Biometrics 000, 1–21 DOI: 000
0000

# Screening and Clustering of Sparse Regressions with Finite Non-Gaussian Mixtures

**Jian Zhang**

School of Mathematics, Statistics and Actuarial Science

University of Kent, Canterbury, Kent CT2 7NF, United Kingdom

*email:* jz79@kent.ac.uk

Summary: This article proposes a method to address the problem that can arise when covariates in a regression setting are not Gaussian, which may give rise to approximately mixture-distributed errors, or when a true mixture of regressions produced the data. The method begins with non-Gaussian mixture-based marginal variable screening, followed by fitting a full but relatively smaller mixture regression model to the selected data with help of a new penalization scheme. Under certain regularity conditions, the new screening procedure is shown to possess a sure screening property even when the population is heterogeneous. We further prove that there exists an elbow-point in the associated scree plot which results in a consistent estimator of the set of active covariates in the model. By simulations, we demonstrate that the new procedure can substantially improve the performance of the existing procedures in the content of variable screening and data clustering. By applying the proposed procedure to motif data analysis in molecular biology, we demonstrate that the new method holds promise in practice.

Key words: Heterogeneity, non-Gaussian mixture regression models, component-wise regularization, simultaneous clustering and variable screening.

## 1. Introduction

The advance of high-throughput technology in science has allowed scientists to collect data of unprecedented size and complexity. Such large-scale data are often characterized by a certain degree of heterogeneity (a concept used in statistics relating to the non-uniformity in the composition of a population) as they may arise from different sources. The large-scale data hold great promise for discovering subtle population patterns that are not possible with small-scale data (Fan et al., 2014). For example, one of the most successful computational tools for finding transcription factor DNA-binding motifs is the linear regression of gene expressions on motif-matching scores (Colon et al., 2003). The homogeneity assumption that the regression coefficients are the same for all observations underpins the above tool. However, the recent study has demonstrated that there exist heterogeneous structures in the data (Khalili et al., 2011). Similarly, in gene microarray expression data, researchers found that only a fraction of conditions (i.e., covariates) may exhibit an influence on the response in a subset of observations (Zhang, 2010). Therefore, the use of homogeneous population models in these studies can be inadequate. Heterogeneity can also arise in high-dimensional regression after variable selection (Fan and Lv, 2008). The aim of variable selection is to screen out variables with weak effects in the model. Although weak variables may have non-zero effects on the response, the existing variable selection procedures such as LASSO often assign zero values to the regression coefficients of these unselected variables in order to have a selection effect (Tibshirani, 1996). After variable selection, many weak variables can be filtered out from the model, resulting in heterogeneous residuals due to aggregate effects of dropping these weak variables. Therefore, the regression model after variable selection can be misspecified, where the use of a heterogeneous model is desirable.

Over the past two decades, much progress has been made on how to incorporate heterogeneous structures into a model with mixture distributions (McLachlan and Peel, 2000). In

particular, Gupta and Ibrahim (2007), Städler et al. (2010), and Khalili et al. (2011) presented a finite Gaussian mixture model for modeling heterogeneity in the regression setting. In these seminal works, the authors either imposed a penalty on the likelihood or introduced a Bayesian prior on the parameters. Despite the above progress, there are still the following practically important issues remained to address. First, Gaussian mixture regression models may not be robust to model misspecifications: slight deviations from normality in mixture-components can lead to spurious groupings. In particular, our simulations suggest that the commonly used marginal models in variable screening may not be Gaussian even the full model is Gaussian. This is a parallel development to Fan et al. (2011) where they addressed the non-linearity of marginal regression functions in the marginal screening when covariates are not normally distributed. Secondly, the standard method for regularizing the above mixture models is to add a composite penalty to the log-likelihood as suggested by Khalili et al.(2011) and Städler et al. (2010). A drawback of their method is that the resulting GEM algorithm has no explicit updating formula for estimating the mixture proportions and thus requires an optimization over a simplex (Städler et al., 2010). Finally, when both the dimension and the sample size are large, the computational cost of the GEM algorithm is prohibitive. To reduce the cost, a fast variable screening is required to search for a smaller mixture model. The commonly used screening method is the so-called correlation screening (Fan and Lv, 2008). However, it is largely unknown in the literature when marginal variable screening can consistently recover the true active covariates in a mixture regression model.

Here, to address the above issues, we propose an exponential power distribution (EPD) based mixture regression model (EPDMIX) as a flexible extension of the standard Gaussian mixture regression model. The proposed model is then used to define a two-stage procedure for carrying out variable selection and data clustering simultaneously. The procedure begins with non-Gaussian mixture-based marginal variable screening, followed by fitting a full

but relatively smaller mixture regression model to the selected data with help of a new penalization scheme. To our knowledge, the idea of using univariate mixture regression models to screen variables is completely new in the literature. The mechanism of the proposed screening procedure can be explained by solving the variable selection problem for the model $y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i, \quad 1 \leqslant i \leqslant n$, where $\varepsilon_i$'s are i.i.d. $N(0,1)$. To screen variables, for each $j$, we single out the $j$-th covariate and rewrite the above equation as $y_i = x_{ij}\beta_j + \varepsilon_i^*$, with $\varepsilon_i^* = \sum_{t\neq j} x_{it}\beta_t + \varepsilon_i$, $1 \leqslant i \leqslant n$. If the covariate observations $\{x_{it} : 1 \leqslant t \leqslant p\}$ follow a multivariate normal distribution, then $\varepsilon_i^*$ is homogeneous with a Gaussian distribution. However, if these observations have a group structure, then $\varepsilon_i^*$'s are heterogeneously distributed. Therefore, a mixture model-based marginal variable screening is necessary. See the Web Appendix $A$, the Web-based Supplementary Material for more details. The new penalization scheme called component-wise regularization on the likelihood is employed to improve the existing penalization scheme (Khalili et al., 2011; Städler et al.,2010). In the proposed scheme, the number of components and the penalty coefficient are simultaneously selected by optimizing the so-called Bayesian Information Criterion (BIC) over a restricted region. A new block-wise GEM algorithm is developed to compute the corresponding maximum penalized likelihood estimators. Unlike the existing GEM algorithms (Khalili et al., 2011; Städler et al.,2010), under the new penalization scheme, explicit updating formulas for estimating mixture proportions are obtained, which speed up the computation. The proposed GEM algorithm is further shown to have the non-descent property with respect to maximizing the penalized log-likelihood. As the main contribution of our paper, we establish the sure screening property for the proposed procedure when the population is heterogeneous. We further prove that there exists an elbow-point in the BIC scree plot which results in a consistent estimator of the set of active covariates in the model.

We conduct a series of simulation studies and a real data analysis to evaluate the perfor-

mance of the proposed procedure with a comparison to the Gaussian mixture-based approach (GAUMIX). There are various ways to summarize the performance of a mixture regression model. Khalili et al. (2011), and Städler et al. (2010) focused on the accuracy of variable selection and not on the clustering. They assumed that the number of components was known and fixed when comparing different mixture regression models. In our simulations, we remove this assumption. We assess the performance of the proposed mixture-based variable screening. We then evaluate the accuracy of the mixture model-based clustering in the Web Appendix E, the Web-based Supplementary Materials. Our simulation results show that EPDMIX can have superior performance in variable screening over GAUMIX, the EPD regression (EPD1) and the Gaussian regression (GAU1, the correlation screening), even when the joint distribution of response and covariates is Gaussian. In particular, the EPD1-based screening can improve the GAU1-based screening without significantly compromising its computational speed. The results also show that the component-wise penalization does improve the quality of the clustering derived from non-penalized mixture regression. Moreover, EPDMIX can accurately identify the number of components most times even in the presence of heavy tailed errors. The proposed EPDMIX method is applied to a motif dataset obtained from a biological study. For many covariates, their EPDMIX-based reciprocal BIC values are much higher than the corresponding GAUMIX-based reciprocal BIC values as shown in Figure 1. This implies that EPDMIX can significantly improve GAUMIX in variable screening. Two clusters of genes with the selected motifs are predicted, which show the links between genes and DNA motifs. The biological implication of clustered genes are also given.

[Put Figure 1 about here.]

The remaining of the paper is organized as follows. The details of the proposed methodology and algorithm are provided in Section 2. A new theory on the proposed screening procedure

is developed in Section 3. The simulation studies and a real data application are presented in Sections 4 and 5. The discussion and conclusion are made in Section 6. The technical details and the proofs of the theory can be found in the Web-based Supplementary Materials.

## 2. Methodology

Let $(y_i, \mathbf{x}_i), i = 1, ..., n$ be independent observations on response $y$ and $p$-dimensional covariate $\mathbf{x}$. Suppose that the conditional density of $y_i$ given $\mathbf{x}_i$ is a $K$-component exponential power mixture which can be written as

$$f(y_i|\mathbf{x}_i, \Theta_K) = \sum_{k=1}^{K} \pi_k \phi(y_i|\mathbf{x}_i^T\boldsymbol{\beta}_k, \sigma_k^2, \alpha_k),$$

where $\Theta_K$ denotes the set of all the parameters, $\phi(y_i|\mathbf{x}_i^T\boldsymbol{\beta}_k, \sigma_k^2, \alpha_k)$ is the $k$-th component density of the form

$$\phi(y_i|\mathbf{x}_i^T\boldsymbol{\beta}_k, \sigma_k^2, \alpha_k) = \frac{\alpha_k}{2\sigma_k\Gamma(1/\alpha_k)} \exp\left(-\frac{|y_i - \mathbf{x}_i^T\boldsymbol{\beta}_k|^{\alpha_k}}{\sigma_k^{\alpha_k}}\right)$$

with regression coefficients $\boldsymbol{\beta}_k = (\beta_{k1}, ..., \beta_{kp})^T \in \mathbb{R}^p$, scale parameter $\sigma_k^2 \in (0, \infty)$, shape parameter $\alpha_k \in (0, \infty)$, proportion $\pi_k \geqslant 0$, and $\sum_{k=1}^{K} \pi_k = 1$. For simplicity, we denote the exponential power distribution $\phi(y|\mu, \sigma^2, \alpha)$ by $\text{epd}(\mu, \sigma, \alpha)$ in the remainder of the paper. Let $\boldsymbol{\theta}_k = (\boldsymbol{\beta}_k, \sigma_k^2, \alpha_k, \pi_k)^T$, $k = 1, ..., K$. Then $\Theta_K = (\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_K)$. The family of exponential power distributions takes the Normal and Laplace distributions as special cases when setting $\alpha_1 = \cdots = \alpha_K = 2$ and $\alpha_1 = \cdots = \alpha_K = 1$ respectively. Our principal interest here is to infer the latent components, to group the observations, and to identify the covariates with non-zero regression coefficients for each component.

### 2.1 *Penalized likelihood estimation and algorithms*

The classical maximum likelihood estimator (MLE) is calculated by maximizing the likelihood function given by

$$L_n(\Theta_K|\mathbf{Y}, \mathbf{X}) = \prod_{i=1}^{n} f(y_i|\mathbf{x}_i, \Theta_K),$$

where $\mathbf{Y} = (y_1, ..., y_n)^T$ and $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n)^T$. When $p = p_n$ is larger than the sample size $n$, the above problem is ill-posed. To tackle the problem, we derive a penalty on the likelihood by a prior distribution as follows.

Following Städler et al.(2010), we first introduce the scale-invariant parameter $\boldsymbol{\eta}_k = \boldsymbol{\beta}_k/\sigma_k$. Then, the $k$-component density can be re-parametrized as

$$\phi(y_i|\mathbf{x}_i^T\eta_k, \sigma_k^2, \alpha_k) = \frac{\alpha_k}{2\sigma_k\Gamma(1/\alpha_k)} \exp\left(-\left|\frac{y_i}{\sigma_k} - \mathbf{x}_i^T\boldsymbol{\eta}_k\right|^{\alpha_k}\right).$$

The re-parametrization and the form of a particular log-prior are used as a basis for determining a scale-invariant penalty function for the original parameters.

For $K = 1$, the Laplace-inverse-gamma priors are set for $(\eta_1, \sigma_1^2, \alpha_1)$:

$$p(\eta_1|\sigma_1^2) \propto \exp(-n\lambda|\boldsymbol{\eta}_1|_1), \quad p(\sigma_1^2) \propto \frac{1}{\sigma_1^2}\exp(-\frac{n\kappa_0}{\sigma_1}), \quad p(\alpha_1) \propto 1,$$

where $\kappa_0$ is a pre-specified constant with default of $\kappa_0 = 0$, and $|\cdot|_1$ denotes the $L_1$ norm. The penalized likelihood can be derived from the posterior

$$p(\Theta_1|\mathbf{Y}, \mathbf{X}) \quad \propto \quad \prod_{i=1}^{n} \phi(y_i|\mathbf{x}_i^T\boldsymbol{\beta}_1, \sigma_1^2, \alpha_1) \exp\left(-\frac{\lambda|\boldsymbol{\beta}_1|_1 + \kappa_0/n}{\sigma_1}\right) \frac{1}{\sigma_1^{2/n}},$$

which is the product of the penalized likelihoods of individual observations.

For $K \geqslant 2$, a traditional penalized likelihood was developed by Khalili et al. (2011), which is proportional to

$$\sum_{k=1}^{K} \pi_k \phi(y_i|\mathbf{x}_i^T\boldsymbol{\beta}_k, \sigma_k^2, \alpha_k) \times \exp\left(-n\sum_{k=1}^{K} \pi_k \left(\lambda|\boldsymbol{\eta}_k|_1 + \frac{\kappa_0}{\sigma_k}\right)\right).$$

As pointed out before, there is no explicit formula for updating mixture proportions $\pi_k$'s in the M-step of the EM algorithm if we use the above penalization. Here, to tackle the issue, we derive an alternative penalized likelihood by use of a non-standard log-posterior below. The basic idea is that for each observation, we first construct the penalized likelihoods for all components and then combine these likelihoods together by a component-wise weighting:

$$\mathrm{pL}_n(\Theta_K|(y_i, \mathbf{x}_i)) = \sum_{k=1}^{K} \pi_k \phi(y_i|\mathbf{x}_i^T\boldsymbol{\beta}_k, \sigma_k^2, \alpha_k) \exp\left(-\frac{\lambda|\boldsymbol{\beta}_k|_1 + \kappa_0}{\sigma_k}\right) \frac{1}{\sigma_k^{2/n}}.$$

The penalized incomplete-data likelihood is then defined by multiplying the above individual

likelihoods and priors of $\pi_k$'s as follows:

$$\mathrm{pL}_n(\Theta_K|(\mathbf{Y},\mathbf{X})) = \prod_{i=1}^{n} \mathrm{pL}_n(\Theta_K|(y_i,\mathbf{x}_i)) \prod_{k=1}^{K} \pi_k^{\delta_k},$$

where $\delta_k$, $k = 1, ..., K$ are pre-specified constants with default $\delta_k = 1/K$. Note that we are unable to observe the group memberships of individual subjects, which are defined by the indicator functions $\mathbf{Z} = (z_{ik})_{1\leqslant i\leqslant n, 1\leqslant k\leqslant K}$,

$$z_{ik} = \begin{cases} 1, & \text{if the } i\text{-th subject belongs to the } k\text{-th group,} \\ 0, & \text{otherwise.} \end{cases}$$

The penalized likelihood for the complete data $(\mathbf{Y},\mathbf{X},\mathbf{Z})$ is then given by

$$\mathrm{pL}_n(\Theta_K|\mathbf{Y},\mathbf{X},\mathbf{Z}) = \prod_{k=1}^{K} \pi_k^{\delta_k} \prod_{i=1}^{n}\prod_{k=1}^{K} \left\{ \pi_k \phi(y_i|\mathbf{x}_i^T\boldsymbol{\beta}_k, \sigma_k^2, \alpha_k) \exp\left(-\frac{\lambda|\boldsymbol{\beta}_k|_1 + \kappa_0}{\sigma_k}\right) \frac{1}{\sigma_k^{2/n}} \right\}^{z_{ik}}.$$

Following McLachlan and Peel (2000) and using the complete-data likelihood, we can implement a block-wise GEM algorithm in the following two steps.

*E-step*: Calculate the conditional expectation of $\mathbf{Z}$ given $(\mathbf{Y},\mathbf{X})$ and the current estimate $\Theta_K^{(v)}$, giving

$$\begin{aligned} \Psi &= \Psi(\Theta_K|\Theta_K^{(v)}) = E\left\{\log(\mathrm{pL}_n(\Theta_K|\mathbf{Y},\mathbf{X},\mathbf{Z}))|\mathbf{Y},\mathbf{X},\Theta_K^{(v)}\right\} \\ &= \sum_{k=1}^{K}\left(\sum_{i=1}^{n}\tau_{ik}^{(v)} + \delta_k\right)\log(\pi_k) + \sum_{k=1}^{K}\sum_{i=1}^{n}\tau_{ik}^{(v)}\log\left(\frac{\alpha_k}{2\sigma_k^{(1+2/n)}\Gamma(1/\alpha_k)}\right) \\ &\quad - \sum_{k=1}^{K}\sum_{i=1}^{n}\tau_{ik}^{(v)}\frac{|y_i - \mathbf{x}_i^T\boldsymbol{\beta}_k|^{\alpha_k}}{\sigma_k^{\alpha_k}} - \sum_{k=1}^{K}\sum_{i=1}^{n}\tau_{ik}^{(v)}\frac{\lambda|\boldsymbol{\beta}_k|_1 + \kappa_0}{\sigma_k}, \end{aligned}$$

where

$$\tau_{ik}^{(v)} = \frac{\pi_k^{(v)}\phi(y_i|\mathbf{x}_i^T\boldsymbol{\beta}_k^{(v)}, \sigma_k^{(v)2}, \alpha_k^{(v)}) \exp(-(\lambda|\boldsymbol{\beta}_k^{(v)}|_1 + \kappa_0)/\sigma_k^{(v)}) \frac{1}{\sigma_k^{(v)2/n}}}{\sum_{k=1}^{K}\pi_k^{(v)}\phi(y_i|\mathbf{x}_i^T\boldsymbol{\beta}_k^{(v)}, \sigma_k^{(v)2}, \alpha_k^{(v)}) \exp(-(\lambda|\boldsymbol{\beta}_k^{(v)}|_1 + \kappa_0)/\sigma_k^{(v)}) \frac{1}{\sigma_k^{(v)2/n}}}.$$

*M-step*: To update the estimate of $\Theta$, we maximize $\Psi(\Theta_K|\Theta_K^{(v)})$ with respect to $\Theta_K$ block by block. In particular, by using the Lagrange multiplier on $\Psi$, we update the block of $\pi_k$'s by

$$\hat{\pi}_k^{(v+1)} = \frac{\sum_{i=1}^{n}\tau_{ik}^{(v)} + \delta_k}{n + \sum_{k=1}^{K}\delta_k}.$$

See Web Appendices B, C and D, the Web-based Supplementary Materials for technical

details, for the way to initialize the GEM algorithm and for the BIC-based approach to choosing the penalty coefficient and the number of components.

### 2.2 *Marginal variable screening by mixtures*

In the previous sections, we build an exponential power mixture model to utilize group structures in the data. However, the computational cost of the GEM algorithm prevents it from applications to data with a large size and a large number of covariates. To mitigate the impact of high-dimension, a marginal variable screening is required to reduce the dimension before fitting a full model to the data. We make the following sparsity assumption: although there are many covariates with varying contributions to the response variable, only a few of them are significantly important and majority of them have only marginal effects. The marginal variable screening aims to filter out variables with marginal effects in the model. In this paper, we considered the following four screening procedures: correlation learning or simple Gaussian linear regression (GAU1), simple EPD linear regression (EPD1), simple Gaussian mixture regression (GAUMIX), and simple EPD mixture regression (EPDMIX). All are with the penalty coefficient $\lambda = 0$. For each $j$, $1 \leqslant j \leqslant p$, we fit the above four models to the data $(y_i, x_{ij}), i = 1, ..., n$ respectively and calculate the corresponding reciprocal BIC value $_{\text{BIC}}{}_{Kj}$, where $K$ is restricted to $1 \leqslant K \leqslant K_n$. Then, we calculate the reciprocal BIC values for EPD1 and GAU1, and the reciprocal minimum BIC value $_{\text{rBIC}}{}_j = 1/\min_{1 \leqslant K \leqslant K_n} {}_{\text{BIC}}{}_{Kj}$ for EPDMIX and GAUMIX. We rank these values in decreasing order $_{\text{rBIC}}{}_{(j)}$ and plot them again index $(j)$. The resulting plot is called scree plot. We choose these covariates with the reciprocal (minimum) BIC values larger than the elbow point of the curve in the scree plot.

## 3. Theory

In this section, we investigate the sure screening property for the proposed procedure. To do so, we need to introduce more notations. As before, we consider an independent sample $(y_i, \mathbf{x}_i), i = 1, ..., n$ with $y_i|\mathbf{x}_i \sim f_K(y|\mathbf{x}_i, \Theta_K^*)$. Let $\Theta_{Kj} = ((\beta_{1j}, \sigma_1^2, \alpha_1, \pi_1)^T, ..., (\beta_{Kj}, \sigma_K^2, \alpha_K, \pi_K)^T)$ denote the parameters that are used to link $y_i$ to the $j$-th covariate $x_j$. To facilitate our technical derivations, we restrict $\Theta_{Kj}$ to taking values in a bounded set $\Xi_{Kj}$ with

$$\Xi_{Kj} = \{\Theta_{Kj} : \pi_b \leqslant \pi_k \leqslant 1, 1 \leqslant \alpha_k \leqslant \alpha_u, \sigma_b \leqslant \sigma_1 \leqslant \sigma_u, |\beta_{kj}| \leqslant \beta_u, 1 \leqslant k \leqslant K; \sum_{k=1}^{K} \pi_k = 1\},$$

where $\pi_b$, $\alpha_u$, $\sigma_b$, $\sigma_u$, and $\beta_u$ are positive constants, and $\pi_b$ and $\sigma_b$ are arbitrarily small. Let $f_{K_0}(y|\mathbf{x}, \Theta_{K_0}^{(0)})$ be the true density of $y$ given $\mathbf{x}$ and $f(\mathbf{x})$ the density of $\mathbf{x}$. Note that for each $K$, the true parameter $\Theta_{K_0}^{(0)}$ may not be in $\Xi_{Kj}$. But we find a value in $\Xi_{Kj}$ which is most close to $\Theta_{K_0}^{(0)}$, namely

$$\Theta_{Kj}^* = \text{argmax}_{\Theta_{Kj} \in \Xi_{Kj}} \int \log \left( f_K(y|x_j, \Theta_{Kj}) / f_{K_0}(y|\mathbf{x}, \Theta_{K_0}^{(0)}) \right) f_{K_0}(y|\mathbf{x}, \Theta_{K_0}^{(0)}) f(\mathbf{x}) dy d\mathbf{x}.$$

Replacing $\{\beta_{1j}^*, ..., \beta_{Kj}^*\}$ in $\Theta_{Kj}^*$ by zeros, we define a background model with zero signals and parameter $\Theta_{Kj}^{*0}$. For $0 < \delta < 1/2$, we consider a neighborhood of $\Theta_{Kj}^*$, defined by $[\delta]_{Kj} = \{\Theta_{Kj} : \Theta_{Kj} \in \Xi_{Kj}, |\Theta_{Kj} - \Theta_{Kj}^*| \leqslant \delta\}$, where $|\cdot|$ is the $L_1$ norm. We define the norm

$$||h||_{P_n} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} h^2(y_i, x_{ij})}$$

and the subset of functions

$$\mathcal{F}_{Kj}(\delta) = \{\log(f_K(y|x_j, \Theta_{Kj})) I_n(y, x_j) : \Theta_{Kj} \in [\delta]_{Kj}\}.$$

Let $H(\cdot, \mathcal{F}_{Kj}(\delta), ||\cdot||_{P_n})$ be the entropy of $\mathcal{F}_{Kj}(\delta)$ equipped with the metric induced by the norm $||\cdot||_{P_n}$.

For any $(K, K_1)$ with $1 \leqslant K, K_1 \leqslant K_n$, we use the following average Kullback-Leibler discrepancy to measure the distance from $\Theta_{Kj}$ to $\Theta_{K_1 j}$:

$$\text{KL}(\Theta_{Kj}|\Theta_{K_1 j}) = -\int \log \left( \frac{f_K(y|x_j, \Theta_{Kj})}{f_{K_1}(y|x_j, \Theta_{K_1 j})} \right) f_{K_0}(y|\mathbf{x}, \Theta_{K_0}^{(0)}) f(\mathbf{x}) dy d\mathbf{x}.$$

By the definition of $\Theta_{Kj}^*$, $\text{KL}(\Theta_{Kj}|\Theta_{Kj}^*) \geqslant 0$.

To obtain the convergence rate of the maximum penalized likelihood estimator, we assume the following conditions of identification used in Städler et al. (2010) and Zhang and Liang (2010).

(C1): There exists a positive constant $d_K$ depending on $K$ such that uniformly for $1 \leqslant K \leqslant K_n$, $1 \leqslant j \leqslant p_n$, and $\Theta_{Kj} \in \Xi_{Kj}$,

$$\mathrm{KL}(\Theta_{Kj}|\Theta^*_{Kj}) \geqslant ||\Theta_{Kj} - \Theta^*_{Kj}||^2/d_K^2.$$

And for $\Theta_{Kj} \in \Xi_{Kj}$ and $\Theta_{K_1j} \in \Xi_{K_1j}$, if $\mathrm{KL}(\Theta_{Kj}|\Theta_{K_1j}) = 0$, then $K = K_1$ and $\Theta_{Kj}$ is equal to $\Theta_{K_1j}$ up to a permutation of $K_1$ components.

In Web Appendix E, the Web-based Supplementary Materials, we showed that Condition (C1) holds when $K_n$ is bounded and the Fisher information matrix is bounded away from zero. Similar to Fan and Song (2010), we also need to impose a sub-exponential restriction on each covariate in the model.

(C2): There exists positive constants $r_0$, $r_1$ and $\nu_0$ independent of $1 \leqslant j \leqslant p$, such that for all $t > 0$ and covariate $x_j$, $P(|x_j| > t) \leqslant r_1 \exp(-r_0 t^{\nu_0})$.

For positive constants $V_n$ and $K_n$, let

$$M_n = O(V_n^{\alpha_u+1/2}\log(V_n)), \quad \delta_n = M_n \log(n)\sqrt{\log(n)/n}, \quad \delta_{K_n} = K_n(1+\alpha_u+\sigma_u+\beta_u).$$

We call a covariate active if its regression coefficients are non-zeros at least in one of mixture components. Let $J_K^*$ denote the set of active covariates, $\{1 \leqslant j \leqslant p_n : \sum_{k=1}^K |\beta^*_{kj}| \neq 0\}$. We assume the following identification condition for active covariates, which says when $x_j$ is not active, the associated parameter $\Theta^*_{Kj}$ must be in the $o(n^{-2\kappa})$- neighborhood of the background parameter $\Theta^{*0}_{Kj}$. When $x_j$ is active, the Kullback-Leibler distance from $\Theta^*_{Kj}$ to $\Theta^{*0}_{Kj}$ has order not less than $O(n^{-2\kappa})$.

(C4): Uniformly for $1 \leqslant j \leqslant p_n$, $K_0 \leqslant K \leqslant K_n$,

$$\begin{aligned}
\mathrm{KL}(\Theta^{*0}_{Kj}|\Theta^*_{Kj}) &= o(n^{-2\kappa}), \text{ if } x_j \text{ is not active,} \\
\mathrm{KL}(\Theta^{*0}_{Kj}|\Theta^*_{Kj}) &\geqslant c_9 n^{-2\kappa}, \text{ if } x_j \text{ is active,}
\end{aligned}$$

where $0 < \kappa < 1/2$ and $c_9 > 0$ are constants.

Note that the BIC index $\text{BIC}_j$ for the covariate $x_j$ is defined by $\text{BIC}_j = \min_{1 \leqslant K \leqslant K_n} \text{BIC}_{Kj}$ with

$$\text{BIC}_{Kj} = -\frac{1}{n}\sum_{i=1}^{n} \log\left(f_K(y_i|x_{ij}, \hat{\Theta}_{Kj})\right) + \frac{(4K-1)\log(n)}{n},$$

where $\hat{\Theta}_{Kj}$ is the marginal maximum likelihood estimator. We rank rBIC (short for the reciprocal BIC) values in decreasing order, say $_{\text{rBIC}}{}_j, 1 \leqslant j \leqslant p_n$, and plot them against their indices. For each $2 \leqslant j \leqslant p_n$, we fit a straight line to $\{(t, {}_{\text{rBIC}}{}_t), j \leqslant t \leqslant p_n\}$, obtaining a predictive value $_{\text{prBIC}}{}_{j-1}$ for $_{\text{rBIC}}{}_j$. For any constant $c_*$, we define a change point (elbow point) $\hat{j}$ on the $_{\text{rBIC}}$ curve by $\hat{j} = \max\{2 \leqslant j \leqslant p_n : {}_{\text{rBIC}}{}_j - {}_{\text{prBIC}}{}_{j-1} > c_* n^{-2\kappa}\}$. The change point $\hat{j}$ divides the covariates into estimated active and non-active groups, namely $\hat{J}_{\text{ac}}$ and $\hat{J}_{\text{na}}$. Let $f_K(y_i) = \sum_{k=1}^{K} \pi_k \phi(y_i|0, \sigma_k^2, \alpha_k)$. Then, similar to the theorem in Web Appendix E, the Web-based Supplementary Materials, we can show that

$$\max_{K_0 \leqslant K \leqslant K_n} \sum_{i=1}^{n} \log(f_K(y_i))/n = \max_{K_0 \leqslant K \leqslant K_n} E\left[\log(f_K(y))\right] + o_p(1).$$

We show in the following theorem that for any constant $c_* > 0$ satisfying

$$P\left(\frac{c_9}{c_*} \geqslant \left(\max_{K_0 \leqslant K \leqslant K_n} \sum_{i=1}^{n} \log(f_K(y_i))/n\right)^2\right) \to 1, \tag{3.1}$$

$\hat{J}_{\text{ac}}$ is consistent with the true active set $J^*_{K_0}$.

THEOREM 1: *We assume that $K_n = O(1)$, $d_{K_n} = O(1)$ and that equation (3.1) holds. Then, under Conditions (C1)$\sim$ (C4), as $n \to \infty$, we have that*

$$P(J_{ac} = J^*_{K_0}) \to 1.$$

## 4. Numerical results on simulated data

By simulations, we aim (a) to examine the performances of various marginal variable screening methods including GAU1, EPD1,GAUMIX, and EPDMIX, and (b) to investigate whether the EPD can accommodate non-normality. We consider various scenarios, where following

Städler et al. (2010), the Signal-to-Noise-Ratio (SNR) in each data set is measured by

$$\text{SNR} = 1 + \frac{\sum_{k=1}^{K_0} \pi_k \boldsymbol{\beta}_k^T \mathbf{X}^T \mathbf{X} \boldsymbol{\beta}_k}{n \sum_{k=1}^{K_0} \pi_k \sigma_k^2},$$

with the $k$-mixture proportion $\pi_k$, the $k$- regression coefficient $\boldsymbol{\beta}_k$, and the variance of the $k$-th error term $\sigma_k^2$.

In marginal variable screening, we want to identify active covariates in the model. We compare the performances of GAU1, EPD1, GAUMIX, and EPDMIX in screening in terms of specificity and sensitivity. Specificity and sensitivity are defined as the survival rates of true active covariates and of true non-active covariates respectively in screening. We consider the following two settings.

*Setting 4.1.1 (multiple linear regression)*: We generated 140 datasets with the sample size $n$ and the dimension $p$. Each dataset contained observations $(y_i, x_{ij})$, $1 \leqslant j \leqslant p, 1 \leqslant i \leqslant n$ satisfying $y_i = \sum_{j=1}^{p} x_{ij}\beta_{0j} + \varepsilon_i$, where $\varepsilon_i$, $1 \leqslant i \leqslant n$ were iid $N(0,1)$, and the regression coefficients

$$\boldsymbol{\beta}_0 = (2 + \eta_1, 1.6 + \eta_2, 1.2 + \eta_3, 0.8 + \eta_4, 0.4 + \eta_5, 0_{p-5}^T)^T,$$

where $\eta_j, 1 \leqslant j \leqslant 5,$ were iid $N(0, 0.1^2)$, and $0_{p-5}$ was a $p-5$ vector of zeros. There were five active covariates in the above model. Each dataset was generated in the following steps. First, for each $i$, following Fan and Song (2010), we simulated the covariates by

$$\tilde{x}_{ij} = \frac{t_j + a_j \times t_0}{\sqrt{1 + a^2}},$$

where $t_0 \sim t_2$ (t distribution), $a_j = 0.8$, $1 \leqslant j \leqslant 15$ and $a = 0, 16 \leqslant j \leqslant p$, and $t_j \sim t_2, 1 \leqslant j \leqslant p/3$, $t_j \sim e_j \times (2b(1/2) - 1)$, $p/3 \leqslant j \leqslant 2p/3$ with $b(1/2)$ being a Bernoulli distribution of success probability $1/2$ and $e_j$ being drawn from the standard exponential distribution, $t_j \sim$ the mixture $0.5N(-1,1) + 0.5N(1, 0.5)$, $j > 2p/3$. Then, we randomly shuffled the columns of the matrix $(\tilde{x}_{ij})_{n\times p}$, followed by column standardization. Note that randomly shuffling the columns of the design matrix is equivalent to randomly shuffling active variables. Finally, we centralized $\mathbf{Y} = (y_1, ..., y_n)^T$ by the sample mean $\sum_{i=1}^{n} y_i/n$. We

considered two scenarios with $(n, p) = (500, 600)$ and $(100, 2000)$ and the average SNR values of 26 and 0.85, respectively. Note that under Setting 4.1.1, the true model is a single Gaussian regression model. By the simulation, we demonstrated that even though the underlying full model was a single Gaussian regression model, the marginal model at each covariate could be non-Gaussian. This is due to the so-called aggregate misspecification effects of unselected covariates as described in the Introduction.

*Setting 4.1.2 (Gaussian mixture regression)*: We generated 140 datasets with the sample size $n$ and the dimension $p$. Each dataset consists of observations $(y_i, x_{ij})$, $1 \leqslant j \leqslant p, 1 \leqslant i \leqslant n$. The covariates $x_{ij}, 1 \leqslant i \leqslant n$, $1 \leqslant j \leqslant p$ were adopted from Setting 4.1.1. Given $\mathbf{x}_i = (x_{i1}, ..., x_{ip})^T$'s, $y_i$, $1 \leqslant i \leqslant p$ were independently sampled from the mixture distribution

$$f(y_i) = \sum_{k=1}^{K_0} \pi_k \phi(y_i - \mathbf{x}_i^T \boldsymbol{\beta}_k),$$

where $\phi(.)$ is the density of the standard Gaussian distribution, $\boldsymbol{\beta}_k$, $1 \leqslant k \leqslant K_0$ are regression coefficients, and $\pi_k, 1 \leqslant k \leqslant K_0$ are mixture proportions. We then centralized $\mathbf{Y}$ by its sample mean. We considered the following two cases of $K_0$:

(1) $K_0 = 2$, where there are two components with

$$\begin{aligned} \boldsymbol{\beta}_1 &= (2 + v_1, 1.6 + v_2, 1.2 + v_3, 0.8 + v_4, 0.4 + v_5, 0_{p-5}^T)^T, \\ \boldsymbol{\beta}_2 &= (0, 0, 0, 4 + v_{21}, 4 + v_{22}, 4 + v_{23}, 4 + v_{24}, 4 + v_{25}, 0_{p-8}^T)^T, \end{aligned}$$

where $v_j, 1 \leqslant j \leqslant 5,$ are iid $N(0, 0.1^2)$ and $0_{p-5}$ is a $p-5$ vector of zeros.

(2) $K_0 = 3$, where there are three components with

$$\begin{aligned} \boldsymbol{\beta}_1 &= (2 + v_{11}, 1.6 + v_{12}, 1.2 + v_{13}, 0.8 + v_{14}, 0.4 + v_{15}, 0_{p-5}^T)^T, \\ \boldsymbol{\beta}_2 &= (0, 0, 0, 4 + v_{21}, 4 + v_{22}, 4 + v_{23}, 4 + v_{24}, 4 + v_{25}, 0_{p-8}^T)^T, \\ \boldsymbol{\beta}_3 &= (0, 0, 0, 0, 0, 0, -4 + v_{31}, -4 + v_{32}, 0_{p-8}^T)^T, \end{aligned}$$

where $v_{kj}, 1 \leqslant j \leqslant 5, k = 1, 2, v_{31}, v_{32}$ are iid $N(0, 0.1^2)$, and $0_{p-8}$ is a $p-8$ vector of zeros.

For each case of $K_0$, we considered $(n, p) = (300, 400)$ and $(500, 600)$. The average SNR values are around 165 and 175 for $K_0 = 2$, and around 88 and 90 for $K_0 = 3$.

For each case, we applied GAU1, EPD1, GAUMIX and EPDMIX to each of the 140 datasets. That is, for $1 \leqslant j \leqslant p$, we fitted EPD1, GAU1, EPDMIX and GAUMIX regression models to the data $(y_i, x_{ij})_{1 \leqslant i \leqslant n}$ respectively and calculated the corresponding BIC values. The results were summarized in Figure 3, Tables 1 and 2. The percentage increases in specificity when the sensitivity was fixed were calculated by using the formula $(s/s_{\text{gau1}} - 1)$, where $s$ is the specificity of EPD1 or EPDMIX or GAUMIX, and $s_{\text{gau1}}$ is the specificity of GAU1.

In Setting 4.1.1, note that the true active covariates were located at $j = 1, 2, ..., 5$, with the BIC values $\text{BIC}_j$, $1 \leqslant j \leqslant 5$. We put them in a decreasing order, say $\text{BIC}_{(1)} \geqslant \text{BIC}_{(2)} \geqslant \cdots \geqslant \text{BIC}_{(5)}$. If we threshold the BIC values $\text{BIC}_j, 1 \leqslant j \leqslant p$ at the levels of $\text{BIC}_{(j)}$, the corresponding sensitivity of screening will be $j/5$. This enable us to calculate the values of specificity when the sensitivity of screening is set to $1/5, 2/5, 3/5$, $4/5$ and $5/5$ respectively. For each of the above four screening methods, we calculated these values. The results are displayed in the first two rows of plots in Figure 3 and Table 1. The results suggest that EPD1 and EPDMIX outperformed their Gaussian counterparts. In the case where $(n, p) = (500, 600)$, EPD1 improved GAU1 by 33%, 17% and 7% increases of specificity when the sensitivity level was fixed at the levels of 5/5, 4/5 and 3/5 respectively. At these sensitivity levels, on average EPDMIX improved GAU1 by 64%, 31%, and 9% increases in specificity. This is slightly better than GAUMIX, which improved GAU1 by 62%, 30% and 9% increases in specificity. In the case where $(n, p) = (100, 2000)$, although the SNR is low, the average percentage increases in specificity by use of EPD1, EPDMIX and GAUMIX compared to use of GAU1 were still visible.

In Settings 4.1.2, we calculated the values of specificity when the sensitivity level was fixed

at $j/8$, $j = 1, 2, ..., 8$. This can be achieved by taking the BIC values at these true active covariates as the thresholds for the BIC's. The results were summarized in the bottom 4 rows of Figure 3 and in Tables 2. The results indicate that the largest percentage increases in specificity were obtained when we used EPDMIX instead of GAU1. For example, in the case where $K_0 = 2$ and $(n, p) = (300, 400)$, on average EPDMIX had 128%, 97%, 51%, 25%, and 10% increases in specificity over GAU1 when the sensitivity level was fixed at 8/8, 7/8, 6/8, 5/8 and 4/8. It outperformed GAUMIX which had 110%, 85%, 44%, 23% and 10% increases in specificity over GAU1. Close to the performance of EPDMIX, EPD1 had 119%, 90%, 46%, and 24% and 9% increases in specificity over GAU1. In the case where $K_0 = 3$ and $(n, p) = (300, 400)$, EPDMIX had 84%, 66%, 44%, 32%, 27%, and 16% increases in specificity when the sensitivity level was fixed at 8/8, 7/8, 6/8, 5/8 and 4/8. Similarly, GAUMIX had 58%, 55%, 39%, 28%, 25%, and 16% increases in specificity.

[Put Figure 3 here.]

[Put Table 1 here. Put Table 2 here.]

The performance of full mixture regression models has also been assessed in terms of the adjusted RAND index. The results are presented in Web Appendix E, the Web-based Supplementary Materials. The aim is to demonstrate how to determine the number of components and specify the penalty coefficient simultaneously, and to illustrate the potential of the proposed method.

## 5. Numerical results on motif data

We assess the performance of the proposed method on a motif regression dataset, which was discussed in detail by Conlon et al.(2003) and explored further in Khalili et al. (2011) and Bühmann and van de Geer (2010). A motif is a candidate for a binding site of a transcription factor on the DNA, typically a 5-15 base pairs long DNA sequence. The dataset consists of the mRNA expressions of 4443 Saccharomyces cerevisiae genes and the corresponding matching

scores of 2155 candidate motifs to these genes. The main goal is to find motifs upstream of genes that undergo expression changes under a given condition via an integrative analysis of gene expressions and motif matching scores. Conlon et al.(2003) presented a motif-regression approach by formulating the problem as variable selection for linear regression. However, the gene population can be heterogeneous as genes may belong to differently regularized genetic pathways. Therefore, as suggested in our simulation studies, using a mixture regression model with more than one component might be more appropriate than using a single regression model (Gupta and Ibrahim, 2007; Khalili et al., 2011). Here, we applied our two-stage approach to the dataset, where we conducted marginal variable screening to filter out the redundant covariates, followed by fitting a full mixture regression to the selected covariates. By this dataset, we made a comparison of the approaches based on Gaussian distributions and exponential power distributions, showing that a non-Gaussian mixture model could substantially improve the analysis in terms of BIC values.

To begin with, let $n = 4443$ and $p = 2155$. We let $\mathbf{y}$ denote the logarithms of the expression levels of $n$ genes, and $\mathbf{X_{n\times p}}$ denote an $n$ by $p$ covariate matrix, the corresponding matching scores of the motifs to the genes. For motif $j, 1 \leqslant j \leqslant p$, we fitted the simple EPD mixture regression EPDMIX and the simple Gaussian mixture regression GAUMIX to the data $(\mathbf{Y}, \mathbf{x}_j)$ respectively and calculated the reciprocals of their BIC values. We arranged these reciprocals in decreasing order for the simple EPDMIX and GAUMIX respectively. These ordered values were plotted against their indices in Figure 1. The elbow points on the curves were 143 and 156 respectively. Each elbow point divided the motifs into two groups: One with higher reciprocals and the other with lower reciprocals. The last plot in Figure 1 suggests that the simple EPDMIX outperformed the simple GAUMIX in the sense that the former had the smaller BIC values than did the latter most times. In light of this fact, we adopted the simple EPDMIX as our working filter, selecting 143 variables (i.e., motifs) of higher

reciprocal BIC values. By Theorem 1, we expected these selected motifs should contain most of the true active motifs.

Finally, we fitted the EPD mixture regression EPDMIX and the Gaussian mixture regression GAUMIX to the data, taking the 143 selected motifs (denoted by $m_j$, $j = 1, 2, ..., 143$ as covariates. To take into the predictability into account in determining $K$ and $\lambda$, we randomly divided the dataset into five blocks. We deleted one block and taking the remaining as the training dataset. We ended up with five training datasets, with the size of $n = 3555$ each and the corresponding test datasets, with the size of 888 each.

After a few pilot tries, we decided to restrict $K$ and $\lambda$ in the EPD mixture regression and the Gaussian mixture regression to $1 \leqslant K \leqslant K_n = 7$ and $\lambda = (25+(t-1)10)/3555, t = 1, 2, ..., 30$. For each $K$ and $\lambda$, we calculated $\text{BIC}(K, \lambda)$ and the cross-validation (CV) function for each training dataset. Then we averaged them over five training datasets. In Web Appendix C, the Web-based Supplementary Materials, we showed that the cross-validation did not work well for this dataset. In the following, we used the BIC to determine $K$ and $\lambda$. For EPDMIX, the CV has the value of 0.2545 when $(K, \lambda) = (2, 0.03516)$, whereas for GAUMIX, when $(K, \lambda) = (2, 0.04923)$, the CV has the value of 0.2764, slightly larger than that of EPDMIX. This suggests that EPDMIX performed better than GAUMIX in fitting to the dataset. We thus focused on EPDMIX below.

[Put Figure 2 about here.]

EPDMIX gave two clusters of genes and their posterior memberships $\hat{\tau}_{ik}$ are plotted in Figure 2. EPDMIX selected 35 and 33 motifs for gene clusters 1 and 2 respectively. Cluster 1 contained these genes with large expressions, whereas Cluster 2 consisted of these genes with small expressions. The estimated parameters $(\hat{\alpha}_1, \hat{\alpha}_2)^T = (1.175, 2.534)^T$, $(\hat{\sigma}_1^2, \hat{\sigma}_2^2)^T = (0.2559, 0.0907)^T$ and $(\hat{\pi}_1, \hat{\pi}_2)^T = (0.205, 0.795)^T$, which indicate the two-components are not Gaussian. To biologically annotate the above gene clusters, for each of them, we selected a list

of genes with posterior membership probabilities no less than 90%. This gave 278 and 1394 genes for clusters 1 and 2 respectively. For each GO attribute, we compared its frequency in the gene list to its background frequency in the yeast Gene Ontology (GO) database (http://www.yeastgenome.org) on the 10th/June/2016. A web-based tool, GO term finder, was used to obtain lists of GO terms that are statistically over-represented among the genes in each list after correction for multiple hypotheses testing. For cluster 1, the significantly associated GO terms were as follows: conjugation (P-value $278\times7.38\times10^{-10}$), cell separation after cytokinesis ($278\times2.79\times10^{-8}$), reproduction ($278\times7.12\times10^{-6}$), mitotic cell cycle ($278\times7.39\times10^{-5}$), siderophore transport ($278\times7.93\times10^{-5}$), single-organism cellular process ($278\times2.1\times10^{-4}$), and cyclin-dependent protein serine ($278\times5.28\times10^{-5}$). Similarly, for gene cluster 2, the associated GO terms were: macromolecule localization ($1394\times8.68\times10^{-5}$) and intracellular organelle ($1394\times1.04\times10^{-5}$). To show the significance of the selected motifs for each gene cluster, we regressed the log-expressions of the genes in the cluster to the matching scores of the selected motifs by use of least squares. The resulting fit had an R-square of 0.584, implying that the 35 motifs might jointly account for 58.4% expression-variation in cluster 1. Note that several selected motifs were highly correlated each other (of Pearson correlation coefficients larger than 90%). So, to account for this effect, we performed a sequential ANOVA decomposition on the above fit, obtaining the P-value of extra variation explained after introducing each motif into the model given the previously introduced. This gave rise to 11 significant motifs with their P-values less than 0.01 after correction for multiple testing: $m_1$ (P-values $< 2.2\times10^{-16}$), $m_6$ ($4.18\times10^{-8}$), $m_8$ ($4.97\times10^{-14}$), $m_{26}$ ($2.3\times10^{-10}$), $m_{33}$ ($1.58\times10^{-5}$), $m_{41}$ ($4.18\times10^{-8}$), $m_{80}$ ($3.78\times10^{-13}$), $m_{121}$ ($4.1\times10^{-3}$), $m_{124}$ ($1.49\times10^{-4}$), $m_{127}$ ($1.15\times10^{-7}$), and $m_{134}$ ($1.02\times10^{-4}$).

Analogously, for gene cluster 2, the associated GO terms were: macromolecule localization (P-value $1394\times8.68\times10^{-5}$) and intracellular organelle ($1394\times1.04\times10^{-5}$). We also regressed

the log-expressions of the genes in the cluster to the matching scores of the selected motifs. The resulting fit had an R-square of 0.225, implying that the 33 motifs might jointly account for 22.5% expression-variation in cluster 2. The ANOVA decomposition gave the following list of highly significant motifs with their P-value less than 0.01 after correction for multiple testing: $m_1$ (P-value $< 2.2\times10^{-16}$), $m_6$ ($< 2.2\times10^{-16}$), $m_8$ ($< 2.2\times10^{-16}$), $m_9$ ($4.96\times10^{-3}$), $m_{26}$ ($2.99 \times 10^{-15}$), $m_{43}$ ($< 2.2 \times 10^{-16}$), $m_{54}$ ($5.29 \times 10^{-8}$), $m_{55}$ ($5.59 \times 10^{-3}$), $m_{80}$ ($< 2.2\times10^{-16}$), $m_{89}$ ($4.52\times10^{-12}$), $m_{99}$ ($2.41\times10^{-4}$), $m_{124}$ ($7.19\times10^{-3}$), and $m_{127}$ ($< 2.2\times10^{-16}$). Interestingly, the two clusters shared 7 motifs $m_j$, $j = 1, 6, 8, 26, 80, 124, 127$, implying that the corresponding transcription factors might have multiple functions by varying strengths of binding. The cluster-specific motifs for clusters 1 and 2 were $j = 33, 41, 121$ and $j = 43, 54, 55, 89, 99$ respectively.

## 6. Discussion and Conclusion

We have proposed a method to address the problem that can arise when covariates in a regression setting are not Gaussian, which gives rise to approximately mixture-distributed errors. Our contributions are four folds: We have extended the conventional Gaussian mixture model by using a more general and robust exponential power mixture distribution family for the component distributions; we have proposed a new penalty term for these models and have proved some appealing asymptotic results; with help of pre-screening, we have developed a GEM algorithm that makes model fitting computationally viable for large problems; and we have shown that BIC-based model selection works well for choosing the number of components while simultaneously performing variable selection. In particular, we have established a sure screening property for the proposed mixture-based procedure when the population is heterogeneous, filling-in a gap between the theory and practice of independence variable screening in the literature.

By simulations, we have demonstrated that the proposed non-Gaussian mixture regression

model can substantially improve the accuracy of marginal variable screening in terms of sensitivity and specificity across a range of cut-offs for screening. We have demonstrated that this holds even when the underlying model is a single high-dimensional regression. In particular, the accuracy of clustering can be dramatically improved by use of the proposed non-Gaussian mixture model when many small covariates are unselected. Our simulations have also shown that the proposed model is robust to the deviations of components from normality. The proposed procedure has been applied to the motif data, identifying two groups of genes with the associated sparse motifs. We have shown that the proposed model can improve the Gaussian mixture regression fit in terms of BIC in both the screening step and the full-model fitting step. This is not surprising as there exist model misspecification effects in both of the steps. The proposed likelihood approach can be directly extended to other penalties. The details can be found in Web Appendix H, the Web-based Supplementary Materials.

## 7. Supplementary Materials

Supplementary Materials, referenced in Sections 2, 3, 4, and 5 are available with this paper at the Biometrics website on Wiley Online Library.

### Acknowledgements

We are grateful to Professor Peter Bähmann from ETH for sharing the motif data with us.

### References

Bühmann, P. and van de Geer, S. (2010). *Statistics for High-Dimensional Data: Method, Theory and Applications*. Springer, New York.

Conlon, E.M., Liu, X., Lieb, J., and Liu, J. (2003). Integrating regulatory motif discovery and genome-wide expression analysis. *Proc. Natl. Acad. Sci. USA,* **100**, 3339-3344.

Fan, J., Feng, Y., Song, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Jour. Ameri. Stat. Assoc.*, **106**, 544-557.

Fan, J., Han, F., and Liu, H. (2014). Challenges of Big Data analysis. *National Science Review*, **1**, 293-314.

Fan, J. and Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *Ann. Stat.*, **38**, 3567-3604.

Fan, J. and Lv, J. (2008). Sure independence screening for ultra-high dimensional feature space (with discussion). *Jour. Roy. Stat. Soc. B*, **70**, 849-911.

Gupta, M. and Ibrahim, J.G. (2007). Variable selection in regression mixture modeling for the discovery of gene regulatory networks. *Jour. Ameri. Stat. Assoc.*, **102**, 867-880.

Khalili, A., Chen, J. and Lin, S. (2011). Feature selection in finite mixture of sparse normal linear models in high-dimensional feature space. *Biostatistics*, **12**, 156-172.

McLachlan, G. and Peel, D. . (2000). Finite Mixture Models. *Wiley*, New York.

Städler, N., Bühlmann, P., and van de Geer, S. (2010). $l_1$-penalization for mixture regression models. *Test*, **19**, 209-256.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B*, **58**, 267-288.

Zhang, J. (2010). A Bayesian model for biclustering with applications, *Appl Statist, JRSS C* **59**, 635656.

Zhang, J. and Liang, F. (2010). Robust clustering using exponential power mixtures. *Biometrics*, **66**,1078-1086.

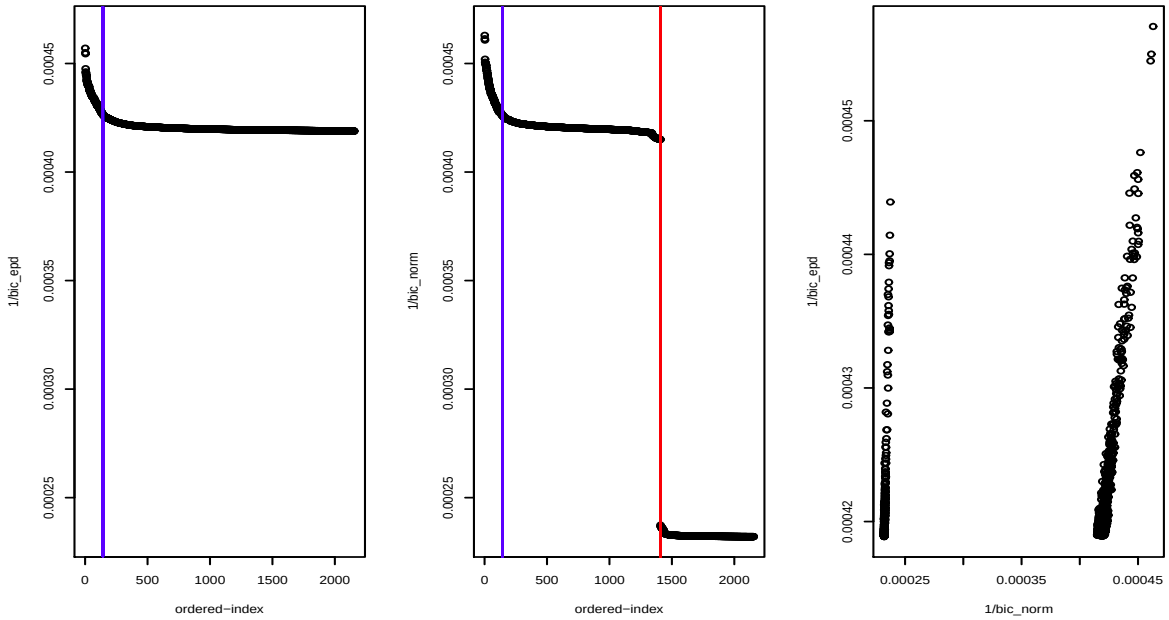


**Figure 1.** Screening plots: From the left to the right, the first two panels show the ordered reciprocal BIC values of 2155 motifs for the EPD case and the Gaussian case respectively. The vertical line in the first panel and the left vertical line in the second panel indicate the elbow points for the EPD and the Gaussian respectively. The right vertical line in the second panel points out the point after which the Gaussian fits were substantially deteriorated compared to the EPD fits. In the last panel the reciprocal BIC values of the simple EPDMIX are plotted against those of the simple GAUMIX. This figure appears in color in the electronic version of this article.

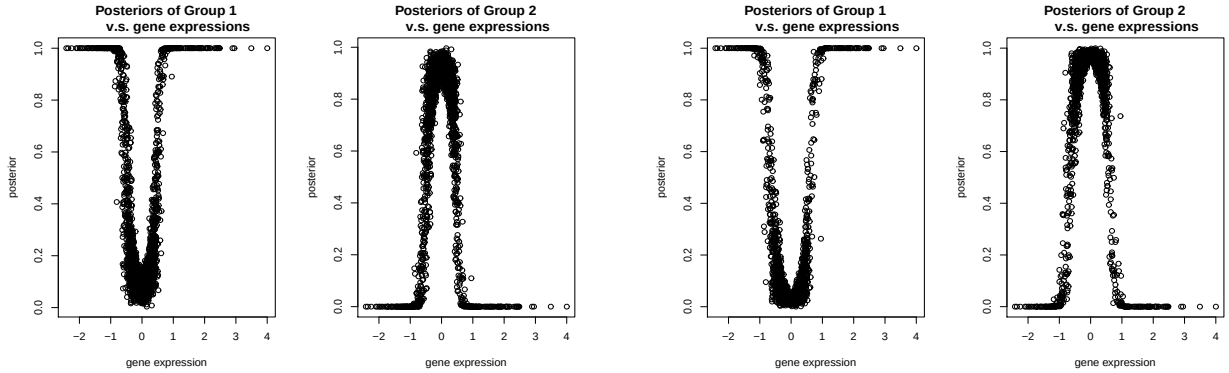


**Figure 2.** Posterior probabilities of gene memberships: the LASSO-based. The left two plots are the membership plots for the gene groups 1 and 2 derived from the LASSO-based EPDMIX while the right two plots were the membership plots for gene groups 1 and 2 derived from the LASSO-based GAUMIX.

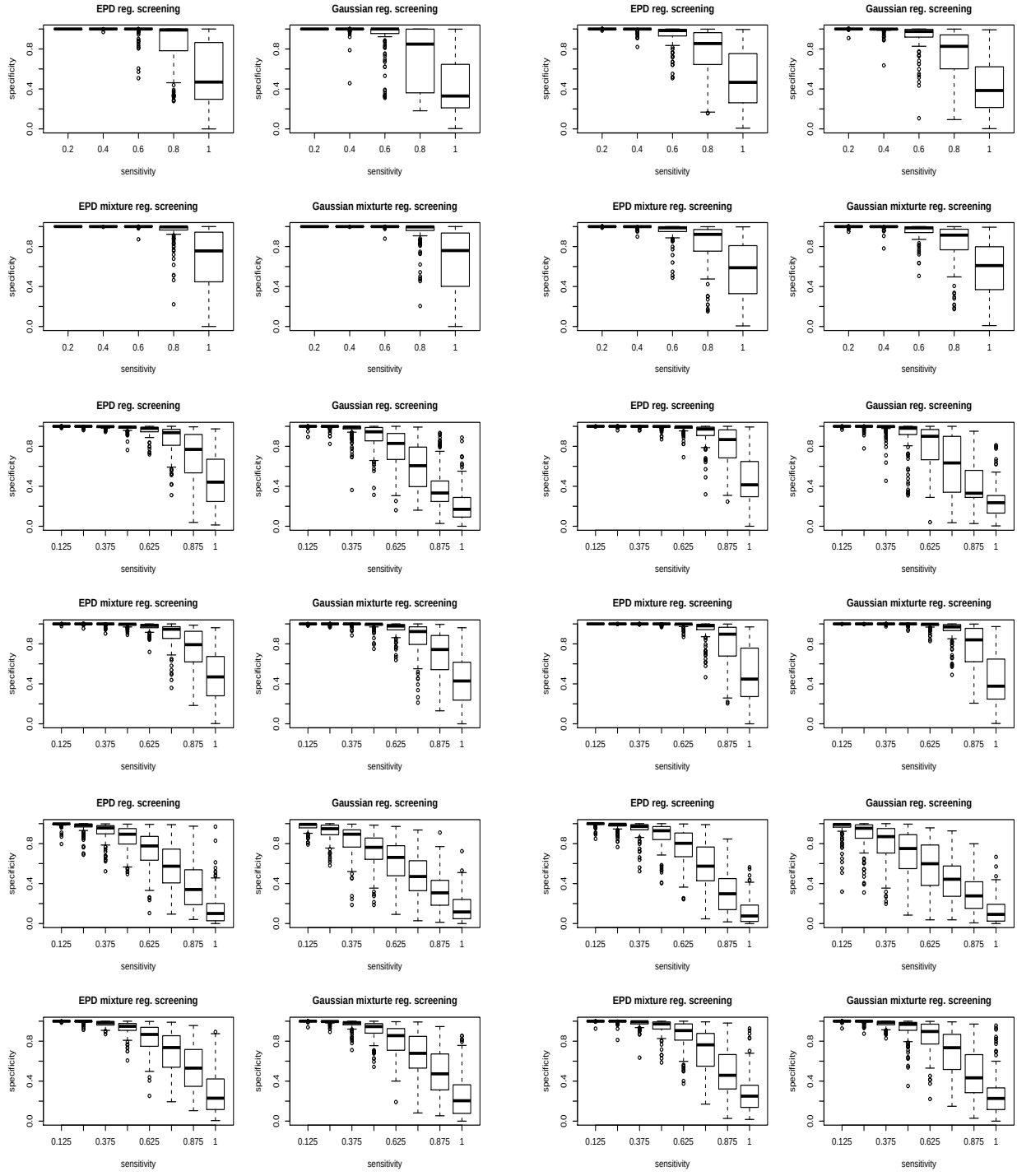


**Figure 3.** The top two rows: The box-whisker plots of the specificity for *Setting 4.1.1* when the sensitivity is fixed at the levels of $j/5$, $j = 1, 2, 3, 4, 5$. The left four plots and the right four plots are respectively for the EPD1, GAU1, EPDMIX, and GAUMIX1 when $(n, p) = (500, 600)$ and $(n, p) = (100, 2000)$. The middle two rows: The box-whisker plots of the specificity for *Setting 4.1.2(1)* when the sensitivity is fixed at the levels of $j/8$, $j = 1, 2, ..., 8$. The left four plots and the right four plots are for $(n, p) = (300, 400)$ and $(500, 600)$ respectively. The bottom two rows are for *Setting 4.1.2(2) with* $(n, p) = (300, 400)$ *and* $(500, 600)$.

**Table 1**

*Percentage increase of average specificity compared to the GAU1 in variable screening*

| Setting 4.1.1: single component | | | | | |
|---|---|---|---|---|---|
| Sensitivity | 5/5 | 4/5 | 3/5 | 2/5 | 1/5 |
| Percentage increase of ave. spe. (%) | | | | | |
| $(n, p) = (500, 600)$ | | | | | |
| GAU1 | 0 | 0 | 0 | 0 | 0 |
| EPD1 | 33 | 17 | 7 | 1 | 0 |
| EPDMIX | 64 | 31 | 9 | 1 | 0 |
| GAUMIX | 62 | 30 | 9 | 1 | 0 |
| $(n, p) = (100, 2000)$ | | | | | |
| GAU1 | 0 | 0 | 0 | 0 | 0 |
| EPD1 | 13 | 4 | 2 | 0 | 0 |
| EPDMIX | 29 | 13 | 4 | 1 | 0 |
| GAUMIX | 30 | 14 | 3 | 0 | 0 |

**Table 2**

*Percentage increase of average specificity compared to the GAU1 in variable screening*

| Setting 4.1.2: multiple components | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Sensitivity | 8/8 | 7/8 | 6/8 | 5/8 | 4/8 | 3/8 | 2/8 | 1/8 |
| Percentage increase of ave. spe. (%) | | | | | | | | |
| Two components: $(n, p) = (300, 400)$ | | | | | | | | |
| GAU1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| EPD1 | 119 | 90 | 46 | 24 | 9 | 3 | 0 | 0 |
| EPDMIX | 128 | 97 | 51 | 25 | 10 | 3 | 0 | 0 |
| GAUMIX | 110 | 85 | 44 | 23 | 10 | 3 | 1 | 0 |
| Two components: $(n, p) = (500, 600)$ | | | | | | | | |
| GAU1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| EPD1 | 98 | 93 | 49 | 25 | 9 | 2 | 0 | 0 |
| EPDMIX | 112 | 95 | 51 | 26 | 10 | 2 | 0 | 0 |
| GAUMIX | 88 | 84 | 49 | 26 | 10 | 2 | 1 | 0 |
| Three components:$(n, p) = (300, 400)$ | | | | | | | | |
| GAU1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| EPD1 | -4 | 15 | 19 | 17 | 17 | 10 | 6 | 2 |
| EPDMIX | 84 | 66 | 44 | 32 | 27 | 16 | 9 | 3 |
| GAUMIX | 58 | 55 | 39 | 28 | 25 | 16 | 8 | 3 |
| Three components:$(n, p) = (500, 600)$ | | | | | | | | |
| GAU1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| EPD1 | -11 | 7 | 34 | 33 | 26 | 19 | 10 | 4 |
| EPDMIX | 111 | 66 | 63 | 50 | 34 | 23 | 12 | 5 |
| GAUMIX | 100 | 58 | 58 | 47 | 32 | 22 | 12 | 5 |