# Divergence based inference for High dimensional GLMM

Lei Li

Department of Statistics
George Mason University

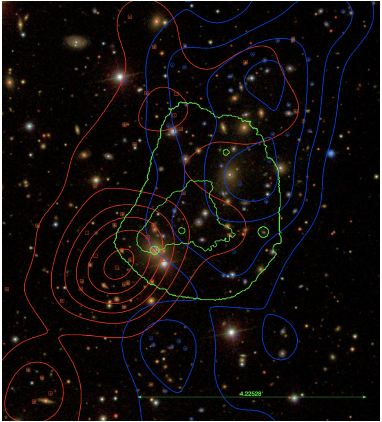
Joint work with Anand N. Vidyashankar

July 2019

# Contents



Merging Cluster Collaboration



# Background and Introduction

- Finite Mixture Regression (FMR) have broad application areas. Eg. Classification, clustering, network analysis, astronomy.
- Challenge:
  - 1. Identifiability.
  - 2. Estimation. Traditional (likelihood based) methods are basically not stable/robust.
- Estimation of FMR includes moment estimator, maximum likelihood estimation, minimum hellinger distance estimation, etc.
- One of the most popular methods is EM algorithm proposed by Dempster, Laird, and Rubin (1977).
- We propose to use minimum conditional divergence method for FMR.

# Finite Mixture Regression

## Definition

*A pair $(\boldsymbol{X}, Y)$ is said to follow an Finite Mixture Regression (FMR) model of order $K$ if the conditional density (or mass) function of $Y$ given $\boldsymbol{X} = \boldsymbol{x}$ is*

$$f(y|\boldsymbol{x}, \theta) = \sum_{k=1}^{K} \pi_k h(y; \lambda_k(\boldsymbol{x}), \xi_k), \tag{1}$$

*where $\pi_k$ are mixing probabilities with $\sum_{k=1}^{K} \pi_k = 1$, and $h(\cdot; \lambda_k(\boldsymbol{x}), \xi_k)$ belongs to a parametric family of density (or mass) functions, such that $\lambda_k(\boldsymbol{x}) = q(\beta_{k0} + \boldsymbol{x}^T \boldsymbol{\beta}_k)$ for a known link function $q(\cdot)$; $\beta_{k0}, \boldsymbol{\beta}_k = (\beta_{k1}, \beta_{k2}, \cdots, \beta_{kp})^T$, and $\xi_k$ are respectively the intercepts, regression coefficients and dispersion parameters.*

# Finite Mixture Regression Examples of GLMM

- Example 1: Poisson Mixture Regression.
  If we take $h(y; \lambda_k(\boldsymbol{x}), \xi_k) = Poi(\lambda_k(\boldsymbol{x}))$, where $\lambda_k(\boldsymbol{x}) = \exp(\beta_{0k} + \boldsymbol{x}^T \boldsymbol{\beta}_k)$ and $\theta_k = (\beta_{0k}, \beta_{1k}, \ldots, \beta_{pk})$.
- Example 2: Poisson Lognormal Mixture Regression.
  If we take $h(y; \lambda_k(\boldsymbol{x}), \xi_k) = Poi(\lambda_k(\boldsymbol{x}))$, and $\lambda_k(\boldsymbol{x}) = \exp(\beta_{0k} + \boldsymbol{x}^T \boldsymbol{\beta}_k + \epsilon_k)$, $\epsilon_k \sim N(0, \sigma_k^2)$, and $\xi_k = (\beta_{0k}, \beta_{1k}, \ldots, \beta_{pk}, \sigma_k^2)$.

# Minimum Divergence Estimation

For models with no regression:

- The population level general divergence between $g(\cdot)$ and $f(\cdot;\theta)$ is given by

$$D(\theta) = D(g(\cdot), f(\cdot;\theta)) \equiv \boldsymbol{E}_Y \left[ G\left( -1 + \frac{g(Y)}{f(Y;\theta)} \right) \right],$$

  where $G(\cdot)$ is a real valued thrice differentiable strictly convex function with $G(0) = 0$. Besides, $\delta(y;\theta) = -1 + \frac{g(y)}{f(y;\theta)}$ is called Pearson's residual between $g(y)$ and $f(y;\theta)$.

- The minimum divergence estimator is then given by

$$\hat{\theta}_{\text{MDE}} = \underset{\theta}{\text{argmin}}\ D_n(\theta), \quad \text{where} \quad D_n(\theta) = D(g_n(\cdot), f(\cdot;\theta)), \qquad (2)$$

  where $g_n(\cdot)$ is a nonparametric density estimate, one choice is kernel density estimate.

- See Basu, Shioya, and Park 2011 for a comprehensive description.

# Minimum Conditional Disparity Estimation

- Intuition: For models with no regression, note that

$$\begin{aligned} D(g(\cdot), f(\cdot;\theta)) &= \int_{\mathbb{R}} G(\delta(y;\theta)) f(y;\theta) dy \\ &= \int_{\mathbb{R}} \left(G(\delta(y;\theta)) + \delta(y;\theta)\right) \left(\frac{f(y;\theta)}{g(y)}\right) g(y) dy \\ &= \boldsymbol{E}_g \left[\frac{G(\delta(Y;\theta)) + \delta(Y;\theta)}{\delta(Y;\theta)+1}\right]. \end{aligned} \tag{3}$$

So one can approximate $D(g(\cdot), f(\cdot;\theta))$ as

$$D_n(g_n(\cdot), f(\cdot;\theta)) = \frac{1}{n}\sum_{i=1}^{n} \left(\frac{G(\delta_n(X_i;\theta)) + \delta_n(X_i;\theta)}{\delta_n(X_i;\theta)+1}\right). \tag{4}$$

- Given $\{(\boldsymbol{X}_i, Y_i)\}_{1\le i\le n}$, the minimum conditional disparity objective function is given by

$$D_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \left(\frac{G(\delta_n(Y_i|\boldsymbol{X}_i;\theta)) + \delta_n(Y_i|\boldsymbol{X}_i;\theta)}{\delta_n(Y_i|\boldsymbol{X}_i;\theta)+1}\right). \tag{5}$$

- The minimum conditional divergence estimator (MCDE) is then given by

$$\hat{\theta}_{\text{MCDE}} = \underset{\theta\in\Theta}{\operatorname{argmin}}\ D_n(\theta).$$

- Q: How to find $\hat{\theta}_{\text{MCDE}}$ for FMR?

# DivMin Algorithm

- Suppose that the pair $(Y, Z)$ has a joint density $p(y, z; \theta)$ that belongs to a parameterized family $\{p(\cdot, \cdot; \theta) : \theta \in \theta\}$, where $\theta \subseteq \mathbb{R}^d$. Only $Y$ is observed. Suppose that the true probability density of $Y$ is $g(\cdot)$, and is postulated as a parametric density $f(\cdot; \theta)$ for $Y$, where

$$f(y; \theta) = \int_{\mathcal{Z}} p(y, z; \theta) dz.$$

  In addition, let $k(z|y; \theta)$ denote the conditional density of $Z$ given $Y$.

- By incorporating the latent data structure, this new divergence for "complete" data $(Y, Z)$ is given by

$$Q(\theta'|\theta) = \boldsymbol{E}_Y \left[ \boldsymbol{E}_{Z|Y} \left[ G \left( -1 + \frac{g(Y)k(Z|Y;\theta)}{f(Y;\theta')k(Z|Y;\theta')} \right) \right] \right]. \tag{6}$$

- Similar to EM algorithm, the DivMin algorithm can be divided into two steps:
  - D-step. Determine $Q(\theta'|\theta)$.
  - M-step. Choose $\theta_{m+1} \in \theta$ so that it minimizes $Q(\theta'|\theta_m)$ over $\theta' \in \Theta$.

  These two steps are repeated until convergence.

# DivMin Algorithm Special Cases and Relation with Other Algorithms

- Special Cases:
  - 1. EM algorithm: Let $G(\delta) = (\delta + 1)\log(\delta + 1)$, $Q(\theta'|\theta)$ becomes the objective function obtained from E-step in the EM algorithm.
  - 2. HMIX algorithm: Taking $G(\delta) = 2[(\delta + 1)^{1/2} - 1]^2$, the corresponding objective function is

    $$Q_{\text{HD}}(\theta'|\theta) = 2\int_{\mathcal{Y}}\int_{\mathcal{Z}}\left[(g(y)k(z|y;\theta))^{\frac{1}{2}} - (f(y;\theta')k(z|y;\theta'))^{\frac{1}{2}}\right]^2 dzdy.$$

  - 3. VNEDMIX algorithm: Taking $G(\delta) = \exp\left(-\frac{1}{1+\delta} + 1\right)(1+\delta) - (2\delta + 1)$, we get the associated DivMin objective function as

    $$Q_{\text{VNED}}(\theta'|\theta) = \int_{\mathcal{Y}}\int_{\mathcal{Z}}\exp\left(-\frac{f(y;\theta')k(z|y;\theta')}{g(y)k(z|y;\theta)}\right)g(y)k(z|y;\theta)dzdy.$$

- DivMin algorithm belongs to the class of the following algorithms:
  - 1. MM algorithm.
    Note that $D(\theta') \leq Q(\theta'|\theta)$ and $D(\theta') = Q(\theta'|\theta')$ for all $\theta' \in \Theta$.
  - 2. Proximal Point algorithm.
  - 3. Coordinate Descent algorithm.

# DivMin Algorithm Properties

- Under mild regularity conditions, DivMin sequence is nonincreasing and finally converges to the stationary point of $D_n(\cdot)$.
- Similar to EM structure, generalized DivMin (GDM) algorithm, divergence conditional minimizaition (DCM) algorithm could also be derived.
- Q1. From Local minima to Global minima?
- Q2. If so, are we able to connect sample size to iteration size?
- $\hat{\theta}_{\mathrm{MCDE}}$ is (i) consistent, (ii) asymptotic normal, and (iii) robust.

# DivMin Algorithm for FMR

- The sample level DivMin objective function for FMR is given by

$$Q_n(\theta'|\theta) = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\left[\left(\frac{G\left(\delta_n(Y_i|\boldsymbol{X}_i;\theta,\xi_k')\right)+\delta_n(Y_i|\boldsymbol{X}_i;\theta,\xi_k')}{\delta_n(Y_i|\boldsymbol{X}_i;\theta,\psi_k')+1}\right)\tau_{ik}\right],$$

where

$$\delta_n(Y_i|\boldsymbol{X}_i;\theta,\psi_k') = -1+\frac{\tau_{ik}g_n(Y_i|\boldsymbol{X}_i)}{\pi_k' h(Y_i|\boldsymbol{X}_i;\xi_k')} \quad \text{and} \quad \tau_{ik} = \frac{\pi_k h(Y_i|\boldsymbol{X}_i;\xi_k)}{\sum_{l=1}^{K}\pi_l h(Y_i|\boldsymbol{X}_i;\xi_l)}.$$

- The estimation equation is

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\left[W_{ik}\tau_{ik}u(Y_i,\boldsymbol{X}_i;\eta_k')\right]=0, \quad \text{where} \quad W_{ik} = \left(\frac{A(\delta_n(Y_i|\boldsymbol{X}_i;\theta))+1}{\delta_n(Y_i|\boldsymbol{X}_i;\theta)+1}\right), \qquad (7)$$

and $u(Y_i,\boldsymbol{X}_i;\eta_k') = \nabla\log(\pi_k' f(Y_i|X_i;\xi_k'))$, $\eta_k' = (\pi_k',\xi_k')$.

- So the update for $\pi_k'$ is

$$\pi_k' = \frac{\sum_{i=1}^{n}W_{ik}\tau_{ik}}{\sum_{k=1}^{K}\sum_{i=1}^{n}W_{ik}\tau_{ik}}.$$

# DivMin Algorithm for FMR (contd.)

DivMin Algorithm:

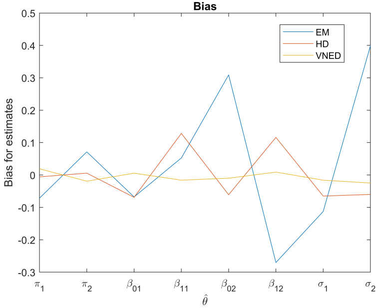
1. D-step: Update posterior probability $\tau_{ik}$ and weight $W_{ik}$.
2. M-step: Choose $\theta_{m+1} \in \theta$ so that it minimizes $Q_n(\theta'|\theta_m)$ over $\theta' \in \Theta$.

Remarks:

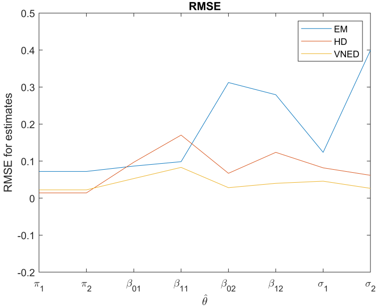
- This is a "weighted EM " structure.
- For EM algorithm, $W_{ik} \equiv 1$.
- For HMIX, $W_{ik} = \left( \frac{\pi_k h(Y_i|\mathbf{X}_i;\theta_k)}{\tau_{ik} g_n(Y_i|\mathbf{X}_i)} \right)^{\frac{1}{2}}$.
- Other special cases can be found in the paper.
- For finite linear mixture regression, close form expression can be obtained (detail skipped).
- We focus on finite mixture Poisson regression and/or Poisson regression with Lognormal random effects.

# Simulation: Finte Poisson Lognormal Model

Suppose that the underlying model is two component Poisson Lognormal (PL) model. We set $n = 2000$,
$\pi_1 = \pi_2 = 0.5, \beta_{01} = 0.6, \beta_{11} = 1.0, \beta_{02} = 2.5, \beta_{12} = 1.5, \sigma_1 = 0.25, \sigma_2 = 0.25$. In addition, 10% of $Y$ are replaced with outlier with value 100. We compare bias and root mean squared error (RMSE) of EM, HD, and VNED methods.




(a) Bias of Parameters




(b) RMSE of Parameters

# Concluding Remarks

- We proposed a robust minimum conditional divergence method for FMR.
- The DivMin algorithm is applied to estimate the parameters of FMR.
- Properties of DivMin algorithm have been investigated.
- Numerical study also supports the methodology.
- High Dimension (Lasso Type):

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left( \frac{G(\delta_n(Y_i|\boldsymbol{X}_i;\theta)) + \delta_n(Y_i|\boldsymbol{X}_i;\theta)}{\delta_n(Y_i|\boldsymbol{X}_i;\theta) + 1} \right) + \lambda \sum_{j=1}^{p} |\beta_j| \right\}.$$

# End

Thank you!