# $\ell_1$-Penalization for Mixture Regression Models

Nicolas Städler, Peter Bühlmann and Sara van de Geer

June 3, 2009

**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 $\ell_1$-penalized maximum likelihood estimator, in an appropriate parameterization, and 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.

## 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 modelled 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. subpopulations). Hence, addressing the issue of heterogeneous data in high-dimensional data is an important need in many practical applications. We will empirically demonstrate with real data in Section 7.2 that prediction improvements are possible by incorporating a heterogeneity structure to the model.

We propose here an $\ell_1$-penalized method, i.e. a Lasso-type estimator (Tibshirani, 1996), for estimating a high-dimensional finite mixture of regressions model where $p \gg n$. Our approach is related to the proposal in Khalili and Chen (2007). In this paper, 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, 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 for non-convex loss functions, as the negative log-likelihood in a mixture model is non-convex, is rather non-standard. Much of the theory

for the high-dimensional Lasso has 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; Bickel *et al.*, 2007) or the negative log-likelihood in a generalized linear model (van de Geer, 2008). From this point of view, we present a non-trivial modification of the mathematical analysis of $\ell_1$-penalized to non-convex but smooth likelihood problems.

The rest of this article is organized as follows: Section 2 describes the finite mixture of regressions (FMR) model with an appropriate parameterization, Section 3 introduces $\ell_1$-penalized maximum-likelihood estimation for FMR models, Sections 4 and 5 present mathematical theory for the low- and high-dimensional case, Section 6 develops some efficient generalized EM algorithm and presents its numerical convergence properties and Section 7 reports on simulations and a real data example.

## 2 Finite mixture of Gaussian regressions model

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

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

Thereby, $X_i \in \mathbb{R}^p$ are fixed or random covariates, $Y_i \in \mathbb{R}$ is a univariate response variable and $\xi = (\beta_1, \ldots, \beta_k, \sigma_1, \ldots, \sigma_k, \pi_1, \ldots, \pi_{k-1})$ denotes the $(p+2)\cdot k - 1$ free parameters and $\pi_k$ 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 $\beta_r$ and error variances $\sigma_r^2$. We are particularly interested in the case where $p \gg n$.

### 2.1 Reparameterized mixture of regressions model

We will 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

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

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

equals:

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

This is the main model we are analysing and working with.

The log-likelihood function of this model equals:

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

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 $\ell_1$-norm penalized MLE which is different from a naive $\ell_1$-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 $\ell_1$-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 $\ell_1$-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 $\varepsilon$, $p \times 1$ vector $\beta$ and $n \times p$ matrix $\mathbf{X}$. In the sequel, $\|.\|$ denotes the Euclidean norm. The $\ell_1$-norm penalized estimator, called the Lasso (Tibshirani (1996)), is defined as:

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

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 $\beta$ and using the $\ell_1$-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 $\sigma^2$.

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

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

Note that we are penalizing only the $\beta$-parameter. However, the scale parameter $\sigma^2$ is influenced indirectly by the amount of shrinkage $\lambda$.

There are two main drawbacks of the estimator in (3.6). First, it is not invariant under scaling of the response, i.e. $b \cdot Y$ $(b > 0)$ leads to a different estimator than with $b = 1$. 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:

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

This estimator is invariant under scaling $b \cdot Y$. It penalizes the $\ell^1$-norm of the coefficients and small variances $\sigma^2$ simultaneously. Furthermore, it is closely related to the Bayesian Lasso (Park and Casella, 2008). They consider a fully Bayesian analysis using 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 the noninformative scale-invariant marginal prior $p(\sigma^2) = 1/\sigma^2$ for $\sigma^2$. They argue that conditioning on $\sigma^2$ 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 invariant under scaling and whose computation involves convex optimization:

$$\hat{\phi}_\lambda, \hat{\rho}_\lambda = \arg\min_{\phi,\rho}(-\log(\rho) + \frac{1}{2n}||\rho Y - X\phi||^2 + \lambda||\phi||_1). \tag{3.7}$$

From an algorithmic point of view, fast algorithms are available to solve the optimization in (3.7). Shooting algorithms (Fu, 1998) with coordinatewise descent are especially suitable, as demonstrated by Meier *et al.* (2008), Friedman *et al.* (2007). We describe in Section

6.1 an algorithm for estimation in a mixture of regressions model, a more complex task than the optimization for (3.7). 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 (3.7) 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 3.1.** *Every solution $(\hat{\phi}, \hat{\rho})$ of (3.7) satisfies:*

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

*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 $\ell_1$-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 (3.7), 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:

$$\hat{\theta}_\lambda^{(\gamma)} = \arg\min_{\theta \in \Theta} -n^{-1} \ell_{pen,\lambda}^{(\gamma)}(\theta), \quad \theta = (\phi_1, \ldots, \phi_k, \rho_1, \ldots, \rho_k, \pi_1, \ldots, \pi_{k-1}), \quad (3.8)$$

$$-n^{-1} \ell_{pen,\lambda}^{(\gamma)}(\theta) = -n^{-1} \sum_{i=1}^n \log \left( \sum_{r=1}^k \pi_r \frac{\rho_r}{\sqrt{2\pi}} \exp(-\frac{1}{2}(\rho_r Y_i - X_i^T \phi_r)^2) \right) + \lambda \sum_{r=1}^k \pi_r^\gamma ||\phi_r||_1, \quad (3.9)$$

$$\Theta = \mathbb{R}^{kp} \times \mathbb{R}_{>0}^k \times \Pi, \quad (3.10)$$

where $\Pi = \{\pi; \pi_r > 0 \text{ for } r = 1, \ldots, k-1 \text{ and } \sum_{r=1}^{k-1} \pi_r < 1\}$. 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 $\pi_r$. As we will see in Sections 6.1 and 6.2, the optimization for computing $\hat{\theta}_\lambda^{(0)}$ is easiest and we are able to 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 $\pi_r$'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 $\ell_1$-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 3.2.** *Assume that $Y_i \neq 0$ for all $i = 1, \ldots, n$. Then the penalized negative likelihood $-n^{-1}\ell^{(0)}_{pen,\lambda}(\theta)$ is bounded from below for all values $\theta \in \Theta$ from (3.10).*

A proof is given in Appendix C. Even though Proposition 3.2 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 $\ell_1$-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 $\lambda$. Thus, we can do variable selection as follows. Denote by

$$\widehat{S} = \Big\{(r, j);\ 1 \leq r \leq k,\ 1 \leq j \leq p, \hat{\phi}_{r,j} \neq 0\Big\}. \tag{3.11}$$

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, $\widehat{S} = \widehat{S}^{(\gamma)}_{\lambda}$ depends on the specification of the regularization parameter $\lambda$ and the type of penalty indicated by $\gamma$.

### 3.3 Adaptive $\ell_1$-norm penalization

A two-stage adaptive $\ell_1$-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 $\ell_1$-norm penalized estimator for a mixture of Gaussian regressions is defined as follows. Consider an initial estimate $\theta^{ini}$, for example from the estimator in (3.8). The adaptive criterion to be minimized involves a re-weighted $\ell_1$-norm penalty term:

$$\begin{aligned} -n^{-1}\ell^{(\gamma)}_{adapt}(\theta) &= -n^{-1}\sum_{i=1}^{n}\log\left(\sum_{r=1}^{k}\pi_r\frac{\rho_r}{\sqrt{2\pi}}\exp(-\frac{1}{2}(\rho_r Y_i - X_i^T\phi_r)^2)\right) \\ &+ \lambda\sum_{r=1}^{k}\pi_r^{\gamma}\sum_{j=1}^{p}w_{r,j}|\phi_{r,j}|, \\ w_{r,j} &= \frac{1}{|\phi^{ini}_{r,j}|}, \quad \theta = (\rho_1, \ldots, \rho_k, \phi_1, \ldots, \phi_k, \pi_1, \ldots, \pi_{k-1}), \end{aligned} \tag{3.12}$$

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

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

where $\Theta$ is as in (3.10).

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 the FMR model in Section 4 and we report some empirical results in Section 7.1.

### 3.4 Selection of the tuning parameters

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

One possibility is to use a modified BIC criterion which minimizes

$$\text{BIC} \quad = \quad -2\ell(\hat{\theta}^{(\gamma)}_{\lambda,k}) + log(n)\text{df}, \tag{3.13}$$

over a grid of candidate values for $k$, $\lambda$ and maybe also $\gamma$. Here, $\hat{\theta}^{(\gamma)}_{\lambda,k}$ denotes the estimator in (3.8) using the parameters $\lambda, k, \gamma$ in (3.9), and $-\ell(\cdot)$ is the negative log-likelihood. Furthermore, $\text{df} = kp + k + (k-1) - \sum_{j=1\ldots p, r=1\ldots k} 1_{\{\hat{\phi}_{r,j}=0\}}$ are the degrees of freedom (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 $\lambda$, we consider $0 \leq \lambda_1 < \ldots < \lambda_M \leq \lambda_{max}$, where $\lambda_{max}$ is given by

$$\lambda_{max} = \max_{j=1,\ldots,p} \left| \frac{\langle Y, X^{(j)} \rangle}{\sqrt{n}||Y||} \right|. \tag{3.14}$$

At $\lambda_{max}$, all coefficients $\hat{\phi}_j$, $(j = 1, \ldots, p)$ of the one-component model are exactly zero. Equation (3.14) easily follows from Proposition 3.1.

For the adaptive $\ell_1$-norm penalized estimator minimizing the criterion in (3.12) we proceed analogously. As initial estimator in the adaptive criterion, we propose to use the estimate in (3.8) 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 (3.9). We assume here that the number of covariates $p$ and the number of mixture components $k$ is 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 $\theta_0$ the true parameter.

**Theorem 4.1.** *(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 (3.9) ($\gamma \in \{0, 1/2, 1\}$) such that*

$$\sqrt{n}\left(\hat{\theta}^{(\gamma)}_{\lambda} - \theta_0\right) \to^{\mathbb{P}} 0 \ (n \to \infty).$$

A proof is given in Appendix A. Theorem 4.1 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 $\theta_0$) can choose a consistent

sequence of roots (van der Vaart, 2007). 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). Denote by $S$ the population analogue of (3.11), 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 4.2.** *(Asymptotic oracle result) Consider model (2.2) with fixed design and fixed $p$ and $k$. If $\lambda = o(n^{-1/2})$, $n\lambda \to \infty$ and if $\theta^{ini}$ 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 (3.12) ($\gamma \in \{0, 1/2, 1\}$) which satisfies:*

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

A proof is given in Appendix A. As in Theorem 4.1, the assertion of the Theorem is only making a statement about *some* local optimum. Furthermore, variable selection consistency and the oracle property hinge on the implicit assumption that the regression parameters are either zero or take a fixed value different from zero which excludes the cases with small non-zero values in e.g. the $n^{-1/2}$-domain.

# 5 Theory for high-dimensional setting

We will present here some theory, entirely different from Theorems 4.2 and 4.1, which reflects some consistency and optimality behaviour of the $\ell_1$-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 $\ell_1$-penalized smooth likelihood problems which are generally non-convex: $\ell_1$-penalized likelihood estimation in FMR models is then a special case.

## 5.1 The setting and notation

Let $\{f_\theta;\ \theta \in \Theta\}$ be a collection of densities with respect to some $\sigma$-finite measure $\mu$, on a measurable space $\mathcal{Y}$ (i.e. the range for the response variable). The parameter space $\Theta$ is assumed to be a bounded subset of some finite-dimensional space, say

$$\Theta \subset \{\theta \in \mathbb{R}^d;\ \|\theta\|_\infty \le K\},$$

where we have equipped (quite arbitrarily) the space $\mathbb{R}^d$ with the sup-norm $\|\theta\|_\infty = \max_{1\le j\le d} |\theta_j|$. 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 $\ell_q$-norm $\|\theta\|_q = (\sum_{j=1}^d |\theta_j|_q)^{1/q}$ $(q \ge 1)$.

We observe a co-variable $X$ in some space $\mathcal{X}$ and a response variable $Y$ in $\mathcal{Y}$. The true conditional density of $Y$ given $X = x$ is assumed to be equal to

$$f_{\theta_0}(\cdot|x) = f_{\theta_0(x)},$$

where

$$\theta_0(x) \in \Theta,\ \forall\ x \in \mathcal{X}.$$

That is, we assume that the true conditional density of $Y$ given $X = x$ is depending on $x$ only through some parameter function $\theta_0(x)$. Of course, the introduced notation also applies to fixed co-variables.

The parameter $\{\theta_0(x);\ x \in \mathcal{X}\}$ is assumed to have a nonparametric part of interest $\{g_0(x);\ x \in \mathcal{X}\}$ and a low-dimensional nuisance part $\eta_0$, i.e.,

$$\theta_0(\cdot)^T = (g_0(\cdot)^T, \eta_0^T),$$

with

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

In case of FMR models, $g(x)^T = (\phi_1^T x, \phi_2^T x, \ldots, \phi_k^T x)$ and $\eta$ involves the parameters $\rho_1, \ldots, \rho_k, \pi_1, \ldots, \pi_{k-1}$. More details are given in Section 5.6.

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

$$\mathcal{E}(\theta|\theta_0) = -\int \log\left[\frac{f_\theta}{f_{\theta_0}}\right] f_{\theta_0} d\mu$$

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

$$\bar{\mathcal{E}}(\theta|\theta_0) = \frac{1}{n}\sum_{i=1}^n \mathcal{E}\left(\theta(x_i)\middle|\theta_0(x_i)\right).$$

## 5.2 A consistency result for FMR models

Denote by $\theta_0 = (\phi_0, \eta_0)$ the true parameter vector in an FMR model, where $\phi_0 = (\phi_{0,1}, \ldots, \phi_{0,k})^T$ with $\phi_{0,r} = \beta_{0,r}/\sigma_r$ $(r = 1, \ldots, k)$ and $\eta_0 = \log(\rho_1), \ldots, \log(\rho_k), \log(\pi_1), \ldots,$ $\ldots, \log(\pi_{k-1})$. Consider the estimator

$$\begin{aligned}
\hat{\theta}_\lambda &= \arg\min_{\vartheta \in \tilde{\Theta}} -n^{-1}\sum_{i=1}^n \log\left(\sum_{r=1}^k \pi_r \frac{\rho_r}{\sqrt{2\pi}} \exp(-\frac{1}{2}(\rho_r Y_i - X_i^T\phi_r)^2)\right) + \lambda\sum_{r=1}^k ||\phi_r||_1, \\
\tilde{\Theta} &= \{\vartheta^T = (\phi_1^T, \ldots, \phi_k^T, \eta^T);\ \sup_{x\in\mathcal{X}} \|\phi^T x\|_\infty \le K,\ \|\eta\|_\infty \le K\}. \qquad (5.15)
\end{aligned}$$

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 oft the penalized negative log-likelihood which is arguably difficult to compute.

**Theorem 5.1.** *(Consistency) Consider model (2.2) with fixed design and fixed $k$ and assume that Condition 5 below 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 $\lambda \asymp C\sqrt{\log^4(n)/n})$ for some $C > 0$ sufficiently large, then any (global) minimizer $\hat{\theta}_\lambda$ as in (5.15) satisfies*

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

A proof is given in Appendix B.

### 5.3 The margin

We develop in the following Sections 5.3-5.5 a non-asymptotic oracle inequality result for the general case of penalized smooth likelihood estimation.

Denote by

$$l_\theta = \log f_\theta$$

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

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

and the Fisher information

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

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

We will assume boundedness of third derivatives.

**Condition 1** *It holds that*

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

*where*

$$\sup_{x\in\mathcal{X}} \int G_3(y) f_{\theta_0}(y|x) d\mu(y) \leq C_3 < \infty.$$

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

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

$$\Lambda_{\min} = \inf_{x\in\mathcal{X}} \Lambda_{\min}(I(g_0(x), \eta_0)) > 0.$$

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

$$\mathcal{E}(\theta|\theta_0) = -\int \log\left[\frac{f_\theta}{f_{\theta_0}}\right] f_{\theta_0} d\mu$$

is the Kullback Leibler information. We will need the following identifiability condition.

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

$$\inf_{x\in\mathcal{X}} \inf_{\substack{\theta\in\Theta \\ \|\theta-\theta_0(x)\|_2>\varepsilon}} \mathcal{E}(\theta|\theta_0(x)) \geq \alpha_\varepsilon.$$

**Lemma 5.1.** *Assume Conditions 1, 2, and 3. Then*

$$\inf_{x\in\mathcal{X}} \frac{\mathcal{E}(\theta|\theta_0(x))}{\|\theta-\theta_0(x)\|_2^2} \geq \frac{1}{c_0^2},$$

*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.

### 5.4 The empirical process

We now specialize to the case where

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

where (with some abuse of notation)

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

We will assume that

$$\sup_{x \in \mathcal{X}} \|\phi^T x\|_\infty = \sup_{x \in \mathcal{X}} \max_{1 \le r \le k} |\phi_r^T x| \le K.$$

We write

$$\theta_\phi(x)^T = (g_\phi(x)^T, \eta^T).$$

Our parameter space is now

$$\tilde{\Theta} \subset \{\vartheta^T = (\phi_1^T, \ldots, \phi_k^T, \eta^T);\ \sup_{x \in \mathcal{X}} \|\phi^T x\|_\infty \le K,\ \|\eta\|_\infty \le K\}. \tag{5.16}$$

Note that $\tilde{\Theta}$ is in principle $(pk + m)$-dimensional. The true parameter $\vartheta_0$ is assumed to be an element of $\tilde{\Theta}$.

Let us define

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

and the empirical process

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

**Condition 4** *For the score function $s_\theta(\cdot)$ we have:*

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

*where $G_1(\cdot)$ satisifies some moment conditions to be specified in Lemma 5.2.*

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 $\Sigma_n$.

**Condition 5** *For a constant $\Lambda_{\max} < \infty$, it holds that $\Lambda_{\max}(\Sigma_n) \le \Lambda_{\max}$.*

Condition 5 is not really a condition, but 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

$$\lambda_0 = M_n \sqrt{\frac{\log^3 n}{n}}. \tag{5.17}$$

Let $\mathbb{P}_{\mathbf{x}}$ denote the conditional probability given $(X_1, \ldots, X_n) = (x_1, \ldots, x_n) = \mathbf{x}$.

**Lemma 5.2.** *Assume Conditions 4 and 5. We have for constants $c_1$ and $c_2$ depending on $\Lambda_{\max}$, $k$, and $m$, and for all $T \geq 1$,*

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

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

$$1 - c_1 \exp\left[-\frac{T^2 n\lambda_0^2}{c_2^2}\right] - \mathbb{P}_{\mathbf{x}}\left(\frac{1}{n}\sum_{i=1}^n G_1(Y_i) \mathrm{l}\{G_1(Y_i) > M_n\} > T\lambda_0^2/(dK)\right).$$

*Regarding the constants $\lambda_0$ and $K$, see (5.17) and (5.16), respectively.*

The result follows along the lines of Lemma 4 in Tsybakov and van de Geer (2005) by developing a suitable entropy bound.

### 5.5 Oracle inequality for the Lasso for non-convex loss functions

We employ the Lasso-type estimator

$$\hat{\vartheta}^T = (\hat{\phi}^T, \hat{\eta}^T) = \underset{\vartheta^T=(\phi^T,\eta^T)\in\tilde{\Theta}}{\arg\min} \left\{ -\frac{1}{n}\sum_{i=1}^n L_\vartheta(x_i, Y_i) + \lambda \sum_{r=1}^k \|\phi_r\|_1 \right\}. \tag{5.18}$$

We omit in the sequel the dependence of $\hat{\vartheta}$ on $\lambda$. 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 $\vartheta$. We then write $\|\phi\|_1 = \sum_{r=1}^k \|\phi_r\|_1$. We let

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

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

$$\theta_0(x)^T = (g_{\phi_0}(x)^T, \eta_0^T).$$

We also define the set

$$\mathcal{T} = \left\{ \sup_{\vartheta^T=(\phi^T,\eta^T)\in\tilde{\Theta}} \frac{\Big|V_n(\vartheta) - V_n(\vartheta_0)\Big|}{(\|\phi-\phi_0\|_1 + \|\eta-\eta_0\|_2) \vee \lambda_0} \leq T\lambda_0 \right\}. \tag{5.19}$$

Let

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

be the active set, i.e. the set of non-zero coefficients, and

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

**Condition 6** *(Compatibility 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 $\theta(\cdot)^T = (g(\cdot)^T, \eta^T)$, we use the notation

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

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

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

**Theorem 5.2.** *(Oracle result for fixed design). Assume Conditions 1-3 and 6, and that $\lambda \geq 2T\lambda_0$ for the estimator in (5.18). Then on $\mathcal{T}$, for the average excess risk (average Kullback-Leibler loss),*

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

A proof is given in Appendix B. The probability of the set $\mathcal{T}$ is large, assuming Conditions 4 and 5. In the case of FMR models, this is shown in detail by Lemma 5.3 below.

The oracle inequality of Theorem 5.2 has the following well-known interpretation. First, we obtain

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

that is, the average Kullback-Leibler risk is of the order $O(s\lambda_0^2 M_n^2) = O(s\log(n)^4/n)$ (see (5.17) and the assumption on $M_n$ in Lemma 5.3 below) which is up to the factor $\log(n)^4$ the optimal convergence rate if one would know the $s$ non-zero coefficients. 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 $\ell_1$-norm converges to zero at rate $O(s\lambda_0)$).

### 5.6 FMR models

In the finite mixture of regressions model from (2.2) with $k$ components, the parameter is $\vartheta^T = (g^T, \eta^T) = (g^T, \log\rho_1, \ldots, \log\rho_k, \log\pi_1, \ldots, \log\pi_{k-1})$, where the $\rho_r = \sigma_r^{-1}$ are the inverse standard deviations in mixture component $r$ and the $\pi_r$ are the mixture coefficients. We let $\tilde{\Theta} \subset \{\|g\|_\infty \leq K, \|\log\rho\|_\infty \leq K, -K \leq \log\pi_1 \leq 0, \ldots, -K \leq \log\pi_{k-1} \leq 0, \ \sum_{r=1}^{k-1} \pi_r < 1\}$. The log-likelihood is

$$l_\theta(y) = \log\left[\sum_{r=1}^{k} \pi_r \rho_r \varphi\Big(\rho_r y - g_r\Big)\right].$$

In this case $d = 3k - 1$.

**Proposition 5.1.** *For fixed design FMR models as in (2.2) with $\tilde{\Theta}$ as above, Conditions 1,2 and 3 are met, for appropriate $C_3$, $\Lambda_{\min}$ and $\{\alpha_\varepsilon\}$, depending on $k$ and $K$. Also Condition 4 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 5.2 and the following result.

**Lemma 5.3.** *For fixed design FMR models as in (2.2) with $\tilde{\Theta}$ as above: 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:*

$$\mathbb{P}_{\mathbf{x}}\left(\frac{1}{n}\sum_{i=1}^{n} G_1(Y_i)\mathrm{l}\{G_1(Y_i) > M_n/(dK)\} > c_5\frac{\log n}{n}\right) \leq \frac{1}{n}.$$

A proof is given in Appendix B.

Hence, the oracle result in Theorem 5.2 for our $\ell_1$-norm penalized estimator in the FMR model holds on a set $\mathcal{T}$, and this set $\mathcal{T}$ has large probability due to Lemma 5.2 and Lemma 5.3.

# 6 Numerical optimization

We will present a generalized EM (GEM) algorithm for optimizing the criterion in (3.9) in Section 6.1. In Section 6.2 we will discuss numerical convergence properties of the algorithm. 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:

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

where $\Delta_{i,r}$ are unobserved, imaginary indicator variables showing the component-membership of the $i$th observation in the FMR model. The expected complete (scaled) negative log-likelihood is then:

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

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

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

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

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

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

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

a) *Improvement with respect to $\pi$:*

fix $\phi$ at the present value $\phi^{(m)}$ and improve

$$-n^{-1} \sum_{i=1}^n \sum_{r=1}^k \hat{\gamma}_{i,r} \log(\pi_r) + \lambda \sum_{r=1}^k \pi_r^{\gamma} ||\phi_r^{(m)}||_1 \tag{6.20}$$

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 $\pi$ 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 (6.20) 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) *Minimization with respect to $\phi$ and $\rho$:*

A simple calculation shows, that the M-Step decouples for each component into $k$ distinct optimization problems of the form

$$-\log(\rho_r) + \frac{1}{2n_r} ||\rho_r \tilde{Y} - \tilde{X}\phi_r||^2 + \frac{n\lambda}{n_r} \left(\pi_r^{(m+1)}\right)^{\gamma} ||\phi_r||_1, \qquad r = 1, \ldots, k \tag{6.21}$$

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 (6.21) has the same form as (3.7): in particular, it involves convex optimization. Closed-form coordinate updates can easily be computed for each component $r$ $(r = 1, \ldots, k)$ using Proposition 3.1:

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

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

where $S_j$ is defined as

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

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

Because we only improve $Q_{pen}(\theta|\theta^{(m)})$, 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). Its numerical properties are discussed in Section 6.2.

**Remark 6.1.** *For the convex penalty function with $\gamma = 0$, a minimization with respect to $\pi$ 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 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 proofed 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,

$$-n^{-1}\ell_{pen,\lambda}^{(\gamma)}(\theta^{(m+1)}) \le -n^{-1}\ell_{pen,\lambda}^{(\gamma)}(\theta^{(m)}). \tag{6.22}$$

Since $-n^{-1}\ell_{pen,\lambda}^{(0)}(\theta)$ is bounded from below (Proposition 3.2), the following result holds.

**Proposition 6.1.** *For the BCD-GEM algorithm, $-n^{-1}\ell_{pen,\lambda}^{(0)}(\theta^{(m)})$ decreases monotonically to some value $\bar{\ell} > -\infty$.*

In Remark 6.1 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 6.1.** *Consider the BCD-GEM algorithm for the criterion function in (3.9) 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 (3.9).*

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

# 7 Simulations and real data example

## 7.1 Simulations

We consider four different simulation setups. Simulation scenario 1 compares the performance of the unpenalized MLE (Flexmix, according to the name of the `R`-package) 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; 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 $\pi_r, \beta_r, \sigma_r$ 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 $\pi_r$), 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. $\lambda$) 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 co-variable $X^{(j)}$ 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 $\sigma_1^2$, $\sigma_2^2$ and hence have different signal to noise ratios. Model M5 has a non-diagonal covariance structure. Furthermore in model M5 the variances $\sigma_1^2$, $\sigma_2^2$ are tuned to achieve the same signal to noise ratio as in model M1.

We compare the performance 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 box-plots 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 the box-plots that the MLE performs very badly when we add 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.

### 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 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:

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

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.

| | M1 | M2 | M3 | M4 | M5 |
|---|---|---|---|---|---|
| n | 100 | 100 | 100 | 150 | 100 |
| $\beta_1$ | (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) |
| $\beta_2$ | (-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) |
| $\beta_3$ | - | - | - | (0,0,0,0,-3,2) | - |
| $\sigma$ | 0.5, 0.5 | 1, 1 | 2, 2 | 0.5, 0.5, 0.5 | 0.95, 0.95 |
| $\pi$ | 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.

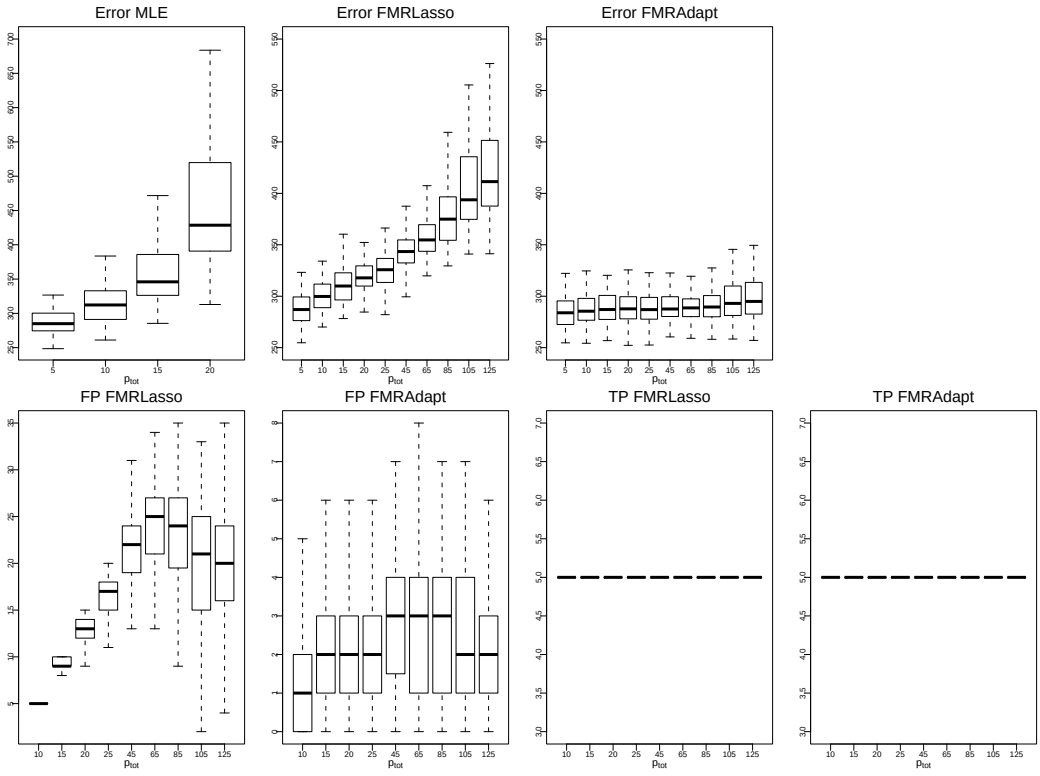


Figure 1: Simulation 1, Model M1. Top: predictive log-likelihood loss (*Error*) for MLE, FMRLasso, FMRAdapt. Bottom: False Positives (*FP*) and True Positives (*TP*) for FMRLasso and FMRAdapt.

| $p_{act}$ | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|
| n | 50 | 100 | 150 | 200 | 250 | 300 | 350 |
| $p_{tot}$ | 10 | 20 | 40 | 80 | 160 | 320 | 640 |
| $\beta_1$ | | | | (3,3,3,0,0,...) | | | |
| $\beta_2$ | | | | (-1,-1,-1,0,0,...) | | | |
| $\sigma$ | | | | 0.5, 0.5 | | | |
| $\pi$ | | | | 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 $\lambda_{opt}$ 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 $\lambda$. 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 $\lambda_{opt}$ using 10-fold cross-validation.

(2) Assume the number of components is given ($k = 2$). Choose $\lambda_{opt}$ by minimizing the

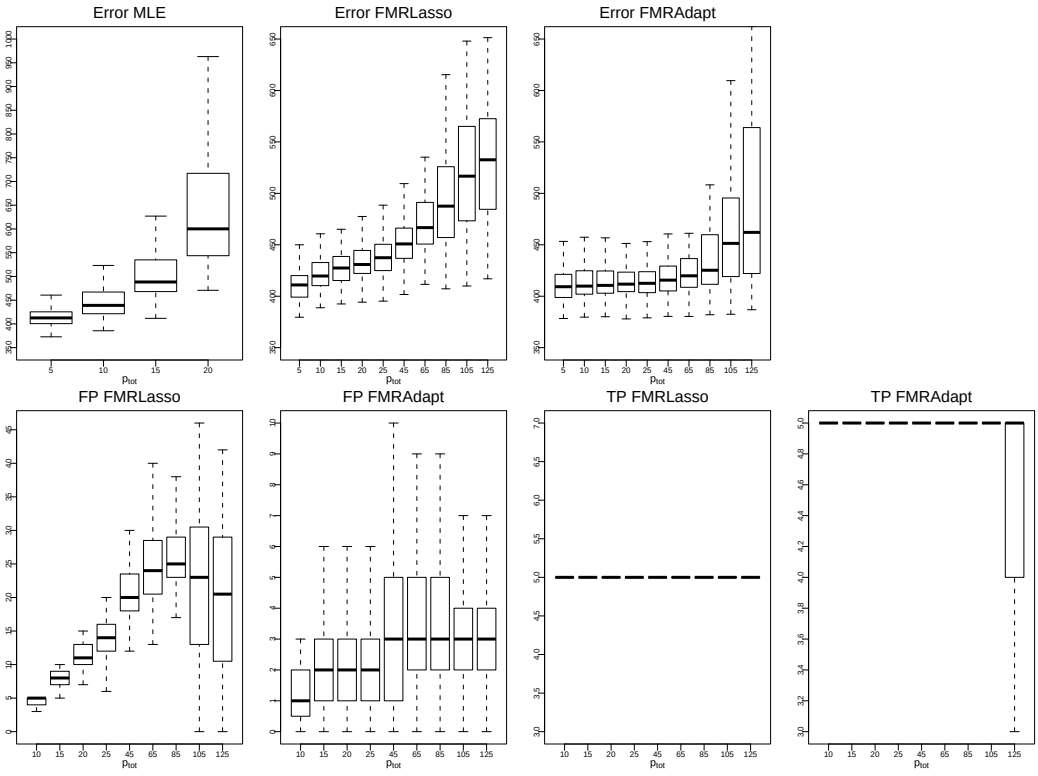


Figure 2: Simulation 1, Model M2. Same notation as in Figure 1.

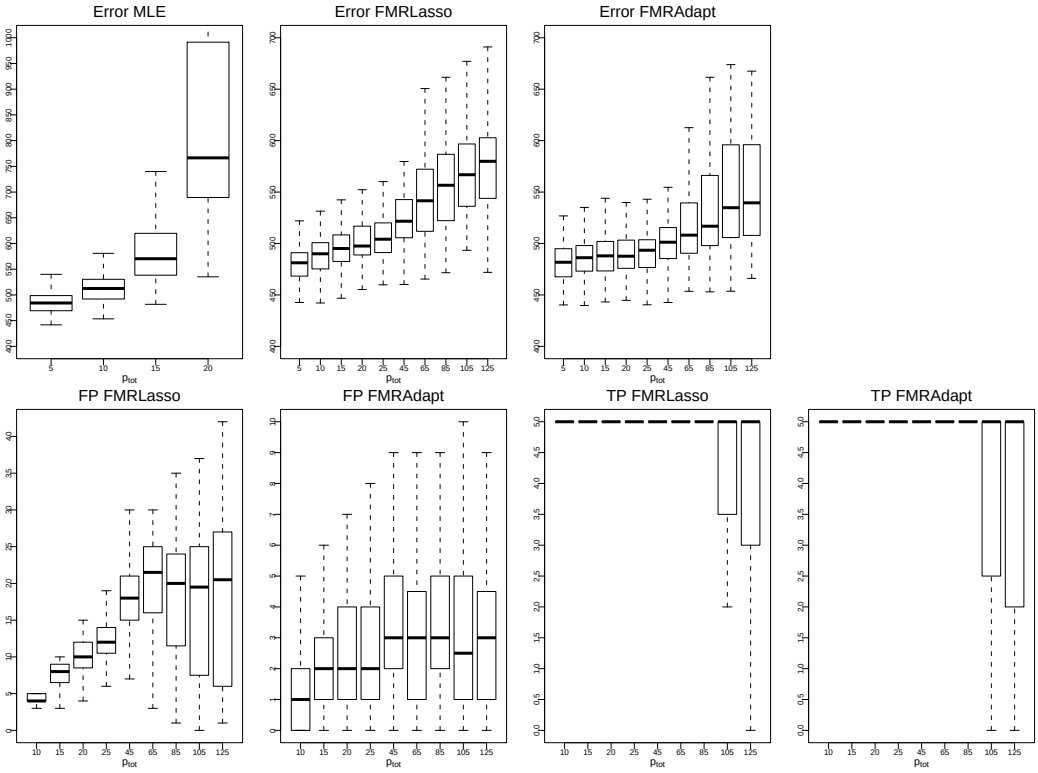


Figure 3: Simulation 1, Model M3. Same notation as in Figure 1.

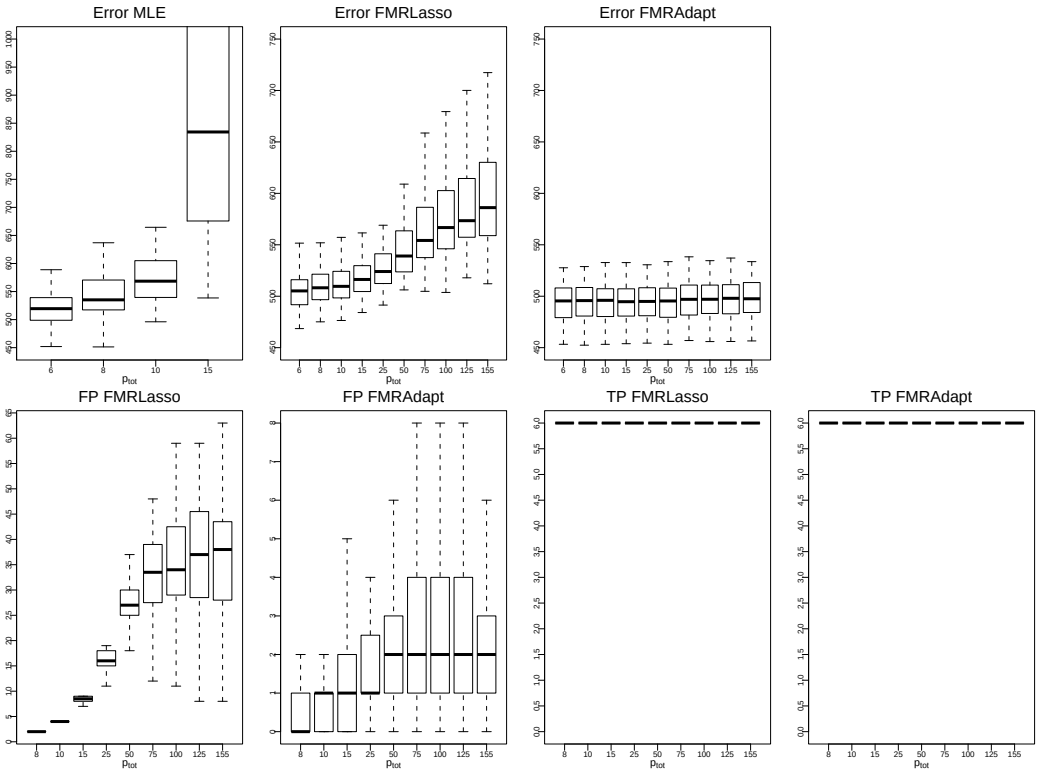


Figure 4: Simulation 1, Model M4. Same notation as in Figure 1.

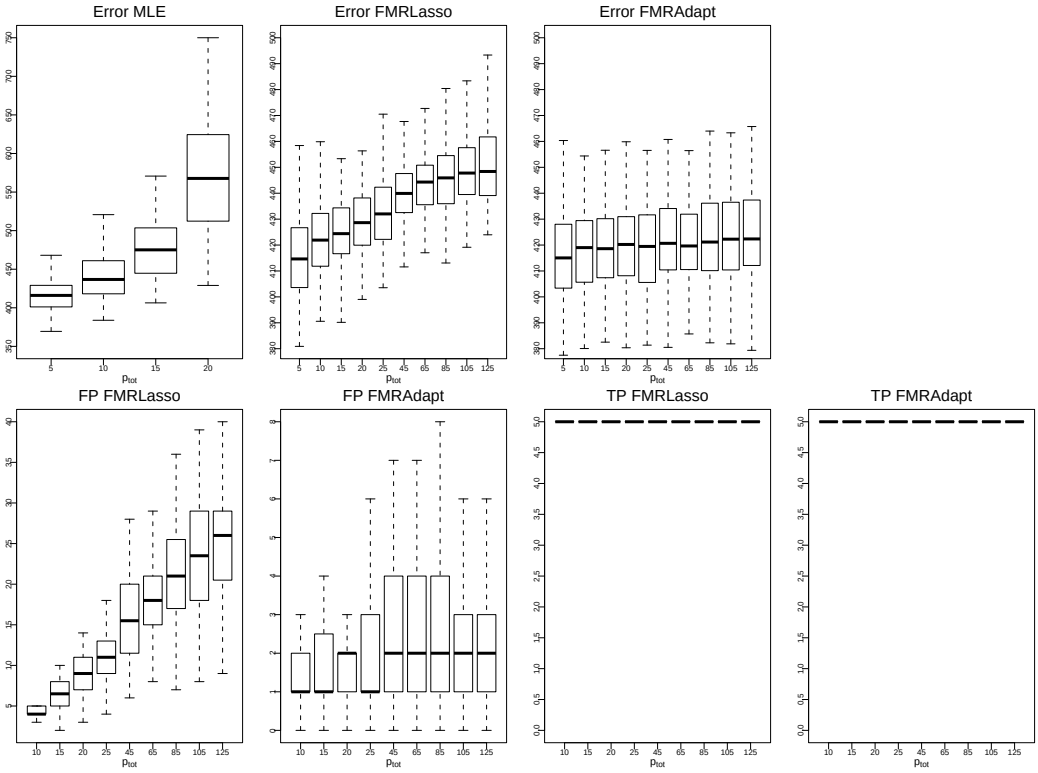


Figure 5: Simulation 1, Model M5. Same notation as in Figure 1.

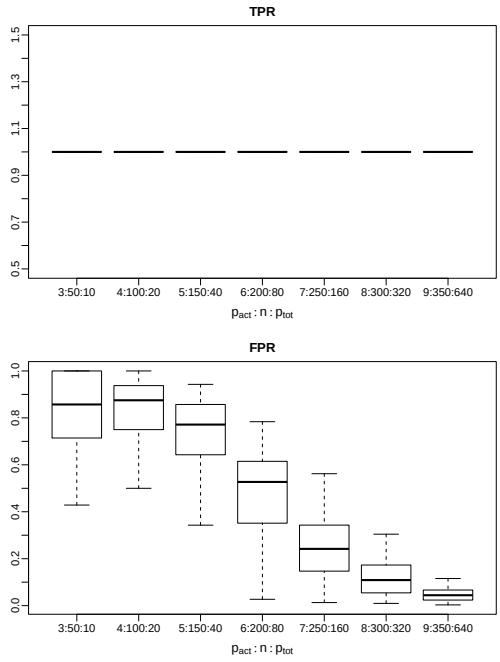


Figure 6: Simulation 2 compares the performance of the FMRLasso for a series of models which gets "sparser" as the sample size grows. Top: True Positive Rate ($TPR$). Bottom: False Positive Rate ($FPR$) over 100 simulation runs.

BIC criterion (3.13).

(3) Choose the number of components $k \in \{1, 2, 3\}$ and $\lambda_{opt}$ by minimizing the BIC criterion (3.13).

The results of this simulation are presented in Figure 7, where box-plots of the log-likelihood loss ($Error$) are shown. For the model with $p_{tot} = 25$ all three strategies perform equally well. The BIC criterion in strategy (3) chooses always $k = 2$. For the model with $p_{tot} = 50$ the BIC strategies (2) and (3) are superior over cross-validation. Strategy (3) chooses in ninety-six simulation runs $k = 2$ and in four runs $k = 3$. With $p_{tot} = 75$ again the BIC strategies (2) and (3) perform better than cross-validation. Strategy (3) chooses once $k = 1$, ninety-five times $k = 2$, and four 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 (3.9). 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 box-plots 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.

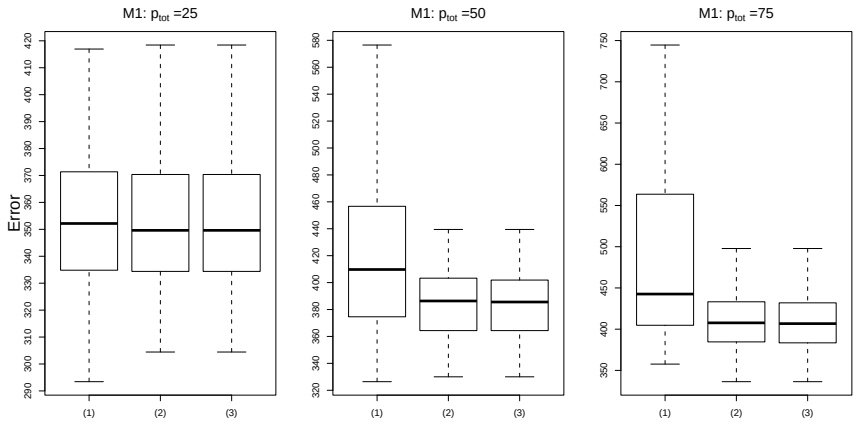


Figure 7: Simulation 3 compares different strategies for choosing the tuning parameters $k$ and $\lambda$. The box-plots show the predictive log-likelihood loss (*Error*) of the FMRLasso, tuned by strategies (1), (2) and (3), for model M1 with $p_{tot} = 25, 50, 75$.

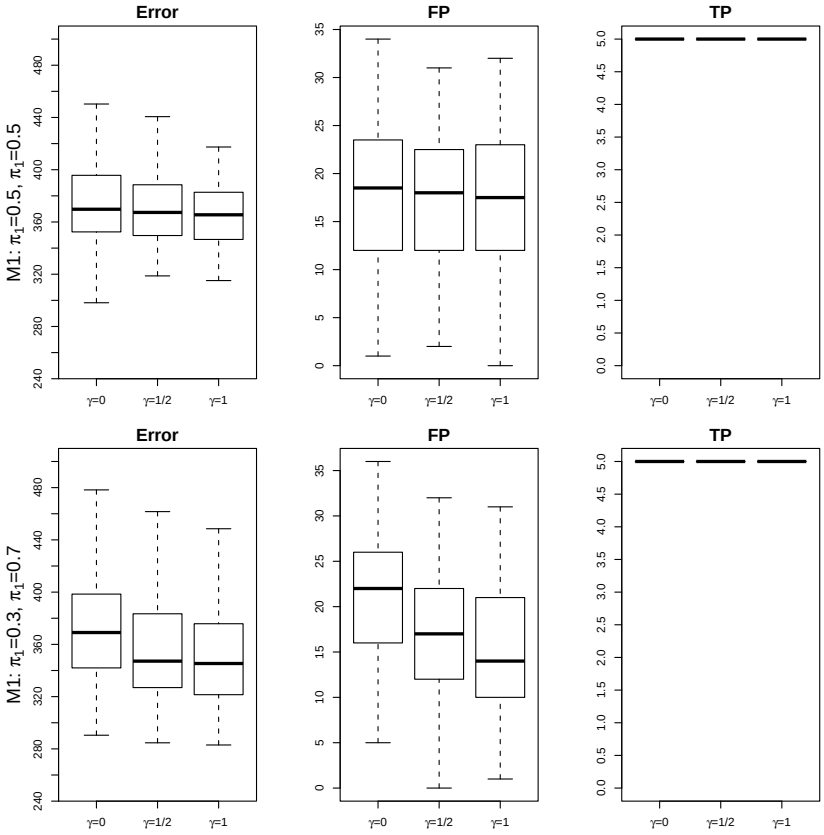


Figure 8: Simulation 4 compares the FMRLasso for different values $\gamma = 0, 1/2, 1$. The upper row of the panels shows the box-plots of the log-likelihood loss (*Error*), the False Positives (*FP*) and the True Positives (*TP*) for model M1 with $p_{tot} = 50$ and $\pi_1 = \pi_2 = 0.5$. The lower row of the panels shows the same box-plots 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. There are measurements of $n = 146$ genetically engineered mutants of *Bacillus Subtilis*. The population seems to be rather heterogenous 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 $k = 1, \ldots, 6$ 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 $\lambda_{opt}$ 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 46 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. The important variables do not show opposite signs 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 22 selected genes from the one-component model are also selected in the the three-component model.

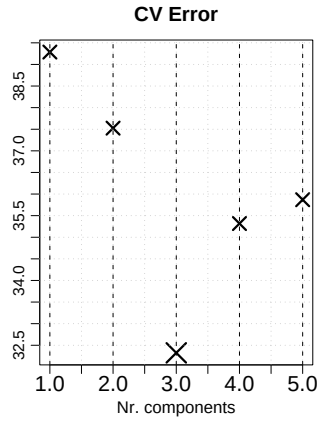


Figure 9: Cross-validated negative log-likelihood loss (*CV Error*) for the FMRLasso estimator when varying over different numbers of components.

# 8 Discussion

We have presented an $\ell_1$-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

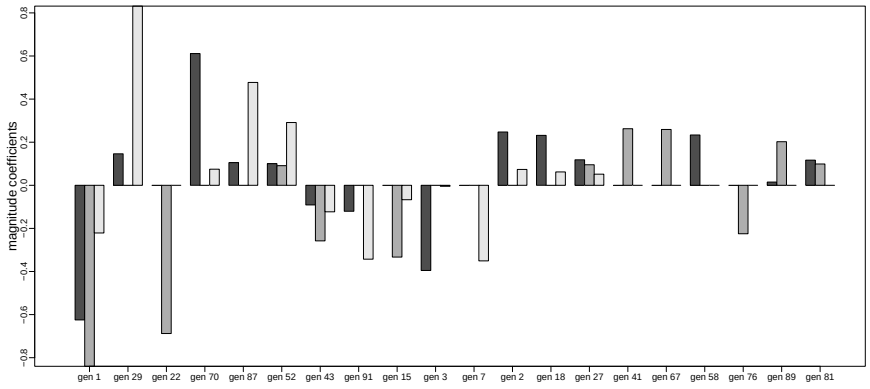


Figure 10: Coefficients of the twenty most important genes, ordered according to $\sum_{r=1}^{3} |\beta_{r,j}|$, for the prediction optimal model with three components.

challenging due to the non-convex behavior of the negative log-likelihood. Regarding the computation, a simple reparameterization is beneficial and the $\ell_1$-penalty term makes the optimization problem numerically 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

*Proof of Theorem 4.1.* 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). □

*Proof of Theorem 4.2.* Without loss of generality consider a two class mixture with $k = 2$.

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 $\theta_0$ 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 of order $O_P(\frac{1}{\sqrt{n}})$ (central limit theorem). Term (2) is of order $O_P(1)$ (law of large numbers). Term (3) is of order $O_P(1)$ (law of large numbers and regularity

condition on 3rd derivatives). 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

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

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 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

$$\left(-I_S(\theta_0) + o_P(1)\right)\sqrt{n}(\hat{\theta}_S - \theta_{0,S}) - \sqrt{n}\lambda O_P(1) = -\frac{1}{\sqrt{n}}\ell'|_{\theta_{0,S}} \tag{A.23}$$

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.23). □

# B Proofs for Section 5

*Proof of Theorem 5.1.* On $\mathcal{T}$, defined in (5.19) with $\lambda_0 = c_4\sqrt{\log^4(n)/n}$ ($c_4$ as in Lemma 5.3; i.e. $M_n = c_4\sqrt{\log(n)}$ in (5.17)), we have the basic inequality

$$\bar{\mathcal{E}}(\hat{\theta}|\theta_0) + \lambda\|\hat{\phi}\|_1 \leq T\lambda_0\Big[(\|\hat{\phi} - \phi_0\|_1 + \|\hat{\eta} - \eta_0\|_2) \vee \lambda_0\Big] + \lambda\|\phi_0\|_1 + \bar{\mathcal{E}}(\theta_0|\theta_0).$$

Note that $\|\hat{\eta} - \eta_0\|_2 \leq 2K$ and $\bar{\mathcal{E}}(\theta_0|\theta_0) = 0$. Hence, for $n$ sufficiently large,

$$\begin{aligned}\bar{\mathcal{E}}(\hat{\theta}|\theta_0) + \lambda\|\hat{\phi}\|_1 &\leq T\lambda_0(\|\hat{\phi} - \phi_0\|_1 + 2K) + \lambda\|\phi_0\|_1 + \bar{\mathcal{E}}(\theta_0|\theta_0) \\ &\leq T\lambda_0(\|\hat{\phi}\|_1 + \|\phi_0\|_1 + 2K) + \lambda\|\phi_0\|_1 + \bar{\mathcal{E}}(\theta_0|\theta_0),\end{aligned}$$

and therefore also

$$\bar{\mathcal{E}}(\hat{\theta}|\theta_0) + (\lambda - T\lambda_0)\|\hat{\phi}\|_1 \leq T\lambda_0 2K + (\lambda + T\lambda_0)\|\phi_0\|_1 + \bar{\mathcal{E}}(\theta_0|\theta_0).$$

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 $\bar{\mathcal{E}}(\hat{\theta}|\theta_0) \to \bar{\mathcal{E}}(\theta_0|\theta_0) = 0$ ($n \to \infty$). Finally, the set $\mathcal{T}$ has large probability, as shown by Lemma 5.2 and using Proposition 5.1 and Lemma 5.3 for FMR models. □

*Proof of Lemma 5.1.* It is clear that

$$\mathcal{E}(\theta|\theta_0) = (\theta - \theta_0)^T I(\theta_0)(\theta - \theta_0)/2 + r_\theta,$$

where

$$\begin{aligned}|r_\theta| &\leq \frac{\|\theta - \theta_0\|_1^3}{6}\int \sup_{\theta\in\Theta}\max_{j_1,j_2,j_3}\left|\frac{\partial^3 l_\theta}{\partial\theta_{j_1}\partial\theta_{j_2}\partial\theta_{j_3}}\right| f_{\theta_0}d\mu \\ &\leq \frac{d^{3/2}C_3}{6}\|\theta - \theta_0\|_2^3.\end{aligned}$$

Hence

$$\mathcal{E}(\theta|\theta_0(x)) \geq \|\theta - \theta_0(x)\|_2^2\Lambda_{\min}^2/2 - d^{3/2}C_3\|\theta - \theta_0(x)\|_2^3/6.$$

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.* 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 $h(z) \geq \alpha_{\varepsilon_0} = K_0^2 \alpha_{\varepsilon_0}/K_0^2 \geq |z|^2 \alpha_{\varepsilon_0}/K_0^2$.

□

*Proof of Theorem 5.2.* On $\mathcal{T}$

$$\bar{\mathcal{E}}(\hat{\theta}|\theta_0) + \lambda\|\hat{\phi}\|_1 \leq T\lambda_0 \Big[ (\|\hat{\phi} - \phi_0\|_1 + \|\hat{\eta} - \eta_0\|_2) \vee \lambda_0 \Big] + \lambda\|\phi_0\|_1 + \bar{\mathcal{E}}(\theta_0|\theta_0).$$

By Lemma 5.1,

$$\bar{\mathcal{E}}(\hat{\theta}|\theta_0) \geq \|\hat{\theta} - \theta_0\|_{Q_n}^2/c_0^2,$$

and $\bar{\mathcal{E}}(\theta_0|\theta_0) = 0$.

**Case 1** Suppose that

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

Then we find

$$\begin{aligned} \bar{\mathcal{E}}(\hat{\theta}|\theta_0) &\leq T\lambda_0^2 + \lambda\|\hat{\phi} - \phi_0\|_1 + \bar{\mathcal{E}}(\theta_0|\theta_0) \\ &\leq (\lambda + T\lambda_0)\lambda_0. \end{aligned}$$

**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

$$\begin{aligned} & \bar{\mathcal{E}}(\hat{\theta}|\theta_0) + (\lambda - T\lambda_0)\|\hat{\phi}_{S^c}\|_1 \leq 2T\lambda_0\|\hat{\eta} - \eta_0\|_2 \\ \leq \;& 4T^2\lambda_0^2 c_0^2 + \|\hat{\eta} - \eta_0\|_2^2/(2c_0^2) \\ \leq \;& 4T^2\lambda_0^2 c_0^2 + \bar{\mathcal{E}}(\hat{\theta}|\theta_0)/2. \end{aligned}$$

So then

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

**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

$$\bar{\mathcal{E}}(\hat{\theta}|\theta_0) + (\lambda - T\lambda_0)\|\hat{\phi}_{S^c}\|_1 \leq 2(\lambda + T\lambda_0)\|\hat{\phi}_S - \phi_0\|_1.$$

So then

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

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

$$\begin{aligned} \bar{\mathcal{E}}(\hat{\theta}|\theta_0) + (\lambda - T\lambda_0)\|\hat{\phi}_{S^c}\|_1 &\leq 2(\lambda + T\lambda_0)\sqrt{s}\|\hat{\phi}_S - \phi_0\|_2. \\ &\leq 2(\lambda + T\lambda_0)\sqrt{s}\kappa\|\hat{g} - g_0\|_{Q_n} \end{aligned}$$

$$\leq 4(\lambda + T\lambda_0)^2 c_0^2 \kappa^2 s + \bar{\mathcal{E}}(\hat{\theta}|\theta_0)/2.$$

So we arrive at

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

□

*Proof of Lemma 5.3.* 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 $M = 2\sqrt{\log n}$,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n |Z_i|1\{|Z_i| > M\} > \frac{4\log n}{n}\right)$$

$$\leq \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n |Z_i|1\{|Z_i| > M\} - E|Z|1\{|Z| > M\} > \frac{2\log n}{n}\right)$$

$$\leq \frac{nE|Z|^2 1\{|Z| > M\}}{4(\log n)^2} \leq \frac{2}{n}.$$

The result follows from this, as

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

and $Y$ has a normal mixture distribution.

□

# C Proofs for Sections 3 and 6

*Proof of Proposition 3.2.* We restrict ourself to a two class mixture with k = 2. Consider the function $u(\xi)$ defined as

$$\begin{aligned} & u(\xi) = \exp(\ell_{pen}^{(0)}(\xi)) \\ \propto \; & \prod_{i=1}^n \left\{ \left( \pi \frac{1}{\sigma_1} e^{\frac{-(Y_i - X_i'\beta_1)^2}{2\sigma_1^2}} + (1-\pi)\frac{1}{\sigma_2} e^{\frac{-(Y_i - X_i'\beta_2)^2}{2\sigma_2^2}} \right) e^{-\frac{\lambda}{n}\frac{||\beta_1||_1}{\sigma_1}} e^{-\frac{\lambda}{n}\frac{||\beta_2||_1}{\sigma_2}} \right\}. \end{aligned} \tag{C.24}$$

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 $-n^{-1}\ell_{pen}^{(0)}(\theta)$ 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 $\beta_1^*$ such that $Y_i - X_i'\beta_1^* = 0$ and let $\sigma_1 \to 0$. Without the penalty term $\exp(-\frac{\lambda}{n}\frac{||\beta_1^*||_1}{\sigma_1})$ in (C.24) the function would tend to infinity as $\sigma_1 \to 0$. But as $Y_i \neq 0$ for

all $i \in 1, \ldots, n$, $\beta_1^*$ cannot be zero and therefore $\exp(-\frac{\lambda}{n}\frac{||\beta_1^*||_1}{\sigma_1})$ 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

$$0 < \min Y_i^2 - \varepsilon_1 \leq (Y_i - X_i\beta_1)^2 \tag{C.25}$$

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

$$0 < \min Y_i^2 - \varepsilon_1 \leq (Y_i - X_i\beta_2)^2 \tag{C.26}$$

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

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

$$\frac{1}{\sigma_1} e^{-\frac{(\min Y_i^2 - \varepsilon_1)}{2\sigma_1^2}} < \varepsilon_2 \quad \text{and} \quad \frac{1}{\sigma_1} e^{-\frac{\lambda}{n}\frac{m}{\sigma_1}} < \varepsilon_2 \tag{C.27}$$

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

$$\frac{1}{\sigma_2} e^{-\frac{(\min Y_i^2 - \varepsilon_1)}{2\sigma_2^2}} < \varepsilon_2 \quad \text{and} \quad \frac{1}{\sigma_2} e^{-\frac{\lambda}{n}\frac{m}{\sigma_2}} < \varepsilon_2 \tag{C.28}$$

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.25)-(C.28) we easily see that $u(\xi)$ is also bounded on $K^c$ and therefore bounded on $\Xi$.

□

*Proof of Theorem 6.1.* 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

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

$$\theta = (\phi_1, \ldots, \phi_k, \rho_1, \ldots, \rho_k, \pi_1, \ldots, \pi_{k-1}) \in \Theta = \mathbb{R}^{kp} \times \mathbb{R}^k_{>0} \times \Pi \subset \mathbb{R}^{kp+k+(k-1)} = \mathbb{R}^D$$
$$\Pi = \{\pi; \pi_r > 0 \text{ for } r = 1, \ldots, k-1 \text{ and } \sum_{r=1}^{k-1} \pi_r < 1\}, \qquad \pi_k = 1 - \sum_{r=1}^{k-1} \pi_r.$$

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

$$\nu_{pen}(\theta) = -n^{-1}\ell^{(0)}_{pen,\lambda}(\theta) = -n^{-1}\log f_{obs}(Y|\theta) + \lambda\sum_{r=1}^{k}||\phi_r||_1 = Q_{pen}(\theta|\theta') - H(\theta|\theta') \tag{C.29}$$

where $Q_{pen}(\theta|\theta') = -n^{-1}\mathbb{E}[\log f_c(Y,\Delta|\theta)|Y,\theta'] + \lambda\sum_{r=1}^{k}||\phi_r||_1$ (compare Section 6.1) and $H(\theta|\theta') = -n^{-1}\mathbb{E}[\log k(Y,\Delta|Y,\theta)|Y,\theta']$.

By Jensen's inequality we get the following important relationship:

$$H(\theta|\theta') \geq H(\theta'|\theta') \qquad \forall \quad \theta \in \Theta, \tag{C.30}$$

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

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

We are now ready to start with the proof which is inspired by Bertsekas (1995).

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.29) 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.30)). 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:

$$Q_{pen,\theta^m}(\theta^m) \geq Q_{pen,\theta^m}(z_1^m) \geq \ldots \geq Q_{pen,\theta^m}(z_{D-1}^m) \geq Q_{pen,\theta^m}(\theta^{m+1}). \tag{C.31}$$

Additionally from the properties of GEM (equation (C.29) and (C.30)) we have:

$$\nu_{pen}(\theta^0) \geq \nu_{pen}(\theta^1) \geq \ldots \geq \nu_{pen}(\theta^m) \geq \nu_{pen}(\theta^{m+1}). \tag{C.32}$$

Equation (C.32) 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:

$$\begin{aligned} 0 \leq Q_{pen,\theta^m}(\theta^m) - Q_{pen,\theta^m}(\theta^{m+1}) &= \nu_{pen}(\theta^m) - \nu_{pen}(\theta^{m+1}) + \underbrace{H_{\theta^m}(\theta^m) - H_{\theta^m}(\theta^{m+1})}_{\leq 0} \\ &\leq \underbrace{\nu_{pen}(\theta^m) - \nu_{pen}(\theta^{m+1})}_{\to \nu_{pen}(\bar{\theta}) - \nu_{pen}(\bar{\theta}) = 0}. \end{aligned} \tag{C.33}$$

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 $\theta^{m_j}$ and $z_1^{m_j}$, and belongs to $\Theta$ because $\Theta$ is convex. As $Q_{pen,\theta^{m_j}}(.)$ is convex and $z_1^{m_j}$ minimizes this function over all values that differ from $\theta^{m_j}$ along the first coordinate, we obtain

$$Q_{pen,\theta^{m_j}}(z_1^{m_j}) = Q_{pen,\theta^{m_j}}(\theta^{m_j} + \gamma^{m_j}s_1^{m_j}) \leq Q_{pen,\theta^{m_j}}(\theta^{m_j} + \varepsilon\bar{\gamma}s_1^{m_j}) \leq Q_{pen,\theta^{m_j}}(\theta^{m_j}). \tag{C.34}$$

From equation (C.31) and (C.34) we conclude

$$\begin{aligned} 0 \quad &\leq \quad Q_{pen,\theta^{m_j}}(\theta^{m_j}) - Q_{pen,\theta^{m_j}}(\theta^{m_j} + \varepsilon\bar{\gamma}s_1^{m_j}) \overbrace{\leq}^{(C.34)} Q_{pen,\theta^{m_j}}(\theta^{m_j}) - Q_{pen,\theta^{m_j}}(z_1^{m_j}) \\ &\overbrace{\leq}^{(C.31)} \quad Q_{pen,\theta^{m_j}}(\theta^{m_j}) - Q_{pen,\theta^{m_j}}(\theta^{m_j+1}). \end{aligned}$$

Using (C.33) and continuity of $Q_{pen,x}(y)$ in both arguments $x$ and $y$ we conclude by taking the limit $j \to \infty$:

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

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 coordinatewise 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 $d$.

□

## References


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

Bickel, P., Ritov, Y. and Tsybakov, A. (2007) Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics, to appear.*

Dempster, A., Laird, N. and 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. and 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. and Tibshirani, R. (2007) Pathwise coordinate optimization. *Annals of Applied Statistics*, **1**, 302–332.

Fu, W. J. (1998) Penalized regression: the bridge versus the lasso. *Journal of Computational and Graphical Statistics*, **7**, 397–416.

Greenshtein, E. and Ritov, Y. (2004) Persistence in high-dimensional predictor selection and the virtue of over-parametrization. *Bernoulli*, **10**, 971–988.

Huang, J., Ma, S. and Zhang, C.-H. (2008) Adaptive Lasso for sparse high-dimensional regression models. *Statista Sinica*, **18**, 1603–1618.

Khalili, A. and Chen, J. (2007) Variable selection in finite mixture of regression models. *Journal of the American Statistical Association*, **102**, 1025–1038.

Lehmann, E. (1983) *Theory of Point Estimation*. Pacific Grove, CA: Wadsworth and Brooks/Cole.

McLachlan and Peel (2000) *Finite mixture models*. Wiley, New York.

Meier, L., van de Geer, S. and Bühlmann, P. (2008) The group lasso for logistic regression. *Journal of the Royal Statistical Society, Series B*, **70**, 53–71.

Meinshausen, N. and Bühlmann, P. (2006) High dimensional graphs and variable selection with the lasso. *Annals of Statistics*, **34**, 1436–1462.

Pan, W. and Shen, X. (2007) Penalized model-based clustering with application to variable selection. *Journal of Machine Learning Research*, **8**, 1145–1164.

Park, T. and 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.

Tsybakov, A. and van de Geer, S. (2005) Square root penalty: adaptation to the margin in classification and in edge estimation. *Annals of Statistics*, **33**, 1203–1224.

van de Geer, S. (2008) High-dimensional generalized linear models and the lasso. *Annals of Statistics*, **36**, 614–645.

van der Vaart, A. (2007) *Asymptotic Statistics*. Cambridge University Press.

Wu, C. (1983) On the convergence properties of the em algorithm. *Annals of Statistics*, **11**.

Zhao, P. and Yu, B. (2006) On model selection consistency of lasso. *Journal of Machine Learning Research*, **7**, 2541–2563.

Zhou, S., van de Geer, S. and Bühlmann, P. (2009) Adaptive Lasso for high dimensional regression and Gaussian graphical modeling. *Arxiv preprint math.ST/0903.2515*.

Zou, H. (2006) The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, **101**, 1418–1429.