Unsupervised Liu-type shrinkage estimators for mixture of regression models

Elsayed Ghanem¹,², Armin Hatefi¹ and Hamid Usefi¹

Abstract
The mixture of probabilistic regression models is one of the most common techniques to incorporate the information of covariates into learning of the population heterogeneity. Despite its flexibility, unreliable estimates can occur due to multicollinearity among covariates. In this paper, we develop Liu-type shrinkage methods through an unsupervised learning approach to estimate the model coefficients in the presence of multicollinearity. We evaluate the performance of our proposed methods via classification and stochastic versions of the expectation-maximization algorithm. We show using numerical simulations that the proposed methods outperform their Ridge and maximum likelihood counterparts. Finally, we apply our methods to analyze the bone mineral data of women aged 50 and older.

Keywords
Bone mineral data, expectation-maximization algorithm, Liu-type penalty, maximum likelihood, mixture models, multicollinearity, ridge penalty

1 Introduction
As a bone metabolic disease, osteoporosis is characterized when the mineral density of bone tissues decreases significantly, leading to various major health problems such as skeletal fragility and osteoporotic fractures. These can occur in different body areas, including the hip, spine, and femur.¹² Bone mineral density (BMD) is considered one of the most reliable predictors in determining osteoporosis status.³ For example, approximately every 1 out of 3 women and 1 out of 5 men aged 50 and older have osteoporosis and its fractures.⁴ Osteoporosis typically occurs without any significant symptoms. In South Korea, for example, almost 75% of patients are unaware of their osteoporosis problem.⁵ BMD score of an individual improves until age 30 and declines as the individual ages. As the aged population is growing, it is essential to study osteoporosis to plan the well-being and life quality of the aged groups of the communities.

BMD values are measured through dual-energy X-ray absorptiometry imaging and require a costly and time-consuming procedure. Unlike BMD scores, clinicians have access to various easily attainable explanatory variables about patients, including BMI, weight, age, and test results from previous years. Linear regression models are well-known statistical tools to investigate the impact of a set of covariates (e.g., patients’ characteristics) on a response variable (e.g., BMD measurement). The least squares (LSs) method is a common method for estimating the regression model coefficients; however, the LS method can lead to extremely unreliable and misleading results in multicollinearity when the covariates are linearly dependent. As a shrinkage method, Ridge regression is a common solution to the multicollinearity problem. A small ridge parameter is not enough to handle the ill-conditioned design matrix when the problem is severe. On the other hand, large values of the ridge parameter render more biases in the estimation. The Liu-type (LT) shrinkage method was proposed in the work of Liu⁶ to address the multicollinearity and was extended in the work of Duran et al.⁷ to the semi-parametric...
regressions. Later on, the Stein-rule LT estimators for elliptical regression models was developed in the work of Arashi et al. and properties of rank-based samples was used in the work of Pearce and Hatefi to improve the LT shrinkage estimators for linear and logistic regressions.

Finite mixture models (FMMs) are probabilistic model-based tools to analyze heterogeneous populations. The maximum likelihood (ML) method is one of the most popular techniques for estimating the parameters of the FMMs. A part of the popularity comes from the expectation-maximization (EM) algorithm that enables computing the ML estimates of the FMMs. The EM algorithm decomposes the ML estimation procedure into E- and M-steps and iteratively estimates the parameters of the FMMs. Stochastic EM (SEM) algorithm was developed as a stochastic teacher of the probabilistic EM algorithm for mixture models. Also, Celeux and Govaert proposed a classification EM (CEM) algorithm in which a classification step is implemented in the iteration for maximization based on classified data. The mixture of regression models incorporates the properties of the FMMs into linear regression models. The ML and EM algorithms to fit a mixture of regression models was investigated in the work of Faria and Soromenho, Hawkins et al., Jones and McLachlan. Alternatively, Bayesian estimation techniques have been developed for finite mixture regression models, as outlined in the work of Frühwirth-Schnatter. One of the key challenges in FMMs, including mixture regression models, is identifiability. This challenge has been discussed in depth by researchers such as Hennig and Ho and Nguyen. FMMs have found many applications in the core of statistical sciences, such as times series data, sampling methods and censored data. For more details about the theory and applications of the FMMs, see and references therein.

Despite the flexibility of the mixture of linear regressions, the ML estimates of the mixture coefficients become unreliable in the presence of multicollinearity. Recently, a shrinkage estimation method for the mixture of logistic regressions was proposed in Ghanem et al. In this paper, we develop the Ridge and LT shrinkage estimation methods for heterogeneous populations. To do so, we focus on the mixture of linear regression models where each homogeneous sub-population is treated by multiple regression to investigate the relationship between explanatory and response variables in multicollinearity. The population estimation is designed using an unsupervised statistical learning approach where the observations’ component membership is unknown and estimated iteratively. To estimate the coefficients of the component regressions and mixing proportions, we developed EM, CEM and SEM algorithms to incorporate the Ridge and LT penalized log-likelihood in each iteration of the algorithms. In addition, we proposed the singular value decomposition to translate the LT estimation method of the parameters of the population to iterative canonical regression analysis in each component. From the orthogonal properties of the eigenvalues in the ill-conditioned component design matrices, we find the optimum tuning parameter of the LT estimation method. We used the root mean square errors and root mean square prediction errors using K-fold cross-validation to compare the estimation and prediction performances of the proposed methods. Through extensive numerical studies, we show that the LT shrinkage method’s classification and SEM algorithms outperform their counterparts and provide more reliable coefficient estimates for component regressions. We finally apply the developed methods to analyze the bone disorder status of women aged 50 and older.

The outline of the paper is as follows: Section 2 develops the shrinkage methods and their EM algorithms. Section 3 evaluates the performance of the methods by simulation studies. The methods are applied for analysis of bone data in Section 4. Section 5 presents summary and concluding remarks.

2 Statistical methods

Let \( \mathbf{x}_i^T = (x_{i,1}, \ldots, x_{i,p}) \) be the vector of \( p \) explanatory variables for the \( i \)th subject in a random sample of size \( n \). Let \( \mathbf{y} = (y_1, \ldots, y_n) \) and \( \mathbf{X} = (\mathbf{x}_1^T, \ldots, \mathbf{x}_n^T)^T \) denote the response vector and the full column rank \((n \times p)\) design matrix with \( \text{rank}(\mathbf{X}) = p < n \). The regression model \( y_i = \mathbf{x}_i^T \beta + \epsilon_i \) for \( i = 1, \ldots, n \) is one of the most common statistical methods to study the relationship between response variable and the set of the explanatory variables.

The mixture of regression models is a generalization of the regression model when the underlying population comprises of heterogeneous subpopulations. The mixture of regression models is defined by

\[
y_i = \begin{cases} 
\mathbf{x}_i^T \beta_1 + \epsilon_{i1} & \text{with probability } \pi_1 \\
\vdots & \\
\mathbf{x}_i^T \beta_J + \epsilon_{ij} & \text{with probability } \pi_j
\end{cases}
\]

where \( \beta_j = (\beta_{j,1}, \ldots, \beta_{j,p}) \) represents the coefficients of \( p \) predictors in the \( j \)th component regression for \( j = 1, \ldots, J \) and \( \pi = (\pi_1, \ldots, \pi_J) \) denotes the vector of mixing proportions with \( \sum_{j=1}^M \pi_j = 1 \) and \( 0 < \pi_j < 1 \). Also \( \epsilon_{ij} \) are independent normal random errors from each component of the mixture; that is \( \epsilon_{ij} \sim N(0, \sigma_j^2) \). Although we assume that the number of
components \( J \) in the mixture model (1) is known throughout the paper, the component memberships of the observations are unknown and should be estimated in an unsupervised approach. Let \( \beta = (\beta_1, \ldots, \beta_J) \). Let \( \theta_j = (\beta_j, \sigma_j^2) \) denote the parameters of the \( j \)th component. Thus, we represent the vector of all unknown parameters of the mixture (1) with \( \Psi = (\pi, \theta_1, \ldots, \theta_J) \).

From regression model (1), the log-likelihood function of \( \Psi \) can be written as

\[
\ell(\Psi) = \sum_{i=1}^{n} \log \left( \sum_{j=1}^{J} \pi_j \phi_j(x_i^\top \beta_j, \sigma_j^2) \right)
\]  
(2)

where \( \phi_j(x_i^\top \beta_j, \sigma_j^2) \) represents the pdf of the univariate normal distribution with mean \( x_i^\top \beta_j \) and variance \( \sigma_j^2 \). We need to maximize (2) in order to obtain the ML estimate of \( \Psi \). The gradient of (2) is not tractable with respect to component parameters \( \theta_j, j = 1, \ldots, J \). We view \((X, y)\) as incomplete data and apply the EM algorithm of Dempster et al.\(^{10}\) using a complete data to find \( \hat{\Psi}_{ML} \). For each subject \((x_i, y_i)\), we introduce latent variables \( Z_i = (Z_{i1}, \ldots, Z_{iJ}) \) for \( i = 1, \ldots, n \) as

\[
Z_{ij} = \begin{cases} 
1 & \text{if the } i \text{th subject comes from the } j \text{th component,} \\
0 & \text{o.w.,}
\end{cases}
\]

where \( Z_i \sim \text{Multi}(1, \pi_1, \ldots, \pi_M) \). From the marginal distribution of latent variables, the conditional distribution of \( Z_i | y_i \) is given by

\[
f(z_i | y_i) = \prod_{j=1}^{J} \left\{ \frac{\pi_j \phi_j(x_i^\top \beta_j, \sigma_j^2)}{\sum_{j=1}^{J} \pi_j \phi_j(x_i^\top \beta_j, \sigma_j^2)} \right\}^{z_{ij}}
\]  
(3)

From the above, it is easily seen that \( Z_i | y_i \sim \text{Multi}(1, \tau_{i1}(\Psi), \ldots, \tau_{iJ}(\Psi)) \) where

\[
\tau_{ij}(\Psi) = \frac{\pi_j \phi_j(x_i^\top \beta_j, \sigma_j^2)}{\sum_{j=1}^{J} \pi_j \phi_j(x_i^\top \beta_j, \sigma_j^2)}
\]  
(4)

Let \((X, y, Z)\) denote the complete data. Thus, the complete log-likelihood function of \( \Psi \) is given by

\[
\ell_c(\Psi) = \sum_{i=1}^{n} \sum_{j=1}^{J} z_{ij} \log(\pi_j) + \sum_{i=1}^{n} \sum_{j=1}^{J} z_{ij} \log(\phi_j(x_i^\top \beta_j, \sigma_j^2))
\]  
(5)

### 2.1 ML estimation method

The EM algorithm is a standard method to find the ML estimates of the mixture model parameters. The EM algorithm employs the latent variables on top of the observed data and decomposes the estimation procedure into an iterative expectation (E) and maximization (M) steps. As an iterative method, EM algorithm begins with an initial value. Let \( \Psi^{(0)} = (\pi^{(0)}, \theta_1^{(0)}, \ldots, \theta_J^{(0)}) \) and \( \Psi^{(r)} \) denote the initial vector and the estimate in the \( r \)th iteration of the EM algorithm, respectively.

On the \((r+1)\)th iteration, we require to compute the conditional expectation of (5) in the E-step. The \( Q(\Psi, \Psi^{(r)}) \) replaces the latent variables by their conditional expectations as

\[
Q(\Psi, \Psi^{(r)}) = Q_1(\pi, \Psi^{(r)}) + Q_2(\theta, \Psi^{(r)})
\]

where

\[
Q_1(\pi, \Psi^{(r)}) = \sum_{i=1}^{n} \sum_{j=1}^{J} \tau_{ij}(\Psi^{(r)}) \log(\pi_j)
\]  
(6)

and

\[
Q_2(\theta, \Psi^{(r)}) = \sum_{i=1}^{n} \sum_{j=1}^{J} \tau_{ij}(\Psi^{(r)}) \log \left\{ \phi_j(x_i^\top \beta_j, \sigma_j^2) \right\}
\]  
(7)
where $\tau_j(\Psi^{(r)})$ is obtained by (4). In the M-step, we maximize $Q(\Psi, \Psi^{(r)})$ with respect to $\pi$ and $\theta$ to update $\Psi^{(r+1)}$. One can update $\pi^{(r+1)}$ by maximizing $Q_1(\pi, \Psi^{(r)})$ subject to $\sum_{j=1}^{J} \pi_j = 1$ as follows

$$
\hat{\pi}^{(r+1)}_j = \frac{\sum_{i=1}^{n} \tau_{ij}(\Psi^{(r)})}{n}; \quad j = 1, \ldots, J - 1
$$

(8)

The maximization of $Q_2(\theta, \Psi^{(r)})$ can be reformulated by the weighted least squares (WLS) method as

$$
\hat{\beta}^{(r+1)}_j = \arg\min_{\beta_j} (y - X\beta_j)^TW_j(y - X\beta_j)/n
$$

(9)

where $W_j$ is $n \times n$ diagonal matrix with diagonal elements $(\tau_{1j}(\Psi^{(r)}), \ldots, \tau_{nj}(\Psi^{(r)}))$ for all $j = 1, \ldots, J$. One can easily update $\hat{\beta}^{(r+1)}_j$ as the solution to (9) by

$$
\hat{\beta}^{(r+1)}_j = (X^TW_jX)^{-1}X^TW_jy, \quad j = 1, \ldots, J
$$

(10)

From (9) and following, we then update $\hat{\sigma}_{j}^{2(r+1)}$ as follows

$$
\hat{\sigma}_{j}^{2(r+1)} = \frac{(y - X\hat{\beta}^{(r+1)}_j)^TW_j(y - X\hat{\beta}^{(r+1)}_j)}{\sum_{i=1}^{n} \tau_{ij}(\Psi^{(r)})}, \quad j = 1, \ldots, J
$$

(11)

To find $\hat{\Psi}_{ML}$, we iteratively alternate the E- and M- steps of the EM algorithm until the stopping criterion $|\ell(\Psi^{(r+1)}) - \ell(\Psi^{(r)})| \leq \varepsilon$ becomes negligible.

**CEM Algorithm:** In the above EM algorithm, we use information from all observations (as membership probabilities) in each iteration to estimate the parameters of the mixture model. Following Celeux and Govaert, we shall employ the classification version of EM algorithm (CEM) to estimate $\Psi$. The CEM algorithm incorporates a classification (C) step between E and M steps so that the component parameters of the mixture are updated in M-step using the classified complete data log-likelihood function.

The E-step here is identical to the E-step of the EM algorithm. In C-step, the observations are then assigned to $J$ mutually exclusive partitions corresponding to the $J$ components of mixture model. Let $P_1^{(r+1)}, \ldots, P_J^{(r+1)}$ denote the partition in the $(r+1)$th iteration. Each subject $(x_i, y_i)$ is assigned to partition $P_h^{(r+1)}$ when

$$
\tau_{ih}(\Psi^{(r)}) = \arg\max_j \tau_{ij}(\Psi^{(r)})
$$

Note when the maximum weight is not unique, the tie is broken at random. Also the CEM algorithm is stopped and $\Psi^{(r)}$ is returned when a partition becomes empty or has only one observation.

In the M-step, we maximize conditional expectation function using $P_1^{(r+1)}$. From (6), the mixing proportion is updated by

$$
\hat{\pi}^{(r+1)}_j = \frac{n_j}{n}; \quad j = 1, \ldots, J
$$

(12)

where $n_j$ is the number of observations allocated to partition $P_j^{(r+1)}$. Applying the WLS to each partition $P_j^{(r+1)}$, we can update the parameters of the $j$th component ($j = 1, \ldots, J$) by

$$
\hat{\beta}^{(r+1)}_j = (X^TW_jX)^{-1}X^TW_jy_j
$$

(13)

$$
\hat{\sigma}_{j}^{2(r+1)} = \frac{(y_j - X\hat{\beta}^{(r+1)}_j)^TW_j(y_j - X\hat{\beta}^{(r+1)}_j)}{\sum_{i=1}^{n} \tau_{ij}(\Psi^{(r)})}
$$

(14)
where $X_i$ and $y_i$ represent, respectively, $(n_i \times p)$ design matrix and vector of responses corresponding to $P_i^{(r+1)}$. Also $W_i^{(r)}$ is the diagonal weight matrix of size $n_i$ with entries $(r_i, \Psi_i^{(r)}), \ldots, r_{n_i}, \Psi_i^{(r)})$. Finally, we alternate repeatedly the E-, C- and M-steps until $|\ell(\Psi_i^{(r+1)}) - \ell(\Psi_i^{(r)})| < \epsilon$.

**SEM Algorithm:** One can also apply the stochastic version of the EM algorithm\(^{11}\) to fit the mixture of regression models. The SEM algorithm implements a stochastic version of the C-step between E- and M-steps. Although the E- and M-steps are identical to the CEM algorithm, the SEM simulates the component membership of each observation using conditional distribution of the latent variable given incomplete data. On the $(r+1)$th iteration, the S-step simulates a random allocation for each observation using one draw out of $J$ components as

$$Z_i^* = (Z_{i1}^*, \ldots, Z_{ij}^*) \sim \text{Multi}(1, \tau_{i1}, \Psi_i^{(r)}), \ldots, \tau_{ij}, \Psi_i^{(r)}) \quad (i = 1, \ldots, n)$$

Then the $(x_i, y_i)$ is classified to partition $P_i^{(r+1)}$ if $Z_i^* = 1; i = 1, \ldots, j = 1, \ldots, J$. Using the stochastic partitions, we update the mixture parameters from (12), (13) and (14) in M-step. From\(^{11,14}\) point-wise convergence in the SEM algorithm is not guaranteed. The algorithm resembles a Markov chain where at stationary state fluctuates around the ML estimate. Hence, we alternate the E-, S- and M-steps until either the criterion $|\ell(\Psi_i^{(r+1)}) - \ell(\Psi_i^{(r)})| < \epsilon$ satisfies or the chain reaches a prespecified maximum number of iterations that is fixed for all the algorithms for fair comparison.

### 2.2 Ridge estimation method

The ML method is a standard tool to estimate $\Psi$; however, the ML estimates are dramatically influenced by multicollinearity where the covariates are linearly dependent. Ridge method\(^{27}\) is one of the most common methods to encounter with the challenges of the LS method. The ridge estimate for (1) can be obtained as a solution to the penalized log-likelihood function given by

$$\ell^R(\Psi) = \ell(\Psi) - k\beta^T \beta / 2$$

where $\ell(\Psi)$ comes from (2) and $k > 0$ is the ridge parameter. In a similar vein to Subsection 2.1, for each observation $(x_i, y_i), i = 1, \ldots, n$, we first introduce $J$ dimensional latent vector $Z_i = (Z_{i1}, \ldots, Z_{ij})$. We then develop an EM algorithm to maximize the complete ridge log-likelihood function and obtain $\hat{\Psi}_R$.

The E-step of the ridge EM algorithm is identical to the E-step of Subsection 2.1. In the M-step, the mixing proportion are updated from (8). To update the coefficients of the component regressions, we require to maximize $Q_2(\theta, \Psi^{(r)})$ subject to the ridge penalty within each component as

$$Q^R_2(\theta, \Psi^{(r)}) = Q_2(\theta, \Psi^{(r)}) - k_j \beta_j^T \beta_j / 2$$

where $Q_2(\theta, \Psi^{(r)})$ is from (7) and $k_j$ is the ridge parameter in the $j$th component. Like ML method, one can re-write the maximization of $Q^R_2(\theta, \Psi^{(r)})$ as a WLS subject to the ridge penalty by

$$\hat{\beta}^{(r+1)}_{Rj} = \arg \min_{\beta_j} \left( y - X\beta_j \right)^T W_j \left( y - X\beta_j \right) + k_j \beta_j^T \beta_j / 2$$

where $W_j$ is $n \times n$ diagonal matrix with elements $(r_i, \Psi^{(r)}), \ldots, r_{n_j}, \Psi^{(r)})$ obtained from (4). Applying (16), $\hat{\beta}_j^{(r+1)}; j = 1, \ldots, J$ is updated by

$$\hat{\beta}_j^{(r+1)} = \left( X^T W_j X + k_j I \right)^{-1} X^T W_j y$$

**Lemma 1.** Under the assumptions of mixture of regression models (1), suppose $\lambda_1, \ldots, \lambda_p$ and $u_1, \ldots, u_p$ are the eigenvalues and orthonormal eigenvectors of $X^T W_j X$ where $W_j$ is $n \times n$ diagonal matrix with entries $(r_i, \Psi^{(r)}), \ldots, r_{n_j}, \Psi^{(r)})$ under ridge EM algorithm. Let $\Lambda_j = \text{diag}(\lambda_1, \ldots, \lambda_p)$ and $U_j = [u_1, \ldots, u_p]$. Then the canonical weighted ridge estimator in each component regression is given by

$$\hat{a}_{Rj} = (\Lambda_j + k_j I)^{-1} \Lambda_j^{1/2} U_j^T W_j^{1/2} y$$

and

$$\hat{\beta}_{Rj} = U_j \hat{a}_{Rj}$$
with $V = [v_1, \ldots, v_p]_j$ where $v_1, \ldots, v_p$ are the orthonormal eigenvectors of $W_j^{1/2}X_j^T W_j^{1/2}$.

From (16), the variance term can be updated by

$$\sigma_{Rj}^{2(r+1)} = \frac{(y - X_j^T \hat{\beta}_{Rj}^{(r+1)})^T W_j^{(r)} (y - X_j^T \hat{\beta}_{Rj}^{(r+1)})}{\sum_{i=1}^n r_j(y_i^T \Psi^{(r)})}$$

There are various methods available in the literature for the estimation of $k_j$. Following Hoerl et al., we estimate the parameter by $\hat{k}_j = p\hat{\sigma}_{Rj}^2/\hat{\sigma}_{ML}^2$, where $\hat{\sigma}_{Rj}^2$ and $\hat{\sigma}_{ML}^2$ are calculated from (11) and (10), respectively. The E- and M-steps are repeatedly computed until $|e_R(\Psi^{(r+1)}) - e_R(\Psi^{(r)})| < \epsilon$.

**Ridge CEM Algorithm:** One can also apply the CEM algorithm to find $\hat{\Psi}_R$. Similar to the CEM algorithm in Subsection 2.1, we require to accommodate a C-step between E- and M-steps in the ridge EM algorithm. Here the E-step remains the same as before. Similar to the C-step of ML method, we classify the observations to partitions $P^{(r+1)} = (P_1^{(r+1)}, \ldots, P_J^{(r+1)})$ based on the maximum probability of memberships; that is

$$P_j^{(r+1)} = \{ (x_i, y_i); r_j(\Psi^{(r)}) = \arg \max_h r_{jh}(\Psi^{(r)}) \}, \quad \forall j = 1, \ldots, J$$

Based on $P_j^{(r+1)}$, we use (12) to update the mixing proportions of the mixture. The ridge parameters are estimated similar to the ridge EM algorithm. We apply (16) to each partition $P_j^{(r+1)}$ and update the coefficients and variance term of each component regression by

$$\hat{\beta}_{Rj}^{(r+1)} = \left( X_j^T W_j X_j + k_j \right)^{-1} X_j^T W_j y_j$$

$$\hat{\sigma}_{Rj}^{2(r+1)} = \frac{(y_j - X_j^T \hat{\beta}_{Rj}^{(r+1)})^T W_j^{(r)} (y_j - X_j^T \hat{\beta}_{Rj}^{(r+1)})}{\sum_{i=1}^n r_j(y_i^T \Psi^{(r)})}$$

where $X_j$ is $(n_j \times p)$ design matrix and $y_j$ is vector of responses from observations classified to $P_j^{(r+1)}$. $W_j^{(r)}$ is the diagonal weight matrix with entries $(r_j(\Psi^{(r)}), \ldots, r_{nj}(\Psi^{(r)}))$ from (4). Finally, the E-, C-, and M-steps under ridge estimation procedure are alternated until convergence criterion is satisfied.

**Ridge SEM Algorithm:** Ridge estimation method can be implemented via the SEM algorithm. Like SEM of ML method, the S-step determines stochastically the component membership of observations under ridge method by $Z_i^* = (Z_{i1}^*, \ldots, Z_{ij}^*) \sim \text{Multi}(1, r_j(\Psi^{(r)}), \ldots, r_{nj}(\Psi^{(r)})), i = 1, \ldots, n$ and updates $P^{(r+1)} = (P_1^{(r+1)}, \ldots, P_J^{(r+1)})$ such that $P_j^{(r+1)} = \{ (x_i, y_i); Z_{ij}^* = 1 \}, \forall j = 1, \ldots, J$. Based on this stochastic partition of S-step, we update the mixture parameters by (12), (19) and (20).

**Lemma 2.** Under the assumptions of mixture of regression models (1) with component regression models $y = X \beta_j + \epsilon$ based on $n_j$ observations and rank($X_j$) = $p$, suppose $\lambda_1, \ldots, \lambda_p$ and $u_1, \ldots, u_p$ are the eigenvalues and orthonormal eigenvectors of $X_j^T W_j X_j$ where $W_j$ is a $n_j \times n_j$ diagonal matrix with entries $(r_j(\Psi^{(r)}), \ldots, r_{nj}(\Psi^{(r)}))$ under the ridge CEM or ridge SEM algorithm. Let $\Lambda_j = \text{diag}(\lambda_1, \ldots, \lambda_p)$ and $U_j = [u_1, \ldots, u_p]$. Then the canonical weighted ridge estimator in each component regression is given by

$$\hat{\alpha}_{Rj} = (\Lambda_j + k \frac{\lambda_j^2}{\lambda_j^2})^{-1} \Lambda_j^{1/2} V_1^T W_1^{1/2} y$$

and

$$\hat{\beta}_{Rj} = U_j \hat{\alpha}_{Rj}$$

with $V_1 = [v_1, \ldots, v_p]_j$ where $v_1, \ldots, v_p$ are the orthonormal eigenvectors of $W_j^{1/2}X_j^T W_j^{1/2}$.

Finally, the E-, S-, and M-steps are iterated until the stopping rule is satisfied or the algorithm reaches a prespecified maximum number of iterations.
LT estimation method

When the design matrix is severely ill-conditioned, adding small values to the diagonal elements by the ridge estimator may be insufficient in addressing the issue. On the other hand, increasing the ridge parameter may result in a more considerable bias in the ridge estimation method. The LT shrinkage method for the multicollinearity in estimating the regression parameters was proposed in Liu. Like the ridge method, the LT method optimizes the estimating equation subject to the LT penalty to control the multicollinearity. The LT penalty is given by

$$\left(-\frac{d}{k^{1/2}}\right) \hat{\beta} = k^{1/2} \hat{\beta} + \epsilon'$$

where $\hat{\beta}$ can be any estimator of coefficients and $d \in \mathbb{R}$ and $k > 0$ are two parameters of the LT method. We develop the LT shrinkage method in estimating the unknown parameters of the mixture model (1). We shall find the LT estimate of $\Psi$ by maximizing the log-likelihood function (2) subject to the LT penalty.

Similar to Section 2.1, the penalized log-likelihood function based on observed data is not tractable with respect to the component parameters. We develop the LT estimation procedure through an unsupervised approach and use the EM algorithm to estimate $\Psi$. We again require latent vectors $Z_i = (Z_{i1}, \ldots, Z_{ij})$ to represent the component membership of the $i$th observation $(x_i, y_i); i = 1, \ldots, n$. Let $(X, y, Z)$ denote the complete data. Then EM algorithm under the LT method proceeds as follows.

In $(r + 1)$th iteration, the E-step remains identical to the E-step of the ML method. The mixing proportions under the LT method are updated by (8). We require to maximize $Q_2(\theta; \Psi^{(r)})$ from (7) under the LT penalty within each component to estimate the component parameters. The LT penalized log-likelihood function can be written as a WLS constrained on the LT penalty as

$$\hat{\beta}_{LT}^{(r+1)} = \arg \min_{\beta_j} \left( y - X_j \beta_j \right)^T W_j \left( y - X_j \beta_j \right) + \left[ \left( -\frac{d}{k_j^{1/2}} \right) \hat{\beta}_j - k_j^{1/2} \beta_j \right]$$

where $\hat{\beta}_j$ can be any estimate and $W_j$ is a weight diagonal matrix with elements $(\tau_{rj}(\Psi^{(r)}), \ldots, \tau_{nj}(\Psi^{(r)}); j = 1, \ldots, J$. From (22), the coefficients and variance terms in each component regression are updated by

$$\hat{\beta}_{LT}^{(r+1)} = (X^T W_j X + k_j) \left( X^T W_j y - d_j \hat{\beta}_j \right)$$

$$\hat{\sigma}_{LT}^{2(r+1)} = \frac{(y - X \hat{\beta}_{LT}^{(r+1)})^T W_j (y - X \hat{\beta}_{LT}^{(r+1)})}{\sum_{i=1}^n \tau_i(\Psi^{(r)})}$$

Equations (23) and (24) require the estimates of the LT parameters $(k_j, d_j)$ for each component regression. From Liu, we can estimate $k_j$ in the $j$th component by $\hat{k}_{LT} = (\lambda_{1j} - 100\lambda_{2j})/99$ where $\lambda_{1j}$ and $\lambda_{2j}$ are the maximum and the minimum eigenvalues of $X^T W_j X$ on the $(r+1)$th iteration of the EM algorithm.

**Lemma 3.** Under the assumptions of Lemma (1), the canonical LT estimator in the $j$th component regression $j = 1, \ldots, J$ under EM algorithm is given by

$$\hat{\alpha}_{LT} = \left( A_j + k_j \right)^{-1} \left( A_j^{1/2} V_j^{1/2} y - d_j \hat{\alpha}_j \right)$$

and

$$\hat{\beta}_{LT} = U_j \hat{\alpha}_{LT}$$

where $\hat{\alpha}_j$ is the canonical estimate of $\beta_j$ and $V_j = [v_{1j}, \ldots, v_{pj}]$ with $v_{1j}, \ldots, v_{pj}$ are the orthonormal eigenvectors of $W_j^{1/2} X^T X W_j^{1/2}$.

Following Liu and Lemma 3, the optimal $d_j$ can be obtained by the next lemma within each component of the mixture of regression models.
Lemma 4. Under the assumptions of Lemma 1 and for each \( k_j > 0 \), we have

(i) when \( \widehat{\alpha}_j = \alpha_{ML,j} \), then

\[
d_j = \sum_{m=1}^{p_j} \left( \frac{(\sigma_j^2 - k_j \alpha_m^2)}{(\lambda_m k_j + \sigma_j^2)} \right) / \sum_{m=1}^{p_j} \left( \frac{\lambda_m \alpha_m^2 + \sigma_j^2}{\lambda_m k_j + \sigma_j^2} \right)
\]

(ii) when \( \widehat{\alpha}_j = \alpha_{R,j} \), then

\[
d_j = \sum_{m=1}^{p_j} \left( \frac{\lambda_m (\sigma_j^2 - k_j \alpha_m^2)}{(\lambda_m^2 k_j + \sigma_j^2)} \right) / \sum_{m=1}^{p_j} \left( \frac{\lambda_m \alpha_m^2 + \sigma_j^2}{\lambda_m^2 k_j + \sigma_j^2} \right)
\]

minimizes the MSE(\( \widehat{\alpha}_{LT,j} \)) within each component (1) in the EM algorithm of the LT method.

Although Lemma 4 paves the path in estimating the optimal LT parameter \( d_j \) within each component regression, the optimal value still depends on the unknown quantities including \( \sigma_j \), \( k_j \), \( \alpha_j \) and \( \lambda_{mj} \) for \( m = 1, \ldots, p \) and \( j = 1, \ldots, J \). From Lemma 4, we propose a practical approach where \( d_j \), \( j = 1, \ldots, J \) can be updated in the \( (r+1) \)th iteration of the EM algorithm by

\[
d_j = \sum_{m=1}^{p_j} \left( \frac{\lambda_m (\sigma_j^2 - k_j \alpha_m^2)}{(\lambda_m^2 k_j + \sigma_j^2)} \right) / \sum_{m=1}^{p_j} \left( \frac{\lambda_m \alpha_m^2 + \sigma_j^2}{\lambda_m^2 k_j + \sigma_j^2} \right)
\]

where \( \hat{k}_j = \hat{k}_{LT,j}, \hat{\alpha}_{R,j} = (\widehat{\alpha}_{R,j_1}, \ldots, \widehat{\alpha}_{R,j_p}) \) is given by Lemma 1 and \( (\lambda_{i_j}, \ldots, \lambda_{j_p}) \) are eigenvalues of \( X_j^T W_j X_j \) with \( \widehat{\alpha}_{R,j} \) from (18). The E- and M-steps are alternated until the stopping criterion is satisfied. As the parameters \( k_j \) and \( d_j \) are updated in each iteration of the EM algorithm, the proposed LT method is henceforth called the iterative LT i.e. Liu_{ITE}.

Unlike the iterative LT method, one can follow Hoerl et al. \textsuperscript{28} to estimate the LT parameters based on ridge estimates \( \widehat{\alpha}_{R,j} \) and \( \hat{\beta}_{R,j} \). In other words, the mixture parameter \( \Psi \) and \( d_j \) are still iteratively updated in the EM algorithm; however, \( k_j \) and \( d_j \) are estimated only once throughout the EM algorithm using the ridge estimates. Here, we estimate the parameters by \( \hat{k}_{LT,j} = \hat{p}_{LT,j} \hat{\beta}_{LT,j} \). This LT estimation method is henceforth called HKP LT i.e. Liu_{HKP}.

**LT CEM algorithm:** Like previous subsections, the CEM algorithm partitions the observations in the C-step and then update the parameters within each partition. In the \( (r+1) \)th iteration of the CEM algorithm, the E-step remains the same as before. The C-step classifies the observations into partition \( P^{(r+1)} = (P_1^{(r+1)}, \ldots, P_j^{(r+1)}) \) where \( P_j^{(r+1)} = \{ (x_i, y_i); r_{ij}(\Psi^{(r)}) = \text{arg max}_r r_{ij}(\Psi^{(r)}) \} \) with \( (r_{ij}(\Psi^{(r)}), \ldots, r_{ij}(\Psi^{(r)})) \) are obtained from (4). Using \( P^{(r+1)} \), we update the mixing proportions from (12). We then require to estimate the LT parameters \( (k_j, d_j) \) in each iteration of the CEM algorithm. Similar to LT EM algorithm, we propose \( \hat{k}_{LT,j} = (\lambda_{i,j} - 100 \lambda_{j,p}) / 99 \) where \( \lambda_{i,j} \) and \( \lambda_{j,p} \) are the maximum and the minimum eigenvalues of \( X_j^T W_j X_j \).

Lemma 5. Under the assumptions of Lemma (1), the canonical LT estimator in the \( j \)th component regression \( j = 1, \ldots, J \) under the CEM algorithm is given by

\[
\widehat{\alpha}_{LT,j} = (\Lambda_j + k_j)^{-1} (\Lambda_j^{1/2} V_1^T W_j^{1/2} y - d_j \hat{\alpha}_j)
\]

and

\[
\widehat{\beta}_{LT,j} = U_j \widehat{\alpha}_{LT,j}
\]

where \( \hat{\alpha}_j \) is the canonical estimate of \( \alpha_j \) and \( V_1 = [v_{ij}, \ldots, v_{pj}] \) with \( v_{ij}, \ldots, v_{pj} \) are the orthonormal eigenvectors of \( W_j^{1/2} X_j \). From Lemmas 4 and 5 , one can estimate parameter \( d_j \) based on partition \( P_j^{(r+1)} \) from (25) where \( (\lambda_{i,j}, \ldots, \lambda_{j,p}) \) are eigenvalues of \( X_j^T W_j X_j \) and estimate \( \widehat{\alpha}_{R,j} \) from equation (20). To estimate the regression parameters, we implement a WLS.
based on the LT penalty as

$$
\hat{\beta}_{LT,j}^{(r+1)} = \arg \min_{\beta_j} \left( y_j - X_j \beta_j \right)^T W_j \left( y_j - X_j \beta_j \right) + \left[ \left( -\frac{\hat{a}_j}{\hat{k}_{LT,j}^{1/2}} \right) \hat{\beta}_j - \hat{k}_{LT,j}^{1/2} \beta_j \right]^T \left[ \left( -\frac{\hat{a}_j}{\hat{k}_{LT,j}^{1/2}} \right) \hat{\beta}_j - \hat{k}_{LT,j}^{1/2} \beta_j \right] \tag{26}
$$

where $y_j$ and $X_j$ are response vector and design matrix under $P_j^{(r+1)}$ and $\hat{\beta}_j$ can be any estimate for $\beta_j$. Also, $W_j$ is a weight diagonal matrix with entries $(r_j, \Psi^{(r)})_j, \ldots, \tau_{nj,j} \Psi^{(r)})_j); j = 1, \ldots, J$. One can easily find the solution to equation (22) and update the regression parameters by

$$
\hat{\beta}_{LT,j}^{(r+1)} = \left( X_j^T W_j X_j + \hat{k}_{LT,j} \right)^{-1} \left( X_j^T W_j y_j - \hat{a}_j \hat{\beta}_j \right) \tag{27}
$$

$$
\hat{\sigma}_{LT,j}^{2(r+1)} = \frac{\left( y_j - X_j \hat{\beta}_{LT,j}^{(r+1)} \right)^T W_j \left( y_j - X_j \hat{\beta}_{LT,j}^{(r+1)} \right)}{\sum_{i=1}^n \tau_{ij} \Psi^{(r)}} \tag{28}
$$

The E-, C-, and M-steps are repeatedly computed until the convergence criterion is satisfied.

Unlike the iterative LT CEM algorithm, one may estimate the parameters of the mixture model via the HKP LT CEM algorithm where the LT parameters $(k_j, d_j), j = 1, \ldots, J$ are updated once throughout the algorithm. From Hoerl et al.,\textsuperscript{28} Liu,\textsuperscript{5} we propose to estimate $\hat{k}_{LT,j} = p \hat{\sigma}_{R_j} / \hat{\beta}_{R_j} \hat{\beta}_{R_j}$ and $\hat{a}_j$ from equation (25) where $\hat{\beta}_{R_j}$ and $\hat{\sigma}_{R_j}$ come from equations (19) and (20), respectively.

**LT SEM algorithm:** Similar to the SEM algorithms, the S-step partitions the observations stochastically from Multi(1, $\tau_{i1}(\Psi^{(r)})_j, \ldots, \tau_{nj,j}(\Psi^{(r)})_j))$ for $i = 1, \ldots, n$. Once the partition is established, the rest of the LT SEM algorithms are implemented similarly to the LT CEM algorithms.

### 3 Simulation studies

In this section, we examine the performance of the ML, Ridge, and LT methods when the underlying population is a mixture of linear regression models with multicollinearity problem. We present two simulations enabling us to study the effect of the sample size, multicollinearity levels, and the number of components of mixture models on the estimation and prediction of the proposed methods.

This paper investigates, for the first time in the literature, the Ridge and LT shrinkage methods for heterogeneous populations using a mixture of regression models. Hence, the main goal of the numerical studies is to compare the estimation performance of the biased shrinkage methods with the commonly used benchmark ML method in the presence of multicollinearity. Although the initial values play an essential role in iterative methods like the EM algorithm, in these simulations as comparative numerical studies, we have $(J \times p)$ matrix of coefficients, mixing proportions and variance terms to deal with the parameters of $J$ component regressions each with $p$ covariates. For these reasons, in all the numerical studies, we chose the initial values fixed, close to the true values of the parameters and the same for all the EM, CEM, and SEM approaches under ML, Ridge, and LT estimation methods for a fair comparison. Also, we used $|\ell'(\Psi^{(i+1)}) - \ell'(\Psi^{(i)})| < 10^{-5}$ as the stopping rule for all the estimation methods. We set the maximum number of iterations to be 2000 for all the EM, CEM, and SEM approaches under the ML, Ridge, and LT methods. We used a single start for all the numerical studies. If the EM algorithms, for any reason, did not converge in 2000 iterations, the estimate in the last iteration was treated as the solution in the EM algorithm.

It is well-known that a FMM’s log-likelihood may appear multimodal; hence, the EM algorithm may converge to local maxima. We designed the numerical studies in a comparative way to compare the estimation performance of the proposed shrinkage methods with those of the commonly used ML method. Since the log-likelihood of all the estimation methods is subject to this multimodality, we did not exclude the convergence to the local maxima for all the estimation methods for a fair comparison. Although the algorithms may converge to either local or global maxima, the overall estimation performance of the methods is evaluated. Therefore, the local maxima estimates in simulations indirectly affect the errors of the estimation algorithms.

In the first study, we simulate data from a population corresponding to a mixture of two regression models with four covariates $(x_1, \ldots, x_4)$. Following Inan and Erdogan,\textsuperscript{29} we use $\rho$ to denote the correlation between covariates to simulate the multicollinearity in the mixture model. To do so, we first generate random numbers $\{w_{ij}; i = 1, \ldots, n; j = 1, \ldots, 5\}$ from
the standard normal distribution. The covariates are then generated by

\[ x_{ij} = (1 - \rho^2)^{1/2} w_{ij} + \rho w_{1.5}, \quad j = 1, \ldots, 4 \]

where the correlation matrix of \( X \) is given by \( 4 \times 4 \) whose diagonal entries are 1 and off-diagonal entries are equal to \( \rho^2 = \text{corr}(x_{ij}, x_{kj}); j, k = 1, \ldots, 4 \). Here, we set \( \rho \in \{0.88, 0.9, 0.95, 0.97, 0.99\} \) to simulate the multicollinearity levels in the mixture of regression models. The responses are then generated from mixture model (1) whose true parameters are given by \( \Psi = (\pi_0, \beta_0, \sigma^2_0, \sigma^2_m) \) with \( \pi_0 = 0.7, \beta_0 = (1, 3, 4, 5, 6) \), \( \beta_0 = (-1, -1, -2, -3, -5) \) and \( \sigma^2_0 = \sigma^2_m = (0.5, 1, 2) \). We measured the estimation performance of the ML, Ridge, and LT methods by root means square errors (RMSE) in estimating \( (\pi, \beta, \sigma^2) \) and computed RMSEP(\( \hat{\beta} \)) = \( [1 \times (\hat{\beta} - \beta_0)^\top (\hat{\beta}_0 - \beta_0)]^{1/2} \), RMSEP(\( \hat{\sigma} \)) = \( \sqrt{(\hat{\sigma} - \sigma_0)^2} \) and RMSEP(\( \hat{\sigma}_m \)) = \( \sqrt{(\hat{\sigma}_m - \sigma_0)^2} \), where \( \hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2), \hat{\sigma} = (\hat{\sigma}_1, \hat{\sigma}_2), \hat{\sigma}_0 = (\hat{\beta}_0, \hat{\beta}_0) \) and \( \sigma_0 = (\sigma^2_0, \sigma^2_0) \) and \( M \) denotes the number of free parameters. We also used the root mean squared errors of prediction (RMSEP) to evaluate the prediction performance of the methods. To do so, we first compute the RMSEP of the \( m \)th replicate through \( K \)-fold cross-validation by

\[ \text{RMSEP}^{(m)} = \left( \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i^{(m)})^2 \right)^{1/2} \]

where \( \hat{y}_i^{(m)} \) is the predicted response of the \( i \)th observation in the \( m \)th replicate.

We computed the estimation and prediction measures for the ML, Ridge, and LT methods as follows. We first generated a sample of size \( n = \{60, 100, 200\} \) from the underlying mixture of regression models as described above. We selected the starting point of parameters as \( \pi^{(0)} = 0.5, \beta^{(0)} = (-1, 1, 2, 3, 4), \beta_0^{(0)} = (1, 1, 0, -1, -3) \) and \( \sigma^{(0)} = (1, 2) \). We then used the EM, CEM, and SEM algorithms to estimate the parameters of the mixture population via ML, Ridge, and LT methods. We applied the idea of \( K = 5 \) cross-validation to assess the prediction performance of the methods. To this end, we divided the sample into \( K \) folds of equal sizes. We used \( K - 1 \) folds for training and the remaining fold for prediction. We repeated the procedure for all \( k = 1, \ldots, K \) to compute the RMSEP(\( m \)). Eventually, we replicated the entire procedures \( m = 2000 \) times and computed the median and 95% confidence interval (CI) for the SSE and RMSEP measures. The lower and upper bounds of the CI correspond to 2.5 and 97.5 percentiles of 2000 replications, respectively.

Due to sampling variability, the generated design matrix and response vector change from one replication to another in the simulation. Given this, Figure 55 (in the Appendix) reports the sampling distributions of the covariates and responses for a randomly selected sample with size \( n = 100 \) and \( \rho = 0.95 \). Due to sampling variability, the sample correlation may change slightly from the nominal correlation. Accordingly, for the selected random sample, the sample variance inflation factors (VIFs) are almost around VIF = 45.70 in Figure 55. We show the results of the simulation study in estimating \( (\pi, \beta, \sigma^2) \) in Figures 1 to 3 and 5 to 28 (in the Appendix). It is observed that the ML methods estimate slightly better the mixing proportion than the ridge and LT methods. This happens because the ridge and LT estimators are biased shrinkage methods where a slight bias is incorporated into the estimation to encounter the multicollinearity problem. We observe that the multicollinearity significantly impacts the ML estimates of the coefficients and results in extremely unreliable estimates for all EM, CEM, and SEM algorithms. Unlike ML estimates, a significant improvement is seen in the performance of the shrinkage methods in estimating the coefficients of the component regressions. The LT methods appear more reliable than the ridge counterparts in the multicollinearity. From a comparison between LT(ISTR) and Liu_HKP, we see that Liu_HKP provides more reliable estimates for \( \sigma^2 \). Among the Liu_HKP estimators, the CEM algorithm almost always outperforms its EM ad SEM counterparts. Tables 1 and 9 (in the Appendix) show the median and the length of 95% CI of the RMSEP for all the developed methods. The tables clearly show that the prediction performances of all the methods and EM algorithms are almost identical. This finding is consistent with Inan and Erdogan\(^{39}\) that multicollinearity seriously affects the estimation of the methods while prediction levels stay almost the same.

In the second simulation study, we evaluate the performance of the estimators when the underlying mixture model consists of three component regressions with two covariates. We set \( \rho \in \{0.9, 0.92, 0.95, 0.97, 0.99\} \) to simulate the multicollinearity in the mixture. As described earlier, we generated the covariates and responses from the mixture model with parameters \( \pi_0 = (\pi_{01}, \pi_{02}, \pi_{03}) = (0.3, 0.4, 0.3), \beta_0 = (1, 3, 4), \beta_0 = (-1, -1, -2), \beta_0 = (-3, 1, -4) \) and \( \sigma^2_0 = (0.25, 1, 0.09) \). For example, Figure 56 (in the Appendix) reports the sampling distributions of the covariates and responses for a randomly selected sample with size \( n = 100 \) and \( \rho = 0.95 \). Accordingly, for the selected random sample, the sample VIF is almost around VIF = 10.25. Similar to the settings of the first simulation study, we replicated 2000 times all the estimation and prediction procedures under the EM, CEM, and SEM algorithms and computed the median and 95% CI for the SSEs and RMSEP for sizes \( n = \{60, 100, 200\} \). Here again, we set fixed and the same starting points for all the estimation methods. The employed starting points include \( \pi^{(0)} = (0.3, 0.4, 0.3), \beta_{01}^{(0)} = \beta_{01} - 1, \beta_{02}^{(0)} = \beta_{02} - 1, \)
Figure 1. The median, lower, and upper bounds of 95% CIs for RMSE(\(\hat{\beta}\)) of the estimators when the population is a mixture of two regression models with \(n = 60\) and \(\sigma^2 = 0.5\). RMSE: root means square error; CI: confidence interval.

Figure 2. The median, lower, and upper bounds of 95% CIs for RMSE(\(\hat{\pi}\)) of the estimators when the population is a mixture of two regression models with \(n = 60\) and \(\sigma^2 = 0.5\). RMSE: root means square error; CI: confidence interval.
Figure 3. The median, lower, and upper bounds of 95% CIs for RMSE($\sigma^2$) of the estimators when the population is a mixture of two regression models with $n = 60$ and $\sigma^2 = 0.5$. RMSE: root means square error; CI: confidence interval.

$\beta^{(0)}_{03} = \beta_{03} - 1$ and $\sigma^{2(0)} = \sigma^2 + 0.02$. The estimation and prediction results are demonstrated in Figures 4 and 29 to 54 (in the Appendix) and Tables 13 to 21 (in the Appendix), respectively. Here, we also observe that the ML method slightly better estimates the mixing proportions; however, the ML methods result in unreliable estimates for the coefficients of component regressions. It is easy to see that shrinkage estimators do better in estimating the component regression parameters. In addition, the Liu_HKP almost always outperforms other methods and provides a more reliable estimate of the mixture of regression models. Therefore, the Liu_HKP method based on the CEM algorithm is recommended to fit the mixture of linear regression models in multicollinearity.

Although it is well known that the convergence of the SEM (SEM) algorithm (for all the estimation methods) is not guaranteed; the algorithm resembles a Markov chain that fluctuates around the ML estimate. For this reason, for a fair comparison between all the estimation algorithms, we set the maximum number of iterations to be 2000 for all the EM, CEM, and SEM approaches. If the EM algorithms, for any reason, did not converge in 2000 iterations, the estimate in the last iteration was treated as the solution in the EM algorithm. It is also well known that the log-likelihood of the mixture model may appear multi-model, and the EM algorithm may trap in the local maxima. Although the algorithms in this manuscript may converge to either local or global maxima for all estimation methods, the overall estimation performance of the EM, CEM, and SEM of the ML, Ridge, and LT methods are evaluated in a comparative way using RMSE, which takes into account the squared distance of all the estimates from the true value of the parameters of the mixture model.

To this end, we designed the third simulation study to compare the convergence of the proposed shrinkage estimation algorithms with the commonly used ML counterparts using the same stopping rule. To do so, we used the tolerance levels $\epsilon \in \{10^{-3}, 10^{-5}\}$, while we kept the maximum number of iterations fixed for all methods and the same as 2000. We replicated 2000 times the entire data collection and the estimation in the same way as described above. We then computed the convergence rates of each estimation algorithm by the proportion of times the algorithm converged before reaching the maximum number of iterations. Table 22 (in the Appendix) demonstrates the convergence rates of the different estimation methods and EM algorithms under two different underlying mixtures of regression models with two and three components. From Table 22, we observe that the convergence rates of the Ridge and LT shrinkage methods outperform that of their ML counterpart using EM, CEM, and SEM algorithms in the presence of the multicollinearity. Moreover, this superiority is consistent across both tolerance levels, indicating the robustness of the shrinkage methods in achieving convergence.
Additionally, the convergence rates highlight the stability and reliability of the proposed shrinkage estimation algorithms, suggesting their suitability for practical applications in scenarios characterized by multicollinearity.

4 Bone data analysis

As a bone metabolic disease, osteoporosis occurs when the bone mineral architecture of the body deteriorates. This deterioration results in skeletal fragility and a high risk of osteoporotic fractures in different body areas such as the hip and femur. Osteoporosis and its related diseases significantly impact the patient’s health and survival. For example, one out of every two patients with osteoporotic hip fractures can no longer live independently, and one out of three may die within one year after the medical complication of the broken bone.\(^2,30\) The BMD of individuals improves until age 30 and then decreases as individuals age. The BMD score of an individual is compared with a BMD norm to determine the bone disorder status. The BMD norm is computed by the mean BMD scores of healthy adults aged 20–30. The bone status of an individual is diagnosed as osteoporosis when the BMD score is less than \(-2.5\) SD from the BMD norm of the population.

The BMD scores are obtained via an expensive and time-consuming procedure. Despite this, the researchers have access to various easy-to-measure patients’ characteristics, such as age, weight, BMI, and test results from earlier surveys.\(^31,32\) Regression models are among the most common methods to investigate the impact of a set of patients’ characteristics on the BMD responses. The impact of the characteristics may differ at different BMD levels. Hence, the inference on BMD measurements can be handled as a problem of the mixture of linear regression models.

The bone mineral data in this section were obtained from the National Health and Nutritional Examination Survey (NHANES III) administrated by the CDC on more than 33999 American adults between the years 1988–1994. The data set is publicly available on the website of the CDC from (https://wwwn.cdc.gov/nchs/nhanes/nhanes3/Default.aspx). One hundred eighty-two White women aged 50 and older participated in both bone examinations during the survey. Due to the substantial impact of osteoporosis on older women, these 182 women were considered our target population. From the second bone examination, we focused on femur BMD as the response variable. We also used two easily attainable physical characteristics as the explanatory variables of the regression. These physical characteristics include arm and bottom circumferences from the first examination. The association between the two explanatory variables is \(\rho = 0.81\). Using the ML method based on the information of all individuals in the population, the BIC criterion suggests that a mixture of two regression models is the best fit with parameters \(\pi_0 = (\pi_{01}, \pi_{02}) = (0.64, 0.36), \beta_{01} = (0.004, 0.006), \beta_{02} = (0.011, 0.005)\),
and \( \sigma^2_0 = (0.0081, 0.0144) \). These parameters were then considered as the true parameters of the mixture population. Figure 57 reports the histograms of the explanatory variables and the responses for the two components. As an unsupervised learning, the true component membership of the observations was obtained as follows. The individuals whose BMD score is less than one standard deviation from the MD norm (the average of the BMD scores from the healthy adults between 20 and 30) were assigned to the first component (osteoporosis) and the others to the second component. We replicated 2000 times the ML, ridge, and LT methods under EM, CEM, and SEM algorithms in estimating the parameters of the bone mineral population based on sample sizes \( n = \{60, 100, 200\} \). Similar to Section 3, we selected the fixed and the same starting points for all estimation methods as \( \pi^{(0)} = (0.7, 0.3), \beta_0^{(0)} = \beta_1 + 0.001, \beta_2^{(0)} = \beta_2 + 0.001, \) and \( \sigma_0^{(0)} = \sigma_0 + 0.001 \). Similar to Section 3, we used \( |\bar{c}(\Psi^{(t+1)}) - \bar{c}(\Psi^{(t)})| < 10^{-5} \) as the stopping rule for all the estimation methods with the maximum number of iterations being 2000 for all the EM, CEM, and SEM approaches. We then computed the estimation and prediction measures RMSE(\( \hat{\beta} \)), \text{RMSE}(\hat{\psi}) \), \text{RMSE}(\sigma^2) \) and MRSEP using 5-fold cross-validation as described in Section 3.

Tables 10 to 12 (in the Appendix) report the median and 95% CI for the above measures in estimating and predicting the bone mineral population. The CI’s lower (L) and upper (U) bounds correspond to 2.5 and 97.5 percentiles, respectively. Although all ML, Ridge, and LT methods almost perform identically in estimating the mixing proportion and component variances, \( \hat{\beta}_{ML} \) become considerably unreliable. Unlike ML methods, the LT and ridge shrinkage methods could appropriately handle the multicollinearity in estimating the coefficients of component regressions. Comparing the shrinkage methods, we observe that the LT estimators appear more reliable than their ridge counterparts in estimating the parameters of the bone mineral population.

We designed the real data analysis this way for several reasons. First, similar to the simulation studies, we require estimating the parameters of the entire bone data so we can fairly calculate and compare the RMSEs of all the ML, Ridge, and LT methods in estimating the bone mixture population. While the original bone data are co-linear, we applied the benchmark ML method to the entire bone data since the ML method is the only method currently available in the literature in estimating the parameters of the mixture of the regressions. In contrast, we propose the Ridge and LT shrinkage methods as a more reliable alternative to the ML method in estimating the mixture of regression models in the presence of collinearity. Note that our proposed shrinkage methods outperformed the ML method in estimating the mixture population of the bone data, on average, over 2000 replications, even in the case where the true parameters of the bone data were estimated by the ML method. Because the simulation studies and the above real data study showed the superiority of the proposed shrinkage methods, we also added a new numerical experiment according to the real-life scenario. In this study, we once applied all the proposed estimation methods to the entire bone data and estimated the parameters of the bone mixture population. Table 23 (in the Appendix) shows the results of this numerical study.

5 Summary and concluding remarks

In medical and environmental research (e.g. osteoporosis research), linear regression models are used as standard statistical methods to investigate the relationship between a set of covariates with the response variable. When the underlying population is heterogeneous, the impact of the covariates on the response may change in different subpopulations. A mixture of linear regression models can be considered a solution to the problem. The ML method is a common technique to fit a mixture of regression models; however, the ML estimates become unreliable when there is multicollinearity between the covariates. We investigated shrinkage methods, including ridge and LT estimators based on the EM, CEM, and SEM algorithms in estimating the parameters of the mixture of regression models. Through extensive numerical studies, we investigated the estimation performance of the proposed methods on 972 different simulation combinations, accounting for 3 sample sizes, 3 standard deviations, 3 EM algorithms, 3 estimation methods, 2 mixture settings, and 6 multicollinearity levels. We showed that LT estimators outperformed their ridge and ML counterparts through extensive numerical studies. Finally, we applied the developed methods to analyze the bone mineral data of women aged 50 and older. While the methods were only applied to osteoporosis research, the developed methods are generic and can be applied to other medical and environmental studies. The code to implement the simulation studies and real data analysis is accessible through our public GitHub repository (https://github.com/ElsayedGhanem/Mixture-of-Linear-Regression).

While the VIF = 2.91 is not too high in the real data analysis, we selected two physical characteristics: the arm and bottom circumferences mainly because the correlation between these two independent variables is relatively high, 0.81, almost close to the minimum correlation value considered in the simulation studies. Furthermore, these two explanatory variables are easily attainable characteristics of the patients. During the analysis, we only had access to the demographic information of individuals from the first examination, as the mass screening examination. For this reason, we had to use patients’ physical characteristics from the first examination as the explanatory variables. Without loss of generality, the patient’s physical characteristics from the examination can also be used. As another example, according to the high association between
different BMD scores from different body tissues, one may consider the different BMD scores as explanatory variables, resulting in more severe multicollinearity in the bone mixture population.

In this paper, we investigated the Ridge and LT shrinkage methods for a mixture of regression models to deal with multicollinearity as one of the most common challenges in regression models. Given these, the proposed estimation methods require extensive iterative algorithms. Consequently, new research is required to investigate the initialization of the iterative algorithms under shrinkage methods and develop reliable and efficient initialization methods\(^{33,34}\) for the shrinkage method for the mixture of regression models. According to the multimodality of the log-likelihood in mixture models, new research is needed to investigate thoroughly this research direction to guarantee the convergence of the EM algorithms to global maxima and possibly from multiple and random starts of the algorithms to use the monotonic properties of the EM algorithms and escape from the local optima. We only considered the continuous, univariate response variable. It is desirable to develop shrinkage methods for the mixture of regression models with multiple response variables and mixed types of data. In this manuscript, we developed shrinkage methods to address the collinearity challenges in the mixture of regression models using a fully parametric framework. Based on recent advancements in semi-parametric mixture models for linear regression,\(^{35}\) we aim to explore the application of developed shrinkage techniques to mitigate collinearity issues in the semi-parametric mixture of regression models. Last but not least, throughout this manuscript, we did not consider the effects of the sampling weights in all the estimation methods and EM algorithms, even for the commonly used ML method available in the literature. Given the impact of the sampling weights in model misspecification and the estimation biases,\(^{36}\) more research studies are needed in future to assess the impacts of the sampling weights on LT, Ridge, and ML methods under different EM algorithms in estimating the parameters of a mixture of regression models.

Acknowledgment

Armin Hatefi and Hamid Usefi acknowledge the research support of the Natural Sciences and Engineering Research Council of Canada (NSERC). The authors would like to thank the two anonymous reviewers whose comments improved the quality and presentation of the manuscript.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

Armin Hatefi and Hamid Usefi acknowledged the financial research support received from Natural Sciences and Engineering Research Council of Canada (NSERC).

ORCID iD

Armin Hatefi https://orcid.org/0000-0002-8357-7125

Supplemental material

Supplemental material for this article is available online.

References

1. Cummings SR, Nevitt MC, Browner WS, et al. Risk factors for hip fracture in White women. New Engl J Med 1995; 332: 767–774.
2. Neuberger J, Currie C, Wakeman R, et al. The impact of a national clinician-led audit initiative on care and mortality after hip fracture in England: An external evaluation using time trends in non-audit data. Med Care 2015; 53: 686.
3. WHO. Assessment of fracture risk and its application to screening for postmenopausal osteoporosis: Report of a WHO study group [meeting held in rome from 22 to 25 june 1992]. 1994.
4. LJ Melton III, Atkinson EJ, O’connor MK, et al. Bone density and fracture risk in men. J Bone Miner Res 1998; 13: 1915–1923.
5. Lim H-S, Kim S-K, Lee H-H, et al. Comparison in adherence to osteoporosis guidelines according to bone health status in Korean adult. J Bone Metab 2016; 23: 143–148.
6. Liu K. Using Liu-type estimator to combat collinearity. Commun Stat-Theory Method 2003; 32: 1009–1020.
7. Duran EA, Hardle WK and Osipenko M. Difference based ridge and Liu type estimators in semiparametric regression models. J Multivar Anal 2012; 105: 164–175.
8. Arashi M, Kibria BMG, Norouzirad M, et al. Improved preliminary test and stein-rule Liu estimators for the ill-conditioned elliptical linear regression model. J Multivar Anal 2014; 126: 53–74.
9. Pearce AD and Hatefi A. Multiple observers ranked set samples for shrinkage estimators. arXiv preprint arXiv:2110.07851, 2021.
10. Dempster AP, Laird NM and Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc: Ser B (Methodological) 1977; 39: 1–22.
11. Celeux G. The SEM algorithm: A probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Comput Stat Quart* 1985; 2: 73–82.
12. Celeux G and Govaert G. A classification EM algorithm for clustering and two stochastic versions. *Comput Stat Data Anal* 1992; 14: 315–332.
13. Quandt RE and Ramsey JB. Estimating mixtures of normal distributions and switching regressions. *J Am Stat Assoc* 1978; 73: 730–738.
14. Faria S and Soromenho G. Fitting mixtures of linear regressions. *J Stat Comput Simul* 2010; 80: 201–225.
15. Hawkins DS, Allen DM and Stromberg AJ. Determining the number of components in mixtures of linear models. *Comput Stat Data Anal* 2001; 38: 15–48.
16. Frühwirth-Schnatter S. *Finite mixture and Markov switching models*. New York: Springer, 2006.
17. Hennig C. Identifiability of models for clustervise linear regression. *J Classif* 2000; 17: 273–296.
18. Ho N and Nguyen X. On strong identifiability and convergence rates of parameter estimation in finite mixtures. *Electron J Stat* 2016; 10: 271–307.
19. Zhang Z, Li WK and Yuen KC. On a mixture GARCH time-series model. *J Time Ser Anal* 2006; 27: 577–597.
20. Wedel M, Ter Hofstede F and M Steenkamp J-BE. Mixture model analysis of complex samples. *J Classif* 1998; 15: 225–244.
21. Peel D and MacLachlan G. *Finite mixture models*. New York: John & Sons, 2000.
22. Karlis D and Xekalaki E. Choosing initial values for the EM algorithm for finite mixtures. *Comput Stat Data Anal* 2003; 41: 577–590.
23. Felson DT, Zhang Y, Hannan MT, et al. Effects of weight and body mass index on bone mineral density in men and women: The Framingham study. *J Bone Miner Res* 1993; 8: 567–573.
24. Kim SJ, Yang W-G, Cho E, et al. Relationship between weight, body mass index and bone mineral density of lumbar spine in women. *J Bone Metab* 2012; 19: 95–102.
25. Furman WD and Lindsay BG. Measuring the relative effectiveness of moment estimators as starting values in maximizing likelihoods. *Comput Stat Data Anal* 1994; 17: 493–507.