Robust mixture modeling using multivariate skew t distributions

Tsung-I Lin

Department of Applied Mathematics and Institute of Statistics

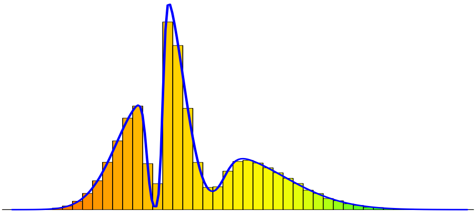
National Chung Hsing University, Taiwan

August 24, 2010

OUTLINE

1. Introduction

- Finite mixture models have become a useful tool for modeling data that are thought to come from several different groups with varying proportions.



- Lin et al. (2007) proposed a novel (univariate) skew t mixture (STMIX) model, which allows for accommodation of both skewness and thick tails for making robust inferences. Drawback: limited to data with univariate outcomes.
- We propose a multivariate version of the STMIX (MSTMIX) model, composed of a weighed sum of g-component multivariate skew t (MST) distributions.

The multivariate skew t (MST) distribution

- The MST distribution, $\boldsymbol{Y} \sim \mathcal{S}t_p(\boldsymbol{\xi}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda}, \nu)$, can be represented by

The stochastic representation of skew t distribution

$$\boldsymbol{Y} = \boldsymbol{\mu} + \frac{\boldsymbol{Z}}{\sqrt{\tau}}, \quad \boldsymbol{Z} \sim \mathcal{SN}_p(\boldsymbol{0}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda}), \quad \tau \sim \Gamma(\nu/2, \nu/2), \quad \boldsymbol{Z} \perp \tau \qquad (1)$$

- $\boldsymbol{Y} \mid \tau \sim \mathcal{SN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}/\tau, \boldsymbol{\Lambda}/\sqrt{\tau})$

Proposition 1.

If $\tau \sim \Gamma(\alpha, \beta)$, then for any $\boldsymbol{a} \in \mathbb{R}^p$

$$E\left(\Phi_p(\boldsymbol{a}\sqrt{\tau}|\boldsymbol{\Delta})\right) = T_p\left(\boldsymbol{a}\sqrt{\frac{\alpha}{\beta}} \,\middle|\, \boldsymbol{\Delta}; 2\alpha\right).$$

- Integrating τ from the joint density of $(\boldsymbol{Y}, \tau)$ yields

$$\psi(\boldsymbol{y}|\boldsymbol{\xi}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda}, \nu) = 2^p t_p(\boldsymbol{y}|\boldsymbol{\xi}, \boldsymbol{\Omega}, \nu) T_p\left(\boldsymbol{q}\sqrt{\frac{\nu+p}{U+\nu}}\,\middle|\boldsymbol{\Delta}; \nu+p\right), \qquad (2)$$

where $\boldsymbol{q} = \boldsymbol{\Lambda}\boldsymbol{\Omega}^{-1}(\boldsymbol{y}-\boldsymbol{\xi})$ and $U = (\boldsymbol{y}-\boldsymbol{\xi})^\top \boldsymbol{\Omega}^{-1}(\boldsymbol{y}-\boldsymbol{\xi})$.

$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}, \quad \boldsymbol{\lambda} = \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix}, \quad \nu = 4$$

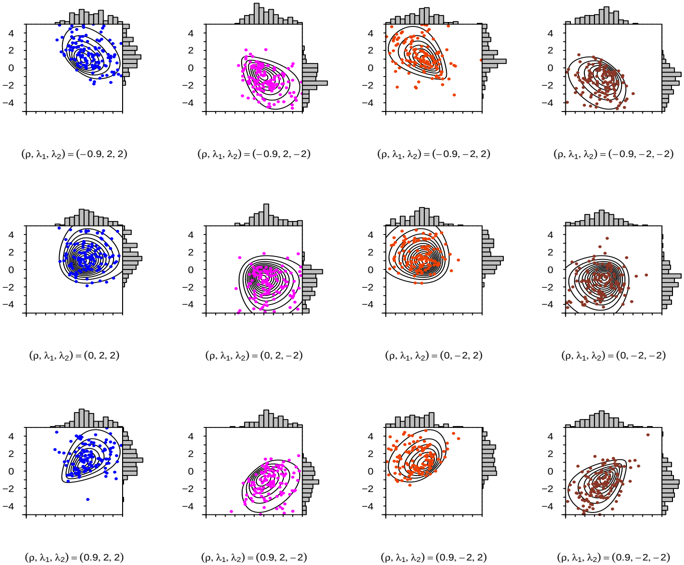


Figure 1: The scatter plots and contours and together with their histograms.

The MSTMIX model

- The MSTMIX model

$$f(\boldsymbol{y}_j|\boldsymbol{\Theta}) = \sum_{i=1}^{g} w_i \psi(\boldsymbol{y}_j \mid \boldsymbol{\xi}_i, \boldsymbol{\Sigma}_i, \boldsymbol{\Lambda}_i, \nu_i), \quad (3)$$

  where $\psi(\boldsymbol{y}_j|\boldsymbol{\xi}_i, \boldsymbol{\Sigma}_i, \boldsymbol{\Lambda}_i, \nu_i)$ represents the MST density, and w_i's are the mixing probabilities satisfying $\sum_{i=1}^{g} w_i = 1$.

- Introduce allocation variables $\boldsymbol{Z}_j = (Z_{1j}, \ldots, Z_{gj})^\top$, $j = 1, \ldots, n$, whose values are a set of binary variables with

$$Z_{ij} = \begin{cases} 1 & \text{if } \boldsymbol{Y}_j \text{ belongs to group } i, \\ 0 & \text{otherwise,} \end{cases}$$

  and satisfying $\sum_{i=1}^{g} Z_{ij} = 1$. Denoted by

$$\boldsymbol{Z}_j \sim \mathcal{M}(1; w_1, \ldots, w_g).$$

- A hierarchical representation of (3) is

$$\begin{aligned} \boldsymbol{Y}_j \mid (\gamma_j, \tau_j, Z_{ij}=1) &\sim \mathcal{N}_p(\boldsymbol{\xi}_i + \boldsymbol{\Lambda}_i\gamma_j, \boldsymbol{\Sigma}_i/\tau_j), \\ \gamma_j \mid (\tau_j, Z_{ij}=1) &\sim \mathcal{HN}_p(\boldsymbol{0}, \boldsymbol{I}_p/\tau_j), \\ \tau_j \mid (Z_{ij}=1) &\sim \Gamma(\nu_i/2, \nu_i/2), \\ \boldsymbol{Z}_j &\sim \mathcal{M}(1; w_1, \ldots, w_g). \end{aligned} \tag{4}$$

- The complete data log-likelihood function of $\boldsymbol{\Theta}$ is

$$\begin{aligned} &\ell_c(\boldsymbol{\Theta}|\boldsymbol{y}, \gamma, \boldsymbol{\tau}, \boldsymbol{Z}) \\ = \;& \sum_{i=1}^{g}\sum_{j=1}^{n} Z_{ij}\Big\{ \log(w_i) + \frac{\nu_i}{2}\log\left(\frac{\nu_i}{2}\right) - \log\Gamma\left(\frac{\nu_i}{2}\right) - \frac{1}{2}\log|\boldsymbol{\Sigma}_i| \\ & +(\frac{\nu_i}{2} + p - 1)\log\tau_j - \frac{\tau_j}{2}\Big((\boldsymbol{y}_j - \boldsymbol{\xi}_i - \boldsymbol{\Lambda}_i\gamma_j)^\top \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{y}_j - \boldsymbol{\xi}_i - \boldsymbol{\Lambda}_i\gamma_j) \\ & +\nu_i + \gamma_j^\top\gamma_j\Big)\Big\}. \end{aligned}$$

Computational aspects of parameter estimation

- The Q function is

$$Q(\boldsymbol{\Theta}|\hat{\boldsymbol{\Theta}}^{(k)}) = E(\ell_c(\boldsymbol{\Theta}|\boldsymbol{y}, \gamma, \boldsymbol{\tau}, \boldsymbol{Z})|\boldsymbol{y}, \hat{\boldsymbol{\Theta}}^{(k)}).$$

- In the MCEM-based algorithm, Q-function can be approximated by

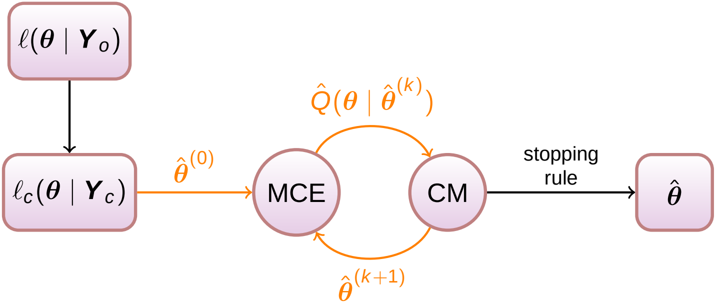
$$\hat{Q}(\boldsymbol{\Theta}|\hat{\boldsymbol{\Theta}}^{(k)}) = \frac{1}{M}\sum_{m=1}^{M} \ell_c(\boldsymbol{\Theta} \mid \boldsymbol{y}, \hat{\gamma}^{*(k)}_{[m]}, \hat{\boldsymbol{\tau}}^{*(k)}_{[m]}, \boldsymbol{Z}), \quad (5)$$

where $\hat{\gamma}^{*(k)}_{[m]} = \{\hat{\gamma}^{*(k)}_{ij,m}\}$ and $\hat{\boldsymbol{\tau}}^{*(k)}_{[m]} = \{\hat{\tau}^{*(k)}_{ij,m}\}$ are independently generated by

1. $\hat{\gamma}^{(k+1)}_{ij,m}|(\boldsymbol{y}_j, Z_{ij} = 1) \sim \mathcal{T}t_p\Big(\hat{\boldsymbol{q}}^{(k)}_{ij}, \frac{\hat{U}^{(k)}_{ij}+\hat{\nu}^{(k)}_i}{p+\hat{\nu}^{(k)}_i}\hat{\boldsymbol{\Delta}}^{(k)}_i, \hat{\nu}^{(k)}_i + p;\ \mathbb{R}^p_+\Big).$
2. $\hat{\tau}^{(k+1)}_{ij,m}|(\hat{\gamma}^{(k+1)}_{ij,m}, \boldsymbol{y}_j, Z_{ij} = 1)$

$$\sim \Gamma\Big(\frac{\hat{\nu}^{(k)}_i + 2p}{2}, \frac{(\hat{\gamma}^{(k+1)}_{ij,m} - \hat{\boldsymbol{q}}^{(k)}_{ij})^\top \hat{\boldsymbol{\Delta}}^{(k)-1}_i (\hat{\gamma}^{(k+1)}_{ij,m} - \hat{\boldsymbol{q}}^{(k)}_{ij}) + \hat{U}^{(k)}_{ij} + \hat{\nu}^{(k)}_i}{2}\Big).$$

The MCECM algorithm



arg max Q	*fix*
$\boldsymbol{\theta}_1$	$\hat{\boldsymbol{\theta}}_2^{(k)}$, $\hat{\boldsymbol{\theta}}_3^{(k)}$
$\boldsymbol{\theta}_2$	$\hat{\boldsymbol{\theta}}_1^{(k+1)}$, $\hat{\boldsymbol{\theta}}_3^{(k)}$
$\boldsymbol{\theta}_3$	$\hat{\boldsymbol{\theta}}_1^{(k+1)}$, $\hat{\boldsymbol{\theta}}_2^{(k+1)}$

CM-steps:

$$\hat{w}_i^{(k+1)} = n^{-1}\sum_{j=1}^{n}\hat{z}_{ij}^{(k)}$$

$$\hat{\boldsymbol{\xi}}_i^{(k+1)} = \frac{\sum_{j=1}^n \hat{\tau}_{ij}^{(k)}\boldsymbol{y}_j - \hat{\boldsymbol{\Lambda}}_i^{(k)}\sum_{j=1}^n \hat{\boldsymbol{\eta}}_{ij}^{(k)}}{\sum_{j=1}^n \hat{\tau}_{ij}^{(k)}}$$

$$\hat{\boldsymbol{\Lambda}}_i^{(k+1)} = \mathrm{diag}\Big\{\big(\hat{\boldsymbol{\Sigma}}_i^{(k)^{-1}}\odot\hat{\boldsymbol{B}}_{1i}^{(k)}\big)^{-1}\big(\hat{\boldsymbol{\Sigma}}_i^{(k)^{-1}}\odot\hat{\boldsymbol{B}}_{2i}^{(k)}\big)\boldsymbol{1}_p\Big\}$$

$$\hat{\boldsymbol{\Sigma}}_i^{(k+1)} = \frac{1}{\sum_{j=1}^n \hat{z}_{ij}^{(k)}}\Big(\sum_{j=1}^n \hat{\tau}_{ij}^{(k)}(\boldsymbol{y}_j-\hat{\boldsymbol{\xi}}_i^{(k+1)})(\boldsymbol{y}_j-\hat{\boldsymbol{\xi}}_i^{(k+1)})^\top + \hat{\boldsymbol{\Lambda}}_i^{(k+1)}\hat{\boldsymbol{B}}_{1i}^{(k)}\hat{\boldsymbol{\Lambda}}_i^{(k+1)} - \hat{\boldsymbol{\Lambda}}_i^{(k+1)}\hat{\boldsymbol{B}}_{2i}^{(k)} - \hat{\boldsymbol{B}}_{2i}^{(k)^\top}\hat{\boldsymbol{\Lambda}}_i^{(k+1)}\Big)$$

- Obtain $\hat{\nu}_i^{(k+1)}$ as the solution of

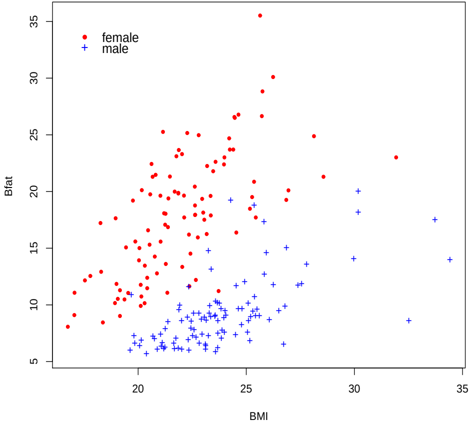
$$\log\Big(\frac{\nu_i}{2}\Big) + 1 - \mathrm{DG}\Big(\frac{\nu_i}{2}\Big) + \frac{1}{\sum_{j=1}^n \hat{z}_{ij}^{(k)}}\sum_{j=1}^n(\hat{\kappa}_{ij}^{(k)} - \hat{\tau}_{ij}^{(k)}) = 0.$$

- If the dfs are assumed to be identical, update $\hat{\nu}^{(k)}$ by

$$\hat{\nu}^{(k+1)} = \underset{\nu}{\mathrm{argmax}}\sum_{j=1}^n \log\Big(\sum_{i=1}^g \hat{w}_i^{(k+1)}\psi(\boldsymbol{y}_j \mid \hat{\boldsymbol{\xi}}_i^{(k+1)}, \hat{\boldsymbol{\Sigma}}_i^{(k+1)}, \hat{\boldsymbol{\Lambda}}_i^{(k+1)}, \nu)\Big).$$

The Australian Institute of Sport (AIS) data

- Data : The AIS data taken by Cook and Weisberg (1994).
- There are 202 athletes which include 100 females and 102 males.
- Variables : BMI (Body mass index; kg/m^2) and Bfat (Body fat percentage).



A two-component MSTMIX model can be written as

$$f(\boldsymbol{y}_j|\boldsymbol{\Theta}) = wf(\boldsymbol{y}_j|\boldsymbol{\xi}_1, \boldsymbol{\Sigma}_1, \boldsymbol{\Lambda}_1, \nu_1) + (1-w)f(\boldsymbol{y}_j|\boldsymbol{\xi}_2, \boldsymbol{\Sigma}_2, \boldsymbol{\Lambda}_2, \nu_2),$$

where

$$\boldsymbol{\xi}_i = (\xi_{i1}, \xi_{i2})^\top, \quad \boldsymbol{\Sigma}_i = \begin{bmatrix} \sigma_{i,11} & \sigma_{i,12} \\ \sigma_{i,12} & \sigma_{i,22} \end{bmatrix} \text{ and } \boldsymbol{\Lambda}_i = \begin{bmatrix} \lambda_{i,11} & 0 \\ 0 & \lambda_{i,22} \end{bmatrix}.$$

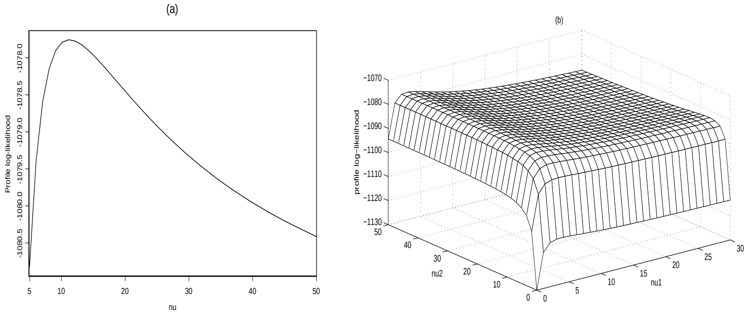


Figure 2: Plot of the profile log-likelihood for ν_1 and ν_2 with a two component MSTMIX model with (a) $\nu_1 = \nu_2 = \nu$ (b) $\nu_1 \neq \nu_2$. ($\hat{\nu}_1 = 4.2$, $\hat{\nu}_2 = 44.1$)

Table 1:Summary results from fitting various mixture models on the AIS data.

$\mathbf{\Theta}$	MVNMIX		MVTMIX		MSNMIX		MSTMIX	
	mle	se	mle	se	mle	se	mle	se
w	0.349	0.044	0.447	0.058	0.451	0.064	0.474	0.065
ξ_{11}	23.109	0.232	23.373	2.084	21.998	2.420	21.676	0.277
ξ_{12}	7.959	0.203	8.320	1.428	5.898	0.141	5.947	0.057
ξ_{21}	22.874	0.393	22.049	0.269	19.319	0.382	19.279	0.345
ξ_{22}	16.477	0.697	17.321	0.579	13.926	1.726	17.134	1.139
$\sigma_{1,11}$	2.878	0.700	3.791	0.873	3.178	2.988	2.730	0.392
$\sigma_{1,12}$	1.551	0.549	2.280	0.614	0.512	0.312	0.579	0.421
$\sigma_{1,22}$	2.111	0.662	3.158	0.573	0.114	0.115	0.140	0.975
$\sigma_{2,11}$	10.971	1.468	5.606	1.098	2.765	1.055	2.420	0.533
$\sigma_{2,12}$	4.946	2.081	6.589	1.839	7.141	2.145	7.047	1.122
$\sigma_{2,22}$	32.103	4.972	24.306	5.225	20.406	9.015	23.844	0.777
$\lambda_{1,11}$	—	—	—	—	1.163	3.223	1.615	0.326
$\lambda_{1,22}$	—	—	—	—	3.413	0.565	3.017	0.139
$\lambda_{2,11}$	—	—	—	—	4.805	0.448	4.192	1.789
$\lambda_{2,22}$	—	—	—	—	4.624	1.910	0.895	6.488
ν	—	—	5.820	1.646	—	—	11.041	5.207
m	11		12		15		16	
$\ell(\hat{\mathbf{\Theta}})$	-1097.790		-1093.585		-1080.647		-1077.760	
AIC	2217.581		2211.170		2191.293		2187.521	
BIC	2253.972		2250.870		2240.917		2240.453	

$\mathrm{AIC} = -2\,\ell(\hat{\mathbf{\Theta}}) + 2\,m$; $\mathrm{BIC} = -2\,\ell(\hat{\mathbf{\Theta}}) + m\log(n)$, $\ell(\hat{\mathbf{\Theta}})$ is the maximized log-likelihood, m is the number of parameters and n is the sample size.

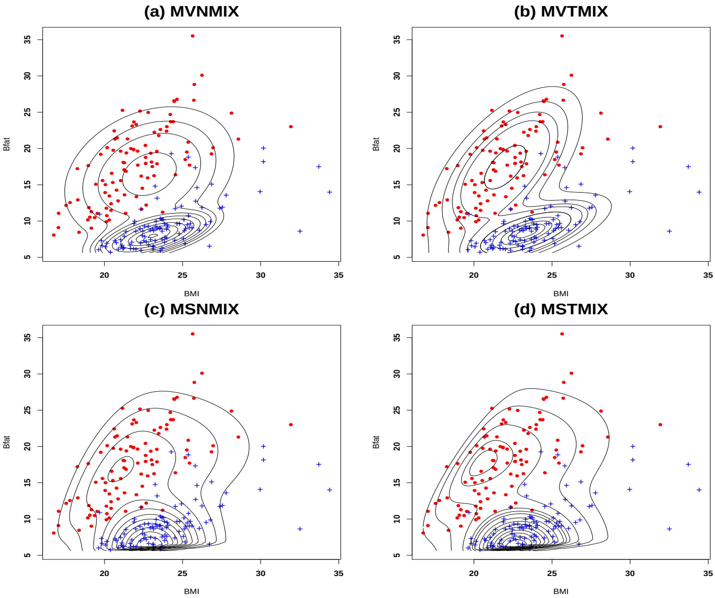


Figure 3: Scatter plot of BMI and Bfat with superimposed contours of two-component various models. The sex are indicated by the female (•) and male (+).

Concluding remarks

- Contributions:
  1. Propose a new robust the MSTMIX model, which offers a great deal of flexibility that accommodates asymmetry and heavy tails simultaneously.
  2. Allow practitioners to analyze heterogeneous multivariate data in a broad variety of considerations.
  3. MCEM-based algorithms are developed for computing ML estimates.
  4. Numerical results show that the MSTMIX model performs reasonably well for the experimental data.