n\lambda \to \infty.$$

Therefore the second term in the bracket of (A.26) dominates the first and the probability of the event

$$\left\{ sgn\left(\frac{1}{n}\frac{\partial \ell_{adapt}}{\partial \phi_{1,s}}|_{\hat{\theta}}\right) = -sgn(\hat{\phi}_{1,s}) \neq 0 \right\}$$

tends to 1. But this contradicts the assumption that $\hat{\theta}$ is a local minimizer (i.e. $\frac{1}{n}\frac{\partial \ell_{adapt}}{\partial \phi_{1,s}}|_{\hat{\theta}} = 0$).

Assertion 2.
Write $\theta = (\theta_S, \theta_{S^c})$. From part 1) it follows that with probability tending to one $\hat{\theta}_S$ is a root-n local minimizer of $-n^{-1}\ell_{adapt}(\theta_S, 0)$.

By using a Taylor expansion:

$$\begin{aligned} 0 = \frac{1}{n}\ell'_{adapt}|_{\hat{\theta}_S} &= \frac{1}{n}\ell'|_{\theta_{0,S}} + \underbrace{\frac{1}{n}\ell''|_{\theta_{0,S}}}_{(1)}\left(\hat{\theta}_S - \theta_{0,S}\right) + \frac{1}{2}\underbrace{\left(\hat{\theta}_S - \theta_{0,S}\right)^T}_{(2)}\underbrace{\frac{1}{n}\ell'''|_{\tilde{\theta}_S}}_{(3)}\left(\hat{\theta}_S - \theta_{0,S}\right) \\ &\quad -\lambda\begin{pmatrix} \gamma\hat{\pi}^{\gamma-1}\sum_{(1,j)\in S} w_{1,j}|\hat{\phi}_{1,j}| - \gamma(1-\hat{\pi})^{\gamma-1}\sum_{(2,j)\in S} w_{2,j}|\hat{\phi}_{2,j}| \\ \hat{\pi}^{\gamma} w_{1,S} sgn(\hat{\phi}_{1,S}) \\ (1-\hat{\pi})^{\gamma} w_{2,S} sgn(\hat{\phi}_{2,S}) \\ 0 \\ 0 \end{pmatrix}. \end{aligned}$$

Now term (1) is of order $-I_S(\theta_0) + o_P(1)$ (law of large numbers); term (2) is of order $o_P(1)$ (consistency); and term (3) is of order $O_P(1)$ (law of large numbers and regularity condition on 3rd derivatives). Therefore we have

$$\sqrt{n}\frac{1}{n}\ell'|_{\theta_{0,S}} + \left(-I_S(\theta_0) + o_P(1)\right)\sqrt{n}(\hat{\theta}_S - \theta_{0,S}) - \sqrt{n}\lambda O_P(1) = 0$$

or

$$\left(-I_S(\theta_0) + o_P(1)\right)\sqrt{n}(\hat{\theta}_S - \theta_{0,S}) - \sqrt{n}\lambda O_P(1) = -\frac{1}{\sqrt{n}}\ell'|_{\theta_{0,S}} \tag{A.27}$$

Notice that $\frac{1}{\sqrt{n}}\ell'|_{\theta_{0,S}} \rightsquigarrow^d \mathcal{N}(0, I_S(\theta_0))$ by the central limit theorem. Furthermore $\sqrt{n}\lambda = o(1)$ as $\lambda = o(n^{-1/2})$.

Therefore $\sqrt{n}(\hat{\theta}_S - \theta_{0,S}) \rightsquigarrow^d \mathcal{N}(0, I_S(\theta_0))$ follows from equation (A.27). □

B Proofs for Section 5

B.1 Proof of Lemma 1

It is clear that

$$\mathcal{E}(\psi|\psi_0) = (\psi - \psi_0)^T I(\psi_0)(\psi - \psi_0)/2 + r_\psi,$$

where

$$|r_\psi| \leq \frac{\|\psi - \psi_0\|_1^3}{6}\int \sup_{\psi\in\Psi}\max_{j_1,j_2,j_3}\left|\frac{\partial^3 l_\psi}{\partial\psi_{j_1}\partial\psi_{j_2}\partial\psi_{j_3}}\right| f_{\psi_0} d\mu$$

$$\leq \frac{d^{3/2}C_3}{6}\|\psi-\psi_0\|_2^3.$$

Hence

$$\mathcal{E}(\psi|\psi_0(x)) \geq \|\psi-\psi_0(x)\|_2^2\Lambda_{\min}^2/2 - d^{3/2}C_3\|\psi-\psi_0(x)\|_2^3/6.$$

Now, apply the auxiliary lemma below, with $K_0^2 = dK^2$, $\Lambda^2 = \Lambda_{\min}^2/2$, and $C = d^{3/2}C_3/6$. □

Auxiliary Lemma. Let $h : [-K_0, K_0] \to [0,\infty)$ have the following properties:

(i) $\forall\ \varepsilon > 0\ \exists\ \alpha_\varepsilon > 0$ such that $\inf_{\varepsilon<|z|\leq K_0} h(z) \geq \alpha_\varepsilon$,

(ii) $\exists\ \Lambda > 0$, $C > 0$, such that $\forall\ |z| \leq K_0$, $h(z) \geq \Lambda^2 z^2 - C|z|^3$.

Then $\forall\ |z| \leq K_0$,

$$h(z) \geq z^2/C_0^2,$$

where

$$C_0^2 = \max\left[\frac{1}{\varepsilon_0}, \frac{K_0^2}{\alpha_{\varepsilon_0}}\right],\ \varepsilon_0 = \frac{\Lambda^2}{2C}.$$

Proof (Auxiliary Lemma)

If $\varepsilon_0 > K_0$, we have $h(z) \geq \Lambda^2 z^2/2$ for all $|z| \leq K_0$.

If $\varepsilon_0 \leq K_0$ and $|z| \leq \varepsilon_0$, we also have $h(z) \geq (\Lambda^2 - \varepsilon_0 C)z^2 \geq \Lambda^2 z^2/2$.

If $\varepsilon_0 \leq K_0$ and $\varepsilon_0 < |z| \leq K_0$, we have $h(z) \geq \alpha_{\varepsilon_0} = K_0^2\alpha_{\varepsilon_0}/K_0^2 \geq |z|^2\alpha_{\varepsilon_0}/K_0^2$. □

B.2 Proof of Lemma 2

In order to prove Lemma 2, we first state and proof a suitable entropy bound:

We introduce the norm

$$\|h(\cdot,\cdot)\|_{P_n} = \sqrt{\frac{1}{n}\sum_{i=1}^n h^2(x_i, Y_i)}.$$

For a collection $\mathcal{H}$ of functions on $\mathcal{X}\times\mathcal{Y}$, we let $H(\cdot, \mathcal{H}, \|\cdot\|_{P_n})$ be the entropy of $\mathcal{H}$ equipped with the metric induced by the norm $\|\cdot\|_{P_n}$.

Define for $\epsilon > 0$,

$$\tilde{\Theta}(\epsilon) = \{\vartheta^T = (\phi_1^T, \ldots, \phi_k^T, \eta^T) \in \tilde{\Theta} :\ \|\phi-\phi_0\|_1 + \ \|\eta-\eta_0\|_2 \leq \epsilon\}.$$

Entropy Lemma For a constant C_0 depending on $\Lambda_{\max}$, k and m, we have for all $u > 0$ and $M_n > 0$,

$$H\left(u, \left\{(L_\vartheta - L_{\vartheta^*})1\{G_1 \leq M_n\} :\ \vartheta \in \tilde{\Theta}(\epsilon)\right\}, \|\cdot\|_{P_n}\right) \leq C_0\frac{\epsilon^2 M_n^2}{u^2}\log\left(\frac{\epsilon M_n}{u}\right).$$

Proof (Entropy Lemma) We have

$$|L_\vartheta(x,y) - L_{\tilde{\vartheta}}(x,y)|^2 \leq G_1^2(y)\left[\sum_{r=1}^k |(\phi_r - \tilde{\phi}_r)^T x| + \|\eta-\tilde{\eta}\|_1\right]^2 \leq dG_1^2(y)\left[\sum_{r=1}^k |(\phi_r-\tilde{\phi}_r)^T x|^2 + \|\eta-\tilde{\eta}\|_2^2\right].$$

It follows that

$$\|(L_\vartheta - L_{\tilde{\vartheta}})1\{G_1 \leq M_n\}\|_{P_n}^2 \leq dM_n^2\left[\sum_{r=1}^k \frac{1}{n}\sum_{i=1}^n |(\phi_r-\tilde{\phi}_r)^T x_i|^2 + \|\eta-\tilde{\eta}\|_2^2\right] \leq dM_n^2\left[\Lambda_{\max}^2\|\phi-\tilde{\phi}\|_2^2 + \|\eta-\tilde{\eta}\|_2^2\right].$$

Let $N(\cdot, \Gamma, \tau)$ denote the covering number of a metric space (Γ, τ) with metric (induced by the norm) τ, and $H(\cdot,\Gamma,\tau) = \log N(\cdot,\Gamma,\tau)$ be its entropy. If Γ is a ball with radius ϵ in Euclidean space $\mathbb{R}^N$, one easily verifies that

$$H(u,\Gamma,\tau) \leq N\log\left(\frac{5\epsilon}{u}\right), \forall\ u > 0.$$

We have for $\|\phi\|_1 \leq \epsilon$,

$$\|\phi\|_2 \leq \|\phi\|_1 \leq \epsilon.$$

Let $u > 0$ be arbitrary. Then for $\|\phi\|_1 \leq \epsilon$,

$$\left\|\phi 1\left\{(j,r) :\ |\phi_{j,r}| \leq u^2/\epsilon\right\}\right\|_2^2 \leq u^2\frac{\|\phi\|_1}{\epsilon} \leq u^2,$$

and moreover

$$\|\phi\|_1 \geq \sum_{|\phi_{j,r}|>u^2/\epsilon} |\phi_{j,r}| \geq u^2 N_u/\epsilon,$$

where $N_u = |\{(j,r):\ |\phi_{j,r}| \leq u^2/\epsilon\}|$. So

$$N_u \leq \frac{\epsilon^2}{u^2}.$$

We have now shown that

$$H(\sqrt{2}u, \{\|\phi\|_1 \leq \epsilon\}, \|\cdot\|_2) \leq \frac{\epsilon^2}{u^2}\log\left(\frac{5\epsilon}{u}\right).$$

Thus, also

$$H(\sqrt{2}u, \{\|\phi-\phi_0\|_1 \leq \epsilon\}, \|\cdot\|_2) \leq \frac{\epsilon^2}{u^2}\log\left(\frac{5\epsilon}{u}\right).$$

We also have

$$H(u, \{\eta\in\mathbb{R}^p:\ \|\eta-\eta_0\|_2 \leq \epsilon\}, \|\cdot\|_2) \leq m\log\left(\frac{5\epsilon}{u}\right).$$

We can therefore conclude that

$$H\Big(\sqrt{d}M_n(\sqrt{2}\Lambda_{\max}+1)u, \Big\{L_\vartheta - L_{\vartheta_0}1\{G_1\leq M_n\}:\ \vartheta\in\tilde{\Theta}(\epsilon)\Big\}, \|\cdot\|_{P_n}\Big) \leq \Big(\frac{\epsilon^2}{u^2}+m\Big)\log\left(\frac{5\epsilon}{u}\right).$$

□

Let's now turn to the main proof of Lemma 2.

In what follows, $\{c_t\}$ are constants depending on $\Lambda_{\max}$, k, m and K. The truncated version of the empirical process is

$$V_n^{\rm trunc}(\vartheta) = \frac{1}{n}\sum_{i=1}^n\Big[L_\vartheta(x_i,Y_i)1\{G_1(Y_i)\leq M_n\} - \mathbb{E}\Big(L_\vartheta(x_i,Y)1\{G_1(Y_i)\leq M_n\}\Big|X=x_i\Big)\Big].$$

Let $\epsilon>0$ be arbitrary. We invoke Lemma 3.2 in van de Geer (2000), combined with a symmetrization lemma (e.g., a conditional version of Lemma 3.3 in van de Geer (2000)). We apply these lemmas to the class

$$\Big\{(L_\theta - L_{\vartheta_0})1\{G_1\leq M_n\}:\ \vartheta\in\tilde{\Theta}(\epsilon)\Big\}.$$

In the notation used in Lemma 3.2 of van de Geer (2000), we take $\delta = c_4 T\epsilon M_n\sqrt{\log^3 n/n}$, and $R = c_5(\epsilon\wedge 1)M_n$. This then gives

$$\mathbb{P}\Big(\sup_{\vartheta\in\tilde{\Theta}(\epsilon)}|V_n^{\rm trunc}(\vartheta)-V_n^{\rm trunc}(\vartheta_0)| \geq c_6\epsilon T M_n\sqrt{\frac{\log^3 n}{n}}\Big) \leq c_7\exp\Big[-\frac{T^2\log^3 n(\epsilon^2\vee 1)}{c_8^2}\Big].$$

Here, we use the bound (for $0<a\leq 1$),

$$\int_a^1 \frac{1}{u}\sqrt{\log\Big(\frac{1}{u}\Big)}du \leq \log^{3/2}\Big(\frac{1}{a}\Big).$$

We then invoke the peeling device: split the set $\tilde{\Theta}$ into sets

$$\{\vartheta\in\tilde{\Theta}:\ 2^{-(j+1)} \leq \|\phi-\phi_0\|_1 + \|\eta-\eta_0\|_2 \leq 2^{-j}\},$$

where $j\in\mathbb{Z}$, and $\leq 2^{-j+1}\geq\lambda_0$. There are no more than $c_9\log n$ indices $j\leq 0$ with $2^{-j+1}\geq\lambda_0$. Hence, we get

$$\sup_{\vartheta^T=(\phi^T,\eta^T)\in\tilde{\Theta}}\frac{\Big|V_n^{\rm trunc}(\vartheta)-V_n^{\rm trunc}(\vartheta_0)\Big|}{(\|\phi-\phi^*\|_1+\|\eta-\eta^*\|_2)\vee\lambda_0} \leq 2c_6 T M_n\sqrt{\frac{\log^3 n}{n}},$$

with $\mathbb{P}_{\mathbf{x}}$ probability at least

$$1 - c_7[c_9\log n]\exp\Big[-\frac{T^2\log^3 n}{c_8^2}\Big] - \sum_{j=1}^\infty c_7\exp\Big[-\frac{T^2 2^{2j}\log^3 n}{c_8^2}\Big] \geq 1 - c_2\exp\Big[-\frac{T^2\log^3 n}{c_{10}^2}\Big].$$

Finally, to remove the truncation, we use

$$|(L_\vartheta(x,y)-L_{\vartheta_0}(x,y))1\{G_1(y)>M_n\}| \leq dKG_1(y)1\{G_1(y)>M_n\}.$$

Hence

$$\frac{\Big|(V_n^{\rm trunc}(\vartheta)-V_n^{\rm trunc}(\vartheta_0)) - (V_n(\vartheta)-V_n(\vartheta_0))\Big|}{(\|\phi-\phi^*\|_1+\|\eta-\eta^*\|_2)\vee\lambda_0}$$
$$\leq \frac{dK}{n\lambda_0}\sum_{i=1}^n\Big[G_1(Y_i)1\{G_1(Y_i)>M_n\} + \mathbb{E}\Big(G_1(Y_i)1\{G_1(Y_i)>M_n\}\Big|X=x_i\Big)\Big].$$

□

B.3 Proof of Theorem 3

On $\mathcal{T}$

$$\bar{\mathcal{E}}(\hat{\psi}|\psi_0) + \lambda\|\hat{\phi}\|_1 \leq T\lambda_0\Big[(\|\hat{\phi}-\phi_0\|_1 + \|\hat{\eta}-\eta_0\|_2) \vee \lambda_0\Big] + \lambda\|\phi_0\|_1 + \bar{\mathcal{E}}(\psi_0|\psi_0).$$

By Lemma 1,

$$\bar{\mathcal{E}}(\hat{\psi}|\psi_0) \geq \|\hat{\psi}-\psi_0\|_{Q_n}^2/c_0^2,$$

and $\bar{\mathcal{E}}(\psi_0|\psi_0) = 0$.

Case 1 Suppose that

$$\|\hat{\phi}-\phi_0\|_1 + \|\hat{\eta}-\eta_0\|_2 \leq \lambda_0.$$

Then we find

$$\begin{aligned}\bar{\mathcal{E}}(\hat{\psi}|\psi_0) &\leq T\lambda_0^2 + \lambda\|\hat{\phi}-\phi_0\|_1 + \bar{\mathcal{E}}(\psi_0|\psi_0)\\ &\leq (\lambda + T\lambda_0)\lambda_0.\end{aligned}$$

Case 2 Suppose that

$$\|\hat{\phi}-\phi_0\|_1 + \|\hat{\eta}-\eta_0\|_2 \geq \lambda_0,$$

and that

$$T\lambda_0\|\hat{\eta}-\eta_0\|_2 \geq (\lambda + T\lambda_0)\|\hat{\phi}_S - (\phi_0)_S\|_1.$$

Then we get

$$\begin{aligned}&\bar{\mathcal{E}}(\hat{\psi}|\psi_0) + (\lambda - T\lambda_0)\|\hat{\phi}_{S^c}\|_1 \leq 2T\lambda_0\|\hat{\eta}-\eta_0\|_2\\ &\leq 4T^2\lambda_0^2c_0^2 + \|\hat{\eta}-\eta_0\|_2^2/(2c_0^2)\\ &\leq 4T^2\lambda_0^2c_0^2 + \bar{\mathcal{E}}(\hat{\psi}|\psi_0)/2.\end{aligned}$$

So then

$$\bar{\mathcal{E}}(\hat{\psi}|\psi_0) + 2(\lambda - T\lambda_0)\|\hat{\phi}_{S^c}\|_1 \leq 8T^2\lambda_0^2c_0^2.$$

Case 3 Suppose that

$$\|\hat{\phi}-\phi_0\|_1 + \|\hat{\eta}-\eta_0\|_2 \geq \lambda_0,$$

and that

$$T\lambda_0\|\hat{\eta}-\eta_0\|_2 \leq (\lambda + T\lambda_0)\|\hat{\phi}_S - (\phi_0)_S\|_1.$$

Then we have

$$\bar{\mathcal{E}}(\hat{\psi}|\psi_0) + (\lambda - T\lambda_0)\|\hat{\phi}_{S^c}\|_1 \leq 2(\lambda + T\lambda_0)\|\hat{\phi}_S - \phi_0\|_1.$$

So then

$$\|\hat{\phi}_{S^c}\|_1 \leq 6\|\hat{\phi}_S - (\phi_0)_S\|_1.$$

We can then apply the restricted eigenvalue condition to $\hat{\phi}-\phi_0$. This gives

$$\begin{aligned}\bar{\mathcal{E}}(\hat{\psi}|\psi_0) + (\lambda - T\lambda_0)\|\hat{\phi}_{S^c}\|_1 &\leq 2(\lambda + T\lambda_0)\sqrt{s}\|\hat{\phi}_S - \phi_0\|_2.\\ &\leq 2(\lambda + T\lambda_0)\sqrt{s}\kappa\|\hat{g}-g_0\|_{Q_n}\\ &\leq 4(\lambda + T\lambda_0)^2c_0^2\kappa^2 s + \bar{\mathcal{E}}(\hat{\psi}|\psi_0)/2.\end{aligned}$$

So we arrive at

$$\bar{\mathcal{E}}(\hat{\psi}|\psi_0) + 2(\lambda - T\lambda_0)\|\hat{\phi}_{S^c}\|_1 \leq 8(\lambda + T\lambda_0)^2c_0^2\kappa^2 s.$$

□

B.4 Proof of Lemma 3

Let Z be a standard normal random variable. Then by straightforward computations, for all $M > 0$,

$$E|Z|1\{|Z| > M\} \leq 2\exp[-M^2/2],$$

and

$$E|Z|^2 1\{|Z| > M\} \leq (M+2)\exp[-M^2/2].$$

Thus, for n independent copies $Z_1, \ldots, Z_n$ of Z, and $M = 2\sqrt{\log n}$,

$$\begin{aligned}&\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n |Z_i|1\{|Z_i| > M\} > \frac{4\log n}{n}\right)\\ &\leq \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n |Z_i|1\{|Z_i| > M\} - E|Z|1\{|Z| > M\} > \frac{2\log n}{n}\right)\\ &\leq \frac{nE|Z|^2 1\{|Z| > M\}}{4(\log n)^2} \leq \frac{2}{n}.\end{aligned}$$

The result follows from this, as

$$G_1(Y) = \mathrm{e}^K|Y| + K,$$

and Y has a normal mixture distribution. □

B.5 Proof of Theorem 5

On $\mathcal{T}$, defined in (5.18) with $\lambda_0 = c_4\sqrt{\log^4(n)/n}$ (c_4 as in Lemma 3; i.e. $M_n = c_4\sqrt{\log(n)}$ in (5.20)), we have the basic inequality

$$\bar{\mathcal{E}}(\hat{\psi}|\psi_0) + \lambda\|\hat{\phi}\|_1 \leq T\lambda_0\Big[(\|\hat{\phi} - \phi_0\|_1 + \|\hat{\eta} - \eta_0\|_2) \vee \lambda_0\Big] + \lambda\|\phi_0\|_1 + \bar{\mathcal{E}}(\psi_0|\psi_0).$$

Note that $\|\hat{\eta} - \eta_0\|_2 \leq 2K$ and $\bar{\mathcal{E}}(\psi_0|\psi_0) = 0$. Hence, for n sufficiently large,

$$\begin{aligned}\bar{\mathcal{E}}(\hat{\psi}|\psi_0) + \lambda\|\hat{\phi}\|_1 &\leq T\lambda_0(\|\hat{\phi} - \phi_0\|_1 + 2K) + \lambda\|\phi_0\|_1 + \bar{\mathcal{E}}(\psi_0|\psi_0)\\ &\leq T\lambda_0(\|\hat{\phi}\|_1 + \|\phi_0\|_1 + 2K) + \lambda\|\phi_0\|_1 + \bar{\mathcal{E}}(\psi_0|\psi_0),\end{aligned}$$

and therefore also

$$\bar{\mathcal{E}}(\hat{\psi}|\psi_0) + (\lambda - T\lambda_0)\|\hat{\phi}\|_1 \leq T\lambda_0 2K + (\lambda + T\lambda_0)\|\phi_0\|_1 + \bar{\mathcal{E}}(\psi_0|\psi_0).$$

It holds that $\lambda \geq 2T\lambda_0$ (since $\lambda = C\sqrt{\log^4(n)/n}$ for some $C > 0$ sufficiently large), $\lambda_0 = O(\sqrt{\log^4(n)/n})$ and $\lambda = O(\sqrt{\log^4(n)/n})$, and due to the assumption about $\|\phi_0\|_1$ we obtain on the set $\mathcal{T}$ that $\bar{\mathcal{E}}(\hat{\psi}|\psi_0) \to \bar{\mathcal{E}}(\psi_0|\psi_0) = 0$ ($n \to \infty$). Finally, the set $\mathcal{T}$ has large probability, as shown by Lemma 2 and using Proposition 3 and Lemma 3 for FMR models. □

C Proofs for Sections 3 and 6

C.1 Proof of Proposition 2

We restrict ourselves to a two class mixture with k = 2. Consider the function $u(\xi)$ defined as

$$\begin{aligned}&u(\xi) = \exp(\ell^{(0)}_{pen}(\xi))\\ &\propto \prod_{i=1}^{n}\left\{\left(\pi\frac{1}{\sigma_1}e^{\frac{-(Y_i - X_i'\beta_1)^2}{2\sigma_1^2}} + (1-\pi)\frac{1}{\sigma_2}e^{\frac{-(Y_i - X_i'\beta_2)^2}{2\sigma_2^2}}\right)e^{-\lambda\frac{||\beta_1||_1}{\sigma_1}}e^{-\lambda\frac{||\beta_2||_1}{\sigma_2}}\right\}.\end{aligned} \tag{C.28}$$

We will show that $u(\xi)$ is bounded from above on $\xi = (\sigma_1, \sigma_2, \beta_1, \beta_2, \pi) \in \Xi = \mathbb{R}^2_{>0} \times \mathbb{R}^{2p} \times [0, 1]$. Then clearly $-n^{-1}\ell^{(0)}_{pen}(\theta)$ is bounded from below on $\theta = (\rho_1, \rho_2, \phi_1, \phi_2, \pi) \in \Theta = \mathbb{R}^2_{>0} \times \mathbb{R}^{2p} \times (0, 1)$.

The critical point for unboundedness is if we choose for an arbitrary sample point $i \in 1, \ldots, n$ a β_1^* such that $Y_i - X_i'\beta_1^* = 0$ and let $\sigma_1 \to 0$. Without the penalty term $\exp(-\lambda\frac{||\beta_1^*||_1}{\sigma_1})$ in (C.28) the function would tend to infinity as $\sigma_1 \to 0$. But as $Y_i \neq 0$ for all $i \in 1, \ldots, n$, β_1^* cannot be zero and therefore $\exp(-\lambda\frac{||\beta_1^*||_1}{\sigma_1})$ forces $u(\xi)$ to tend to 0 as $\sigma_1 \to 0$.

Let's give a more formal proof for boundedness of $u(\xi)$. Choose a small $0 < \varepsilon_1 < \min Y_i^2$ and $\varepsilon_2 > 0$. As $Y_i \neq 0$, $i = 1 \ldots n$, there exists a small constant $m > 0$ such that

$$0 < \min Y_i^2 - \varepsilon_1 \leq (Y_i - X_i\beta_1)^2 \tag{C.29}$$

holds for all $i = 1 \ldots n$ as long as $||\beta_1||_1 < m$ and

$$0 < \min Y_i^2 - \varepsilon_1 \leq (Y_i - X_i\beta_2)^2 \tag{C.30}$$

holds for all $i = 1 \ldots n$ as long as $||\beta_2||_1 < m$.

Furthermore there exists a small constant $\delta > 0$ such that

$$\frac{1}{\sigma_1}e^{-\frac{(\min Y_i^2 - \varepsilon_1)}{2\sigma_1^2}} < \varepsilon_2 \quad \text{and} \quad \frac{1}{\sigma_1}e^{-\lambda\frac{m}{\sigma_1}} < \varepsilon_2 \tag{C.31}$$

holds for all $0 < \sigma_1 < \delta$ and

$$\frac{1}{\sigma_2}e^{-\frac{(\min Y_i^2 - \varepsilon_1)}{2\sigma_2^2}} < \varepsilon_2 \quad \text{and} \quad \frac{1}{\sigma_2}e^{-\lambda\frac{m}{\sigma_2}} < \varepsilon_2 \tag{C.32}$$

holds for all $0 < \sigma_2 < \delta$.

Define the set $K = \{(\sigma_1, \sigma_2, \beta_1, \beta_2, \pi) \in \Xi;\ \delta \leq \sigma_1, \sigma_2\}$. Now $u(\xi)$ is trivially bounded on K. From the construction of K and equations (C.29)-(C.32) we easily see that $u(\xi)$ is also bounded on K^c and therefore bounded on Ξ. □

C.2 Proof of Theorem 6

The density of the complete data is given by

$$f_c(Y,\Delta|\theta)=\prod_{i=1}^{n}\prod_{r=1}^{k}\pi_r^{\Delta_{i,r}}\left(\frac{\rho_r}{\sqrt{2\pi}}e^{-\frac{1}{2}(\rho_r Y_i-X_i'\phi_r)^2}\right)^{\Delta_{i,r}},$$

whereas the density of the observed data is given by

$$f_{obs}(Y|\theta)=\prod_{i=1}^{n}\sum_{r=1}^{k}\pi_r\frac{\rho_r}{\sqrt{2\pi}}e^{-\frac{1}{2}(\rho_r Y_i-X_i'\phi_r)^2}$$

$$\theta=(\rho_1,\ldots,\rho_k,\phi_{1,1},\phi_{1,2},\ldots,\phi_{k,p},\pi)\in\Theta=\mathbb{R}^k_{>0}\times\mathbb{R}^{kp}\times\Pi\subset\mathbb{R}^{k(p+2)-1}$$
$$\Pi=\{\pi=(\pi_1,\ldots,\pi_{k-1});\pi_r>0\text{ for }r=1,\ldots,k-1\text{ and }\sum_{r=1}^{k-1}\pi_r<1\},\qquad \pi_k=1-\sum_{r=1}^{k-1}\pi_r.$$

Furthermore the conditional density of the complete data given the observed data is given by $k(Y,\Delta|Y,\theta)=f_c(Y,\Delta|\theta)/f_{obs}(Y|\theta)$. Then, the penalized negative log-likelihood fulfills the equation

$$\nu_{pen}(\theta)=-n^{-1}\ell^{(0)}_{pen,\lambda}(\theta)=-n^{-1}\log f_{obs}(Y|\theta)+\lambda\sum_{r=1}^{k}||\phi_r||_1=Q_{pen}(\theta|\theta')-H(\theta|\theta') \tag{C.33}$$

where $Q_{pen}(\theta|\theta')=-n^{-1}\mathbb{E}_{\theta'}[\log f_c(Y,\Delta|\theta)|Y]+\lambda\sum_{r=1}^{k}||\phi_r||_1$ (compare Section 6.1) and $H(\theta|\theta')=-n^{-1}\mathbb{E}_{\theta'}[\log k(Y,\Delta|Y,\theta)|Y]$. By Jensen's inequality we get the following important relationship:

$$H(\theta|\theta')\geq H(\theta'|\theta')\qquad\forall\quad\theta\in\Theta, \tag{C.34}$$

see also Wu (1983). $Q_{pen}(\theta|\theta')$ and $H(\theta|\theta')$ are continuous functions in θ and θ'. If we think of them as functions in θ with fixed θ' we write also $Q_{pen,\theta'}(\theta)$ and $H_{\theta'}(\theta)$. Furthermore $Q_{pen,\theta'}(\theta)$ is a convex function in θ and strictly convex in each coordinate of θ. As a last preparation we give a definition of a stationary point for non-differentiable functions (see also Tseng (2001)):

Definition 1 Let u be a function defined on a open set $U\subset\mathbb{R}^{k(p+2)-1}$. $x\in U$ is called stationary point if $u'(x;d)=\lim_{\alpha\downarrow 0}\frac{u(x+\alpha d)-u(x)}{\alpha}\geq 0\quad\forall d\in\mathbb{R}^{k(p+2)-1}$.

We are now ready to start with the proof which is inspired by Bertsekas (1995). We write $\theta=(\theta_1,\ldots,\theta_D)=(\rho_1,\ldots,\rho_k,\phi_{1,1},\phi_{1,2},\ldots,\phi_{k,p},\pi)$ where $D=k+kp+1$ denotes the number of coordinates. Remark that the first D-1 coordinates are univariate, whereas $\theta_D=\pi$ is a "block coordinate" of dimension $k-1$.

Proof Let $\theta^m=\theta^{(m)}$ be the sequence generated by the BCD-GEM algorithm. We need to prove that for a converging subsequence $\theta^{m_j}\to\bar\theta\in\Theta$, $\bar\theta$ is a stationary point of $\nu_{pen}(\theta)$. Taking directional derivatives of equation (C.33) yields

$$\nu'_{pen}(\bar\theta;d)=Q'_{pen,\bar\theta}(\bar\theta;d)-\langle\bigtriangledown H_{\bar\theta}(\bar\theta),d\rangle.$$

Note that $\bigtriangledown H_{\bar\theta}(\bar\theta)=0$ as $H_{\bar\theta}(x)$ is minimized for $x=\bar\theta$ (equation (C.34)). Therefore it remains to show that $Q'_{pen,\bar\theta}(\bar\theta;d)\geq 0$ for all directions d. Let

$$z_i^m=(\theta_1^{m+1},\ldots,\theta_i^{m+1},\theta_{i+1}^m,\ldots,\theta_D^m).$$

Using the definition of the algorithm we have:

$$Q_{pen,\theta^m}(\theta^m)\geq Q_{pen,\theta^m}(z_1^m)\geq\ldots\geq Q_{pen,\theta^m}(z_{D-1}^m)\geq Q_{pen,\theta^m}(\theta^{m+1}). \tag{C.35}$$

Additionally from the properties of GEM (equation (C.33) and (C.34)) we have:

$$\nu_{pen}(\theta^0)\geq\nu_{pen}(\theta^1)\geq\ldots\geq\nu_{pen}(\theta^m)\geq\nu_{pen}(\theta^{m+1}). \tag{C.36}$$

Equation (C.36) and the converging subsequence imply that the sequence $\{\nu_{pen}(\theta^m);m=0,1,2,\ldots\}$ converges to $\nu_{pen}(\bar\theta)$. Further we have:

$$\begin{aligned}0\leq Q_{pen,\theta^m}(\theta^m)-Q_{pen,\theta^m}(\theta^{m+1})&=\nu_{pen}(\theta^m)-\nu_{pen}(\theta^{m+1})+\underbrace{H_{\theta^m}(\theta^m)-H_{\theta^m}(\theta^{m+1})}_{\leq 0}\\&\leq\underbrace{\nu_{pen}(\theta^m)-\nu_{pen}(\theta^{m+1})}_{\to\nu_{pen}(\bar\theta)-\nu_{pen}(\bar\theta)=0}.\end{aligned} \tag{C.37}$$

We conclude that the sequence $\{Q_{pen,\theta^m}(\theta^m) - Q_{pen,\theta^m}(\theta^{m+1}); m = 0, 1, 2, \ldots\}$ converges to zero.

We now show that $\{\theta_1^{m_j+1} - \theta_1^{m_j}\}$ converges to zero ($j \to \infty$). Assume the contrary, in particular that $\{z_1^{m_j} - \theta^{m_j}\}$ does not converge to 0. Let $\gamma^{m_j} = ||z_1^{m_j} - \theta^{m_j}||$. Without loss of generality (by restricting to a subsequence) we may assume that there exists some $\bar{\gamma} > 0$ such that $\gamma^{m_j} > \bar{\gamma}$ for all j. Let $s_1^{m_j} = \frac{z_1^{m_j} - \theta^{m_j}}{\gamma^{m_j}}$. $s_1^{m_j}$ differs from zero only along the first component. As $s_1^{m_j}$ belongs to a compact set ($||s_1^{m_j}|| = 1$) we may assume that $s_1^{m_j}$ converges to $\bar{s}_1$. Let us fix some $\varepsilon \in [0, 1]$. Notice that $0 \leq \varepsilon\bar{\gamma} \leq \gamma^{m_j}$. Therefore, $\theta^{m_j} + \varepsilon\bar{\gamma}s_1^{m_j}$ lies on the segment joining θ^{m_j} and $z_1^{m_j}$, and belongs to Θ because Θ is convex. As $Q_{pen,\theta^{m_j}}(.)$ is convex and $z_1^{m_j}$ minimizes this function over all values that differ from θ^{m_j} along the first coordinate, we obtain

$$Q_{pen,\theta^{m_j}}(z_1^{m_j}) = Q_{pen,\theta^{m_j}}(\theta^{m_j} + \gamma^{m_j}s_1^{m_j}) \leq Q_{pen,\theta^{m_j}}(\theta^{m_j} + \varepsilon\bar{\gamma}s_1^{m_j}) \leq Q_{pen,\theta^{m_j}}(\theta^{m_j}). \tag{C.38}$$

From equation (C.35) and (C.38) we conclude

$$\begin{aligned} 0 \quad &\leq \quad Q_{pen,\theta^{m_j}}(\theta^{m_j}) - Q_{pen,\theta^{m_j}}(\theta^{m_j} + \varepsilon\bar{\gamma}s_1^{m_j}) \overbrace{\leq}^{(C.38)} Q_{pen,\theta^{m_j}}(\theta^{m_j}) - Q_{pen,\theta^{m_j}}(z_1^{m_j}) \\ &\overbrace{\leq}^{(C.35)} Q_{pen,\theta^{m_j}}(\theta^{m_j}) - Q_{pen,\theta^{m_j}}(\theta^{m_j+1}). \end{aligned}$$

Using (C.37) and continuity of $Q_{pen,x}(y)$ in both arguments x and y we conclude by taking the limit $j \to \infty$:

$$Q_{pen,\bar{\theta}}(\bar{\theta} + \varepsilon\bar{\gamma}\bar{s_1}) = Q_{pen,\bar{\theta}}(\bar{\theta}) \quad \forall\varepsilon \in [0, 1].$$

Since $\bar{\gamma}\bar{s_1} \neq 0$ this contradicts the strict convexity of $Q_{pen,\bar{\theta}}(x_1, \bar{\theta}_2, \ldots, \bar{\theta}_D)$ as a function of the first block-coordinate. This contradiction establishes that $z_1^{m_j}$ converges to $\bar{\theta}$.

From the definition of the algorithm we have:

$$Q_{pen}(z_1^{m_j}|\theta^{m_j}) \leq Q_{pen}(x_1, \theta_2^{m_j}, \ldots, \theta_D^{m_j}|\theta^{m_j}) \qquad \forall x_1.$$

By continuity and taking the limit $j \to \infty$ we obtain:

$$Q_{pen,\bar{\theta}}(\bar{\theta}) \leq Q_{pen,\bar{\theta}}(x_1, \bar{\theta}_2, \ldots, \bar{\theta}_D) \qquad \forall x_1.$$

Repeating the argument we conclude that $\bar{\theta}$ is a coordinate-wise minimum. Therefore, following Tseng (2001), $\bar{\theta}$ is easily seen to be a stationary point of $Q_{pen,\bar{\theta}}(.)$, in particular $Q'_{pen,\bar{\theta}}(\bar{\theta}; d) \geq 0$ for all directions d. □

Acknowledgements We would like to thank some referees for constructive comments and the co-editors Ricardo Cao and Domingo Morales for inviting us to present this discussion paper. N.S. acknowledges financial support from Novartis International AG, Basel, Switzerland.

References

Bertsekas D (1995) Nonlinear programming. Athena Scientific, Belmont, MA

Bickel P, Ritov Y, Tsybakov A (2009) Simultaneous analysis of Lasso and Dantzig selector. Annals of Statistics 37:1705–1732

Bunea F, Tsybakov A, Wegkamp M (2007) Sparsity oracle inequalities for the Lasso. Electronic Journal of Statistics 1:169–194

Cai T, Wang L, Xu G (2009a) Stable recovery of sparse signals and an oracle inequality. Tech. rep., Department of Statistics, University of Pennsylvania

Cai T, Xu G, Zhang J (2009b) On recovery of sparse signals via ℓ_1 minimization. IEEE Transactions on Information Theory 55:3388–3397

Candès E, Plan Y (2009) Near-ideal model selection by ℓ_1 minimization. Annals of Statistics 37:2145–2177

Candès E, Tao T (2005) Decoding by linear programming. IEEE Transactions on Information Theory 51:4203–4215

Candès E, Tao T (2007) The Dantzig selector: statistical estimation when p is much larger than n (with discussion). Annals of Statistics 35:2313–2404

Dempster A, Laird N, Rubin D (1977) Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B 39:1–38

Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association 96:1348–1360

Friedman J, Hastie T, Hoefling H, Tibshirani R (2007) Pathwise coordinate optimization. Annals of Applied Statistics 1:302–332

Friedman J, Hastie T, Tibshirani R (2008) Regularized paths for generalized linear models via coordinate descent. Tech. rep., Department of Statistics, Stanford University

Fu WJ (1998) Penalized regression: the Bridge versus the Lasso. Journal of Computational and Graphical Statistics 7:397–416
van de Geer S (2000) Empirical Processes in M-Estimation. Cambridge University Press
van de Geer S (2008) High-dimensional generalized linear models and the Lasso. Annals of Statistics 36:614–645
van de Geer S, Bühlmann P (2009) On the conditions used to prove oracle results for the Lasso. Arxiv preprint mathST/09100722
Greenshtein E, Ritov Y (2004) Persistence in high-dimensional predictor selection and the virtue of over-parametrization. Bernoulli 10:971–988
Grün B, Leisch F (2007) Fitting finite mixtures of generalized linear regressions in R. Computational Statistics & Data Analysis 51:5247–5252, DOI 10.1016/j.csda.2006.08.014
Grün B, Leisch F (2008) FlexMix version 2: Finite mixtures with concomitant variables and varying and constant parameters. Journal of Statistical Software 28:1–35, URL http://www.jstatsoft.org/v28/i04/
Huang J, Ma S, Zhang CH (2008) Adaptive Lasso for sparse high-dimensional regression models. Statista Sinica 18:1603–1618
Khalili A, Chen J (2007) Variable selection in finite mixture of regression models. Journal of the American Statistical Association 102:1025–1038
Koltchinskii V (2009) The Dantzig selector and sparsity oracle inequalities. Bernoulli 15:799–828
Lehmann E (1983) Theory of Point Estimation. Pacific Grove, CA: Wadsworth and Brooks/Cole
Leisch F (2004) FlexMix: A general framework for finite mixture models and latent class regression in R. Journal of Statistical Software 11:1–18, URL http://www.jstatsoft.org/v11/i08/
McLachlan, Peel (2000) Finite mixture models. Wiley, New York
Meier L, van de Geer S, Bühlmann P (2008) The group Lasso for logistic regression. Journal of the Royal Statistical Society, Series B 70:53–71
Meinshausen N, Bühlmann P (2006) High dimensional graphs and variable selection with the Lasso. Annals of Statistics 34:1436–1462
Meinshausen N, Yu B (2009) Lasso-type recovery of sparse representations for high-dimensional data. Annals of Statistics 37:246–270
Pan W, Shen X (2007) Penalized model-based clustering with application to variable selection. Journal of Machine Learning Research 8:1145–1164
Park T, Casella G (2008) The bayesian Lasso. Journal of the American Statistical Association 103:681–686
Tibshirani R (1996) Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society, Series B 58:267–288
Tseng P (2001) Convergence of a block coordinate descent method for nondifferentiable minimization. Journal of Optimization Theory and Applications 109:475–494
Tseng P, Yun S (2008) A coordinate gradient descent method for nonsmooth separable minimization. Mathematical Programming, Series B 117:387–423
Tsybakov A (2004) Optimal aggregation of classifiers in statistical learning. Annals of Statistics 32:135–166
van der Vaart A (2007) Asymptotic Statistics. Cambridge University Press
Wainwright M (2009) Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). IEEE Transactions on Information Theory 55:2183–2202
Wu C (1983) On the convergence properties of the EM algorithm. Annals of Statistics 11:95–103
Zhang CH (2009a) Nearly unbiased variable selection under minimax concave penalty. Annals of Statistics, to appear
Zhang CH, Huang J (2008) The sparsity and bias of the Lasso selection in high-dimensional linear regression. Annals of Statistics 36:1567–1594
Zhang T (2009b) Some sharp performance bounds for least squares regression with L1 regularization. Annals of Statistics 37:2109–2144
Zhao P, Yu B (2006) On model selection consistency of Lasso. Journal of Machine Learning Research 7:2541–2563
Zhou S, van de Geer S, Bühlmann P (2009) Adaptive Lasso for high dimensional regression and Gaussian graphical modeling. Arxiv preprint mathST/09032515
Zou H (2006) The adaptive Lasso and its oracle properties. Journal of the American Statistical Association 101:1418–1429