Unsupervised Shrinkage Estimation Methods for Mixture of Regression Models

Armin Hatefi

A joint work with **Hamid Usefi** and **Elsayed Ghanem**

Department of Mathematics and Statistics,
Memorial University

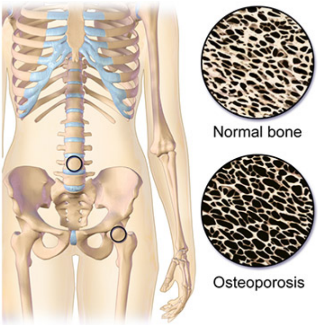
CMS Summer Meeting – June 5, 2022

Why mixture of regressions?

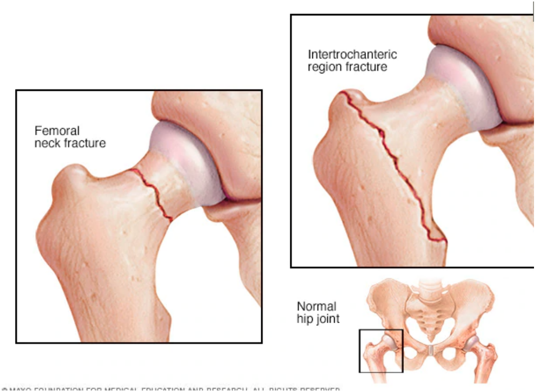
- Mixture of regressions: when the underlying population comprises of heterogeneous subpopulations.
- The Maximum likelihood method is one of the most common method estimating the regression parameters.
- More predictors but multicollinearity problem: unreliable estimates, wide confidence intervals, test of significance,
- Many characteristics can not even be considered as predictors of regression? Rank information for sampling designs.
- In this talk, we develop biased but more reliable methods including ridge and Liu-type (LT) for mixture of regressions.

Osteoporosis

- reduced bone mineral density (BMD) with deterioration of bone architecture
- increased risk of fracture, skeletal fragility, other related – bone disorders
- **WHO:** BMD measurement is one of the most important predictors for osteoporosis diagnosis & bone disorders.
- BMDs are obtained via dual X-ray absorptiometry (DXA).



- 3 out of 4 patients with osteoporosis are not aware of their disease.
- At least 1 in 3 women and 1 in 5 men aged 50 and older will experience osteoporotic fractures (such hip).
- 53% of patients with osteoporotic hip fracture can no longer live independently.
- And 28% die within one year of the complication.



Mixture of Regression Models

- Let $\mathbf{x}_i^\top = (x_{i,1}, \ldots, x_{i,p})$ be the vector of p predictors for the i-th subject for $i = 1, \ldots, n$.
- Let $\mathbf{y} = (y_1, \ldots, y_n)$ and $\mathbf{X} = (\mathbf{x}_1^\top, \ldots, \mathbf{x}_n^\top)^\top$ denote the respect vector and $(n \times p)$ design matrix with $\text{rank}(\mathbf{X}) = p < n$.
- The mixture of regression models:

$$y_i = \begin{cases} \mathbf{x}_i^\top \beta_1 + \epsilon_{i1}, & \text{with probability } \pi_1 \\ \vdots & \\ \mathbf{x}_i^\top \beta_J + \epsilon_{iJ}, & \text{with probability } \pi_J \end{cases} \tag{1}$$

- $\beta_j = (\beta_{j,1}, \ldots, \beta_{J,p})$ and $\sum_{j=1}^M \pi_j = 1$ and $0 < \pi_j < 1$. Also $\epsilon_{ij} \overset{iid}{\sim} N(0, \sigma_j^2)$ for $j = 1, \ldots, J$.
- We assume that the number of components J in the mixture model (1) is known; however, the component memberships are unknown and should be estimated in an unsupervised approach.
- Let $\beta = (\beta_1, \ldots, \beta_J)$. Let $\theta_j = (\beta_j, \sigma_j^2)$. Thus, $\mathbf{\Psi} = (\pi, \theta_1, \ldots, \theta_J)$.

Mixture of Regression Models

From regression model (2), the log-likelihood function of $\boldsymbol{\Psi}$ can be written as

$$\ell(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \log\left(\sum_{j=1}^{J} \pi_j \phi_j(\mathbf{x}_i^\top \beta_j, \sigma_j^2)\right), \tag{3}$$

where $\phi_j(\mathbf{x}_i^\top \beta_j, \sigma_j^2)$ represents the pdf of normal distribution. For each subject $(\mathbf{x}_i, y_i)$, we introduce latent variables $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{iJ})$ for $i = 1, \ldots, n$ as

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

It is easy to see $\mathbf{Z}_i | y_i \overset{iid}{\sim} \text{Multi}(1, \tau_{i1}(\boldsymbol{\Psi}), \ldots, \tau_{iJ}(\boldsymbol{\Psi}))$ where

$$\tau_{ij}(\boldsymbol{\Psi}) = \frac{\pi_j \phi_j(\mathbf{x}_i^\top \beta_j, \sigma_j^2)}{\sum_{j=1}^{J} \pi_j \phi_j(\mathbf{x}_i^\top \beta_j, \sigma_j^2)}. \tag{4}$$

Let $(\mathbf{X}, \mathbf{y}, \mathbf{Z})$ denote the complete data. Then

$$\ell_c(\boldsymbol{\Psi}) = \sum_{i=1}^{n}\sum_{j=1}^{J} z_{ij} \log(\pi_j) + \sum_{i=1}^{n}\sum_{j=1}^{J} z_{ij} \log\{\phi_j(\mathbf{x}_i^\top \beta_j, \sigma_j^2)\}. \tag{5}$$

ML Method

- The EM algorithm turns the estimation into an iterative expectation (E) and maximization (M) steps.
- Let $\mathbf{\Psi}^{(0)} = (\pi^{(0)}, \theta_1^{(0)}, \ldots, \theta_J^{(0)})$ and $\mathbf{\Psi}^{(r)}$ be the r-th iteration of the EM algorithm.
- **E-step**: The conditional expectation of complete data log-likelihood is given by

$$\mathbf{Q}(\mathbf{\Psi}, \mathbf{\Psi}^{(r)}) = \mathbf{Q}_1(\pi, \mathbf{\Psi}^{(r)}) + \mathbf{Q}_2(\theta, \mathbf{\Psi}^{(r)}),$$

where

$$\mathbf{Q}_1(\pi, \mathbf{\Psi}^{(r)}) = \sum_{i=1}^{n} \sum_{j=1}^{J} \tau_{ij}(\mathbf{\Psi}^{(r)}) \log(\pi_j), \tag{9}$$

and

$$\mathbf{Q}_2(\theta, \mathbf{\Psi}^{(r)}) = \sum_{i=1}^{n} \sum_{j=1}^{J} \tau_{ij}(\mathbf{\Psi}^{(r)}) \log\left\{\phi_j(\mathbf{x}_i^{\top}\beta_j, \sigma_j^2)\right\}. \tag{10}$$

ML Method

M-step:

$$\widehat{\pi}_j^{(r+1)} = \sum_{i=1}^{n} \tau_{ij}(\mathbf{\Psi}^{(r)})/n; \qquad j = 1, \ldots, J-1. \tag{13}$$

The maximization of $\mathbf{Q}_2(\theta, \mathbf{\Psi}^{(r)})$ can be reformulated as the weight least square method as follows

$$\widehat{\beta}_j^{(r+1)} = \arg\min_{\beta_j} \, (\mathbf{y} - \mathbf{X}\beta)^\top \mathbf{W}_j (\mathbf{y} - \mathbf{X}\beta)/n, \tag{14}$$

where $\mathbf{W}_j$ is $n \times n$ diagonal matrix with diagonal elements $(\tau_{ij}(\mathbf{\Psi}^{(r)}), \ldots, \tau_{nj}(\mathbf{\Psi}^{(r)}))$ for all $j = 1, \ldots, J$.

$$\widehat{\sigma}_j^{2(r+1)} = \frac{(\mathbf{y} - \mathbf{X}\widehat{\beta}^{(r+1)})^\top \mathbf{W}_j^{(r)} (\mathbf{y} - \mathbf{X}\widehat{\beta}^{(r+1)})}{\sum_{i=1}^{n} \tau_{ij}(\mathbf{\Psi}^{(r)})}, \quad j = 1, \ldots, J. \tag{15}$$

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

Classification EM Algorithm: The CEM algorithm incorporates a classification (C) step between E and M steps.

- **C-step**: Let $\mathbf{P}^{(r+1)} = (P_1^{(r+1)}, \dots, P_J^{(r+1)})$ denote the partition in the $(r+1)$-th iteration.

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

$$\widehat{\pi}_j^{(r+1)} = n_j/n; \quad j = 1, \dots, J, \tag{20}$$

$$\widehat{\beta}_j^{(r+1)} = \left(\mathbf{X}_j^\top \mathbf{W}_j \mathbf{X}_j\right)^{-1} \mathbf{X}_j^\top \mathbf{W}_j \mathbf{y}_j, \tag{21}$$

$$\widehat{\sigma}_j^{2(r+1)} = \frac{(\mathbf{y}_j - \mathbf{X}_j \widehat{\beta}^{(r+1)})^\top \mathbf{W}_j^{(r)} (\mathbf{y}_j - \mathbf{X}_j \widehat{\beta}^{(r+1)})}{\sum_{i=1}^n \tau_{ij}(\boldsymbol{\Psi}^{(r)})}, \tag{22}$$

- $\mathbf{W}_j^{(r)}$ is the diagonal weight matrix of size n_j with diagonal entries $(\tau_{ij}(\boldsymbol{\Psi}^{(r)}), \dots, \tau_{n_j,j}(\boldsymbol{\Psi}^{(r)}))$.
  Stochstic EM Algorithm: The S-step simulates a random allocation for tehobservations.

$$\mathbf{Z}_i^* = (Z_{i1}^*, \dots, Z_{iJ}^*) \overset{iid}{\sim} \text{Multi}(1, \tau_{i1}(\boldsymbol{\Psi}^{(r)}), \dots, \tau_{iJ}(\boldsymbol{\Psi}^{(r)})) \quad (i = 1, \dots, n).$$

Ridge Method

- When $\ell(\mathbf{\Psi})$ is the incomplete log-likelihood, and $k > 0$ is the ridge parameter, then

$$\ell^R(\mathbf{\Psi}) = \ell(\mathbf{\Psi}) - k\beta^\top \beta/2 \tag{26}$$

From the weighted least square, we have

$$\widehat{\beta}_{R,j}^{(r+1)} = \arg\min_{\beta_j} \, (\mathbf{y} - \mathbf{X}\beta)^\top \mathbf{W}_j (\mathbf{y} - \mathbf{X}\beta) + k_j \beta_j^\top \beta_j/2, \tag{27}$$

$$\widehat{\sigma}_{R,j}^{2(r+1)} = \frac{(\mathbf{y} - \mathbf{X}\widehat{\beta}_R^{(r+1)})^\top \mathbf{W}_j^{(r)} (\mathbf{y} - \mathbf{X}\widehat{\beta}_R^{(r+1)})}{\sum_{i=1}^n \tau_{ij}(\mathbf{\Psi}^{(r)})}, \tag{28}$$

where $\widehat{\beta}_R^{(r+1)} = (\widehat{\beta}_{R,1}^{(r+1)}, \dots, \widehat{\beta}_{R,J}^{(r+1)})$. There are various methods available in the literature for estimation of k_j. Following [1],

$$\widehat{k_j} = p\widehat{\sigma}^2_{ML,j}/\widehat{\beta}^\top_{ML,j}\widehat{\beta}_{ML,j}$$

Lemma 1. *Under the assumptions of mixture of regression models (2), suppose* $\lambda_{1j}, \ldots, \lambda_{pj}$ *and* $u_{1j}, \ldots, u_{pj}$ *be eigenvalues and orthonormal eigenvectors of* $\mathbf{X}^\top \mathbf{W}_j \mathbf{X}$ *where* $\mathbf{W}_j$ *is* $n \times n$ *diagonal matrix with entries* $(\tau_{ij}(\boldsymbol{\Psi}^{(r)}), \ldots, \tau_{nj}(\boldsymbol{\Psi}^{(r)}))$ *under ridge EM algorithm. Let* $\boldsymbol{\Lambda}_j = diag(\lambda_{1j}, \ldots, \lambda_{pj})$ *and* $\mathbf{U}_j = [u_{1j}, \ldots, u_{pj}]$*. Then The canonical weighted ridge estimator in each component regression is given by*

$$\widehat{\alpha}_{R,j} = (\boldsymbol{\Lambda}_j + k_j)^{-1} \boldsymbol{\Lambda}_j^{1/2} \mathbf{V}_1^\top \mathbf{W}_j^{1/2} \mathbf{y}.$$

and

$$\widehat{\beta}_{R,j} = \mathbf{U}_j \widehat{\alpha}_{R,j}$$

with $\mathbf{V}_1 = [v_{1j}, \ldots, v_{pj}]$ *where* $v_{1j}, \ldots, v_{pj}$ *are the orthonormal eigenvectors of* $\mathbf{W}_j^{1/2} \mathbf{X} \mathbf{X}^\top \mathbf{W}_j^{1/2}$*.*

Ridge CEM (or SEM) Algorithm:

- **C-step (S-step):** Let $\mathbf{P}^{(r+1)} = (P_1^{(r+1)}, \ldots, P_J^{(r+1)})$ denote the partition in the $(r+1)$-th iteration.

$$\widehat{\beta}_{R,j}^{(r+1)} = \arg\min_{\beta_j} \, (\mathbf{y}_j - \mathbf{X}_j\beta)^\top \mathbf{W}_j(\mathbf{y}_j - \mathbf{X}_j\beta) + k_j\beta_j^\top\beta_j/2, \tag{31}$$

$$\widehat{\sigma}_{R,j}^{2(r+1)} = \frac{(\mathbf{y}_j - \mathbf{X}_j\widehat{\beta}_R^{(r+1)})^\top \mathbf{W}_j^{(r)}(\mathbf{y}_j - \mathbf{X}_j\widehat{\beta}_R^{(r+1)})}{\sum_{i=1}^n \tau_{ij}(\mathbf{\Psi}^{(r)})}, \tag{32}$$

Lemma 3. *Under the assumptions of mixture of regression models* (2)*, with component regression models* $\mathbf{y}_j = \mathbf{X}_j\beta_j + \epsilon$ *based on* n_j *observations with rank*$(\mathbf{X}_j) = p$*. Suppose* $\lambda_{1j}, \ldots, \lambda_{pj}$ *and* $u_{1j}, \ldots, u_{pj}$ *be eigenvalues and orthonormal eigenvectors of* $\mathbf{X}_j^\top\mathbf{W}_j\mathbf{X}_j$ *where* $\mathbf{W}_j$ *is* $n_j \times n_j$ *diagonal matrix with entries* $(\tau_{ij}(\mathbf{\Psi}^{(r)}), \ldots, \tau_{nj}(\mathbf{\Psi}^{(r)}))$ *under ridge CEM or ridge SEM algorithm. Let* $\mathbf{\Lambda}_j = diag(\lambda_{1j}, \ldots, \lambda_{pj})$ *and* $\mathbf{U}_j = [u_{1j}, \ldots, u_{pj}]$*. Then The canonical weighted ridge estimator in each component regression is given by*

$$\widehat{\alpha}_{R,j} = (\mathbf{\Lambda}_j + k_j)^{-1}\,\mathbf{\Lambda}_j^{1/2}\mathbf{V}_1^\top\mathbf{W}_j^{1/2}\mathbf{y}.$$

and

$$\widehat{\beta}_{R,j} = \mathbf{U}_j\widehat{\alpha}_{R,j}$$

with $\mathbf{V}_1 = [v_{1j}, \ldots, v_{pj}]$ *where* $v_{1j}, \ldots, v_{pj}$ *are the orthonormal eigenvectors of* $\mathbf{W}_j^{1/2}\mathbf{X}_j\mathbf{X}_j^\top\mathbf{W}_j^{1/2}$.

LT Method

- When $\ell(\mathbf{\Psi})$ is the incomplete log-likelihood, and $k > 0, d \in \mathbb{R}$ is the ridge parameter, then

$$\ell^{LT}(\mathbf{\Psi}) = \ell(\mathbf{\Psi}) - \left[(-\frac{d}{k^{1/2}})\widehat{\beta} - k^{1/2}\beta\right]^{\top}\left[(-\frac{d}{k^{1/2}})\widehat{\beta} - k^{1/2}\beta\right]. \tag{38}$$

From the weighted least square, we have

$$\widehat{\beta}_{LT,j}^{(r+1)} = \underset{\beta_j}{\arg\min}\ (\mathbf{y} - \mathbf{X}\beta)^{\top}\mathbf{W}_j(\mathbf{y} - \mathbf{X}\beta) + \left[(-\frac{d_j}{k_j^{1/2}})\widehat{\beta}_j - k_j^{1/2}\beta_j\right]^{\top}\left[(-\frac{d_j}{k_j^{1/2}})\widehat{\beta}_j - k_j^{1/2}\beta_j\right]. \tag{39}$$

$$\widehat{\sigma}_{LT,j}^{2(r+1)} = \frac{(\mathbf{y} - \mathbf{X}\widehat{\beta}_{LT}^{(r+1)})^{\top}\mathbf{W}_j^{(r)}(\mathbf{y} - \mathbf{X}\widehat{\beta}_{LT}^{(r+1)})}{\sum_{i=1}^{n}\tau_{ij}(\mathbf{\Psi}^{(r)})}, \tag{40}$$

$$\widehat{k}_{LT,j} = \frac{\lambda 1j - 100\lambda_{pj}}{99}$$

where λ_{1j} and λ_{pj} are max and min eigenvalues of $\mathbf{X}^{\top}\mathbf{W}_j\mathbf{X}$.

Lemma 5. *Under the assumptions of Lemma 1, the canonical weighted LT estimator in each component regression under LT EM algorithm is given by*

$$\widehat{\alpha}_{LT,j} = (\mathbf{\Lambda}_j + k_j)^{-1} \left(\mathbf{\Lambda}_j^{1/2} \mathbf{V}_1^\top \mathbf{W}_j^{1/2} \mathbf{y} - d_j \widehat{\alpha}_j \right).$$

and

$$\widehat{\beta}_{LT,j} = \mathbf{U}_j \widehat{\alpha}_{LT,j}$$

with $\mathbf{V}_1 = [v_{1j}, \ldots, v_{pj}]$ *where* $v_{1j}, \ldots, v_{pj}$ *are the orthonormal eigenvectors of* $\mathbf{W}_j^{1/2} \mathbf{X} \mathbf{X}^\top \mathbf{W}_j^{1/2}$.

Lemma 6. *Under the assumptions of Lemma 1, for each component regression $j = 1, \ldots, J$ and $k_j > 0$,*

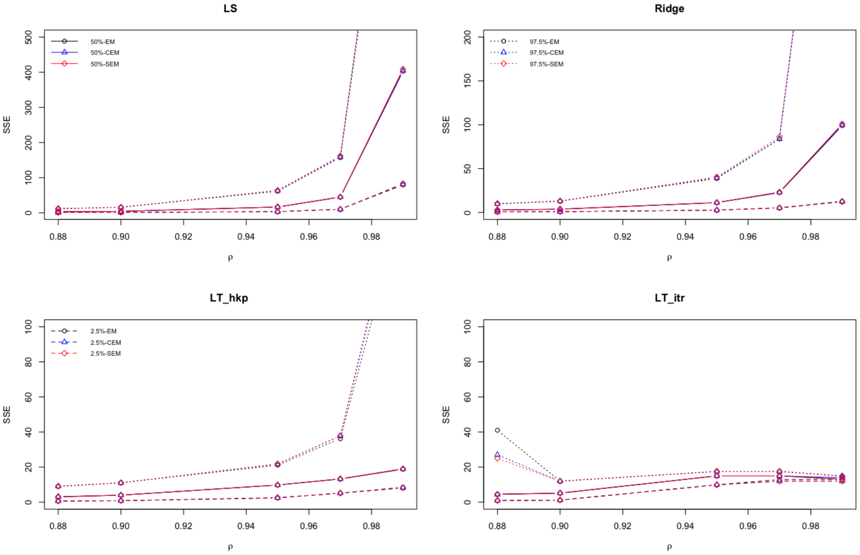
$$d_j = \sum_{m=1}^{p}((\sigma_j^2 - k_j\alpha_{lj}^2)/(\lambda_{lj} + k_j)^2)/\sum_{l=1}^{p}((\lambda_{lj}\alpha_{lj}^2 + \sigma_j^2)/\lambda_{lj}(\lambda_{lj} + k_j)^2)$$

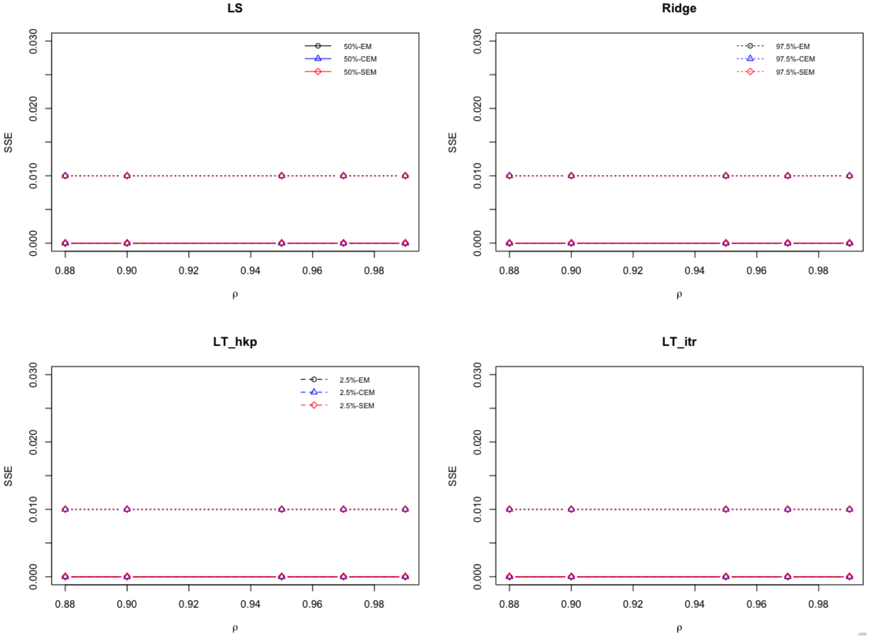
and

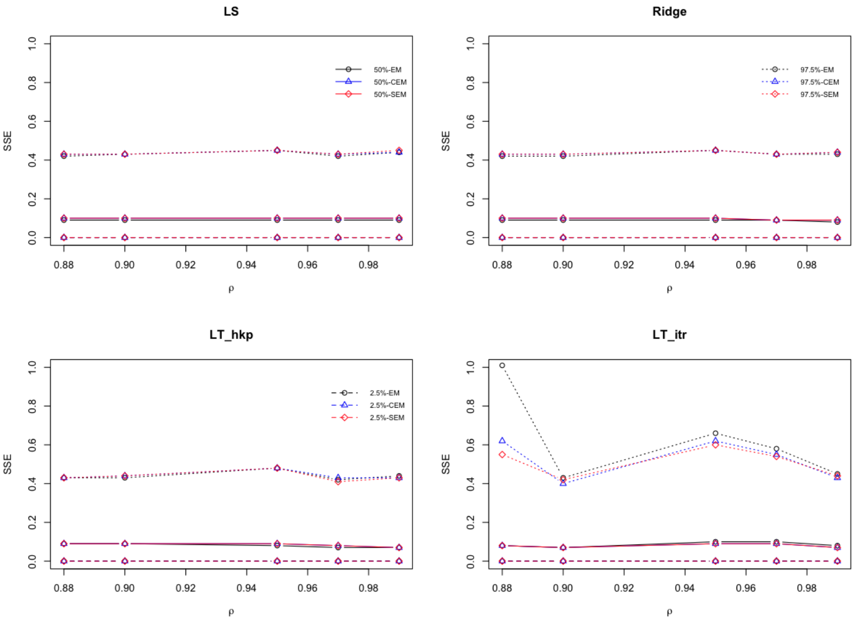
$$\widehat{\beta}_{LT,j} = \mathbf{U}_j\widehat{\alpha}_{LT,j}$$

minimizes the $MSE(\widehat{\alpha}_{LT,j})$.

- Iterative LT method uses Lemma (6) and iteratively updates d_j and k_j.
- As a modified approach, $\mathrm{LT}_K HPonlyupdates\mathrm{d}_j$ and k_j once using the final ridge estimates.
- The same results can be obtained for LT CEM and SEM algorithm.


LS
Ridge
LT_hkp
LT_itr
SSE
ρ
50%-EM
50%-CEM
50%-SEM
97.5%-EM
97.5%-CEM
97.5%-SEM
2.5%-EM
2.5%-CEM
2.5%-SEM
0.88
0.90
0.92
0.94
0.96
0.98


LS
SSE
0.000
0.010
0.020
0.030
0.88
0.90
0.92
0.94
0.96
0.98
ρ
50%-EM
50%-CEM
50%-SEM
Ridge
97.5%-EM
97.5%-CEM
97.5%-SEM
LT_hkp
2.5%-EM
2.5%-CEM
2.5%-SEM
LT_itr


LS
SSE
ρ
50%-EM
50%-CEM
50%-SEM
Ridge
97.5%-EM
97.5%-CEM
97.5%-SEM
LT_hkp
2.5%-EM
2.5%-CEM
2.5%-SEM
LT_itr
0.0
0.2
0.4
0.6
0.8
1.0
0.88
0.90
0.92
0.94
0.96
0.98

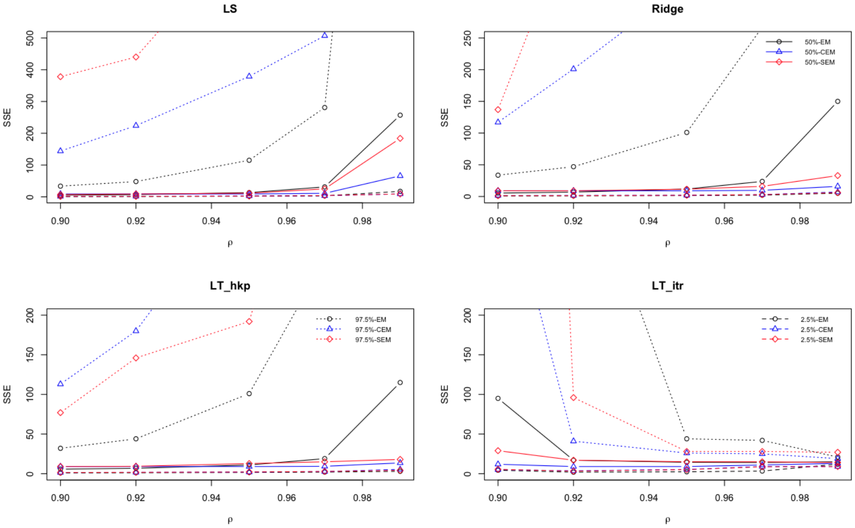

LS
Ridge
LT_hkp
LT_itr
SSE
ρ
0.90
0.92
0.94
0.96
0.98
50%-EM
50%-CEM
50%-SEM
97.5%-EM
97.5%-CEM
97.5%-SEM
2.5%-EM
2.5%-CEM
2.5%-SEM

Table 5: Prediction Error based on MRSEP for mixture of two regressions with $n = 100$

		$\rho = 0.88$		$\rho = 0.90$		$\rho = 0.95$		$\rho = 0.97$		$\rho = 0.99$	
EST	METH	M	L	M	L	M	L	M	L	M	L
ML	EM	16.6	8.2	16.9	8.8	17.7	9.0	18.1	8.9	18.5	9.3
	CEM	16.5	8.6	16.8	8.7	17.6	9.4	18.1	9.0	18.5	9.9
	SEM	16.6	8.5	16.9	8.9	17.7	9.0	18.0	9.4	18.5	9.8
Ridge	EM	16.6	8.8	16.8	8.6	17.7	9.1	18.1	9.2	18.6	9.4
	CEM	16.5	8.6	16.8	8.6	17.6	9.4	18.0	9.4	18.4	9.3
	SEM	16.5	8.7	16.7	8.5	17.8	9.0	18.0	9.3	18.5	9.4
LT(HKP)	EM	16.5	8.6	16.8	8.7	17.6	9.0	18.1	8.9	18.5	9.9
	CEM	16.4	8.8	16.8	8.5	17.6	8.9	18.1	9.4	18.4	9.6
	SEM	16.5	8.6	16.9	8.8	17.7	8.9	18.0	9.2	18.4	9.6
LT(ITE)	EM	16.6	8.5	16.9	8.4	17.5	9.3	17.9	9.3	18.2	9.0
	CEM	16.5	8.6	16.8	8.8	17.3	9.0	17.8	9.2	18.2	9.3
	ESM	16.6	8.4	16.8	8.5	17.4	8.8	17.8	9.7	18.2	9.5

Real Data Analysis: Osteoporosis

- We consider available BMD data on women aged 50 and over in the NHANES III (conducted on $\sim$ 30K American adults) data set as our population with size $n = 181$.
- We consider total BMD of the second examination as response. Arm and Botton circumference as two easy-to-measure predictors, $\rho = 0.81$
- Goodness of fit tests using the BIC criterion shows that a mixture of two regressions was the best fit.
- The performance of methods weas evaluated via 5 folds cross-validation with 2000 replicates.

Table 1: Estimation Performance of bone population with $n = 60$

		CEM			SEM			EM		
Methods	Ψ	M	L	U	M	L	U	M	L	U
ML	β	.010	.002	.165	.019	.003	.213	.018	.003	.134
	π	.333	.100	.366	.333	.183	.366	.218	.015	.365
	σ^2	.003	.000	.014	.006	..000	.014	.004	.000	.014
Ridge	β	.009	.002	.165	.013	.002	.166	.012	.002	.118
	π	.333	.100	.366	.333	.166	.366	.214	.019	.366
	σ^2	.003	.000	.014	.006	.000	.014	.004	.000	.014
LT(HKP)	β	.009	.002	.165	.010	.003	.183	.010	.003	.067
	π	.333	.100	.366	.333	.150	.366	.205	.013	.372
	σ^2	.003	.000	.014	.006	.000	.014	.004	.000	.016
LT(ITERATIVE)	β	.009	.002	.010	.009	.006	.011	.009	.007	.010
	π	.300	.100	.366	.350	.116	.566	.575	.032	.599
	σ^2	.002	.000	.014	.005	.000	.014	.003	.000	.009

Thanks!

- Ghanem, Hatefi and Usefi (2022). Unsupervised shrinkage methods for mixture of regressions.
- Ghanem, Hatefi and Usefi (2022). Liu-type Shrinkage Estimators for Mixture of Logistic Regressions: An Osteoporosis Study.
- Pearce and Hatefi (2021+). Multiple Observers Ranked Set Samples for Shrinkage Estimators.
- Liu (2003). Using liu-type estimator to combat collinearity. Communications in Statistics-Theory and Methods.
- Faria and Soromenho (2010). Fitting mixtures of linear regressions. Journal of Statistical Computation and Simulation.
- Alvandi and Hatefi (2021). Estimation of Ordinal Population with Multi-observer Ranked Set Samples. Statistical Methods in Medical Research.