`rfmm`: A Stata command for the Minimum Density Power Divergence estimation of finite mixtures of regression models

Federico Belotti⋆
Partha Deb◦

⋆ CEIS, University of Rome Tor Vergata
◦ Hunter College and NBER

2013 Italian Stata Users Group Meeting
Florence, November 14, 2013

Outline

Where we start from...

- Suppose $Y \sim N(\mu, 1)$

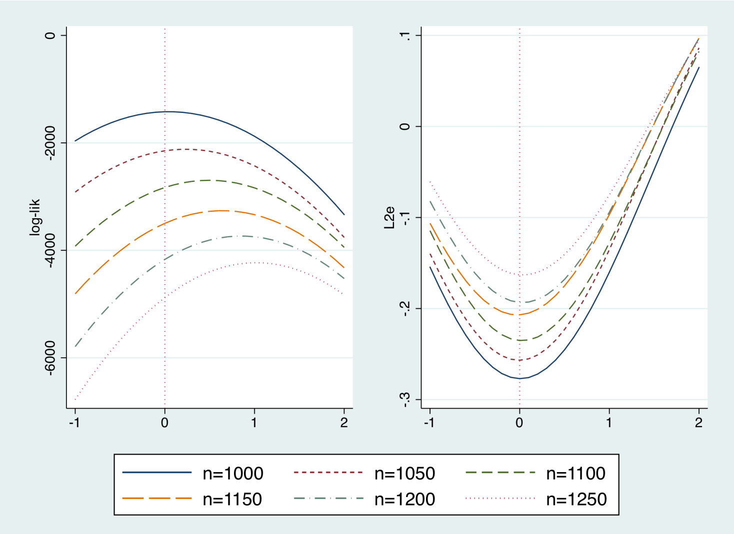
$$\hat{\mu}_{ML} = \underset{\mu}{\text{argmax}} \sum_{i=1}^{n} \log \phi(y_i|\mu, 1)$$

$$\mu_{L_2 e} = \underset{\theta}{\text{argmin}} \left[\frac{1}{2\sqrt{\pi}} - \frac{2}{n} \sum_{i=1}^{n} \phi(y_i|\mu, 1) \right]$$

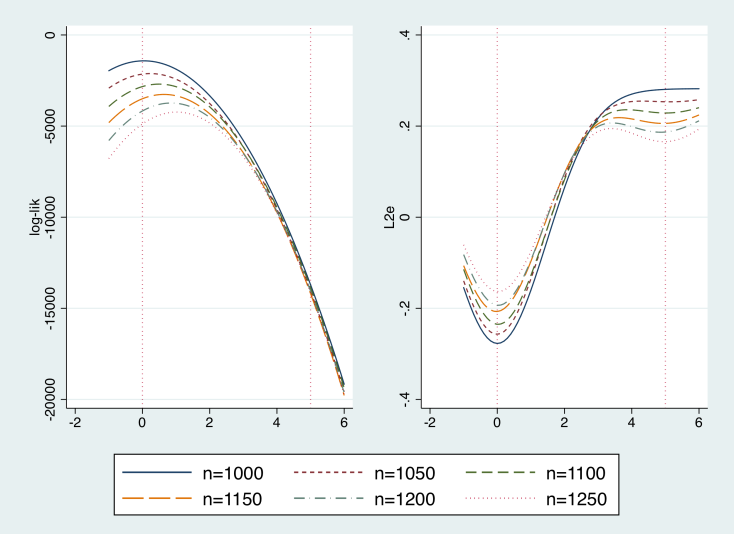
where $\phi(.)$ represents the Gaussian density function

- Consider a sample of size 1000 from $N(0, 1)$ with up to 300 additional observations sampled from a contamination density, $N(5, 1)$

Figure: Log-likelihood and L_2e criteria profiles



... L_2e illuminating behaviour ...



Finite mixtures of regression models

- Let $f_0(y|\boldsymbol{x})$ denote the true conditional density function of Y given $X = \boldsymbol{x}$ and $g_0(y, \boldsymbol{x}) = f_0(y|\boldsymbol{x}) f_X(\boldsymbol{x})$ denote the corresponding (true) joint density

- By assuming that $f_0(y|\boldsymbol{x})$ belongs to a parametric family $\mathcal{F}_m = \{f_{\boldsymbol{\theta}_m}(y|\boldsymbol{x}) : \boldsymbol{\theta}_m \in \boldsymbol{\Theta}_m \subseteq \boldsymbol{R}^p\}$ with $m < \infty$, a finite mixture of regression models can be defined as

$$f_{\boldsymbol{\theta}_m}(y_i|\boldsymbol{x}_i) = \sum_{j=1}^{m} \pi_j f_j(y_i|\boldsymbol{x}_i) \quad \text{with} \quad \boldsymbol{\theta}_m = (\boldsymbol{\pi}, \beta) \tag{1}$$

Common parametric families of probability distributions

$$f_{\boldsymbol{\theta}_m}(y_i|\boldsymbol{x}_i) = \sum_{j=1}^{m} \pi_j f_j(y_i|\boldsymbol{x}_i) \quad \text{with} \quad \boldsymbol{\theta}_m = (\boldsymbol{\pi}, \boldsymbol{\beta}) \tag{2}$$

Poisson $f_j(y_i|\mathbf{x}_i) = f_j(y_i|\lambda_{ij}) = e^{-\lambda_{ij}} \lambda_{ij}^{y_i} / y_i!$

with $\lambda_{ij} = exp(\mathbf{x}_i^T \boldsymbol{\beta}_j)$

NB $f_j(y_i|\mathbf{x}_i) = f_j(y_i|\mu_{ij}, \alpha_j) = \frac{\Gamma(y_i + \frac{1}{\alpha_j})}{\Gamma(y_i+1)\Gamma(\frac{1}{\alpha_j})} \left(\frac{\frac{1}{\alpha_j}}{\frac{1}{\alpha_j} + \mu_{ij}} \right)^{\frac{1}{\alpha_j}} \left(\frac{\mu_{ij}}{\frac{1}{\alpha_j} + \mu_{ij}} \right)^{y_i}$

with $\mu_{ij} = exp(\mathbf{x}_i^T \boldsymbol{\beta}_j)$ and $\alpha_j \geq 0$

Gaussian $f_j(y_i|\mathbf{x}_i) = f_j(y_i|\mu_{ij}, \sigma_j^2) = \frac{1}{\sigma_j \sqrt{2\pi}} \exp\left(- \frac{(\mathbf{y}_i - \mu_{ij})^2}{2\sigma_j^2} \right)$

with $\mu_{ij} = \mathbf{x}_i^T \boldsymbol{\beta}_j$

ML estimation

- Likelihood-based (ML) methods still remain the most widely used procedures to estimate finite mixtures of regression models
- Their strengths are: computational simplicity and asymptotic efficiency
- However, a well known drawback of ML is represented by its sensitivity to extreme values and/or components' contamination (see, among the others, Aitkin and Wilson (1980), Wang et al. (1996))

... robust alternatives ...

- Several robust approaches have been proposed to estimate finite mixtures of regression models. They exploit the intrinsic robustness of the minimum distance estimation framework

- Better performances by:

  1. Hellinger Divergence (Beran, 1977)

  2. Density-based Divergence (Basu et al., 1998)

- The L_2 estimator is a special case of the robust estimation approach proposed by Basu et al. (1998)

Minimum Hellinger Divergence (MHD)

- A robust estimator of $\boldsymbol{\theta}_m$ can be obtained by minimizing the integrated Hellinger divergence

$$\theta_m^{MHD} = \underset{\boldsymbol{\theta}_m}{\operatorname{argmin}} \int\int \left[\hat{f}_n(y|\boldsymbol{x})^{1/2} - f_{\boldsymbol{\theta}_m}(y|\boldsymbol{x})^{1/2}\right]^2 f_X(\boldsymbol{x}) d\boldsymbol{x} dy$$

where $\hat{f}_n(y|\boldsymbol{x})$ is a conditional non-parametric estimator of the true conditional density $f_0(y|\boldsymbol{x})$

- When applicable, MHD estimation allows to achieve efficiency in correctly specified models and robustness in presence of outliers. However:
  1. Computational complexity. Extensions are not straightforward for mixtures of regression models (see for instance Lu et al. (2003))
  2. Results strongly depend on the used conditional density estimators and related bandwidths
  3. Generally more sensitive to the choice of initial values than the MDPD approach (see Karlis and Xekalaki (1998))

Minimum Density Power Divergence (MDPD)

- A robust estimator of $\boldsymbol{\theta}_m$ can be obtained by minimizing the density-power divergence

$$\theta_m^{MDPD} = \underset{\boldsymbol{\theta}_m}{\operatorname{argmin}} \int \left[\int f_{\boldsymbol{\theta}_m}(y|\boldsymbol{x})^{1+\alpha} dy \right] f_X(\boldsymbol{x}) d\boldsymbol{x} - \left(1 + \frac{1}{\alpha}\right) n^{-1} \sum_{i=1}^{n} f_{\boldsymbol{\theta}_m}(y_i|\boldsymbol{x}_i) \quad (3)$$

- Even if not for mixture models, Basu et al. (1998) show that, by choosing a small value of α, MDPD estimation has strong robustness properties with a negligible loss in terms of efficiency relative to ML

- Lee and Sriram (2012) show that L_2 estimator (MDPD with $\alpha = 1$) is a very useful, attractive and viable alternative to the MHD for finite mixtures of Poisson or negative binomial regression models

Our contribution

1. We extend the L_2 estimator of Lee and Sriram (2012) to the MDPD estimation framework for finite mixtures of Poisson, NB-2 and Gaussian regression models
2. We investigate the properties of the proposed estimators through an extensive Monte Carlo study focusing on the Poisson distribution
3. We provide the new Stata command `rfmm`

MDPD estimation of finite mixtures of Poisson and NB2 regression models

- We propose to approximate $\int \left[\int f_{\boldsymbol{\theta}_m}(y|\boldsymbol{x})^{1+\alpha} dy\right] f_X(\boldsymbol{x}) d\boldsymbol{x}$ in equation (3) with $n^{-1}\sum_{i=1}^{n}\sum_{y=0}^{\infty} f_{\theta_m}^{1+\alpha}(y|\mathbf{x}_i)$

- We define the MDPD estimator for a finite mixture of count regression models the minimizer of the following divergence

$$\hat{\theta}_m^{MDPD} = \underset{\theta_m}{\operatorname{argmin}} \left[n^{-1} \sum_{i=1}^{n} \sum_{y=0}^{\infty} f_{\theta_m}^{1+\alpha}(y|\mathbf{x}_i) - (1+\frac{1}{\alpha}) n^{-1} \sum_{i=1}^{n} f_{\theta_m}^{\alpha}(y_i|\mathbf{x}_i) \right] \quad (4)$$

- Estimation is straightforward replacing $\sum_{y=0}^{\infty}$ with $\sum_{y=0}^{\max(y)}$

MDPD estimation of finite mixtures of Gaussian regression models

- As noted in Lee and Sriram (2012), there is no closed-form expression for the integral in equation (3) in the case of Gaussian mixtures of regression models

- Consider the following 2-components mixture of Gaussian regression models

$$f_{\boldsymbol{\theta}_m}(y_i|\boldsymbol{x}_i) = \pi N(y_i, \boldsymbol{\mu}_{i,1}, \sigma_1) + (1-\pi)N(y_i, \boldsymbol{\mu}_{i,2}, \sigma_2) \tag{5}$$

where $\boldsymbol{\mu}_{i,1} = \boldsymbol{x}_i'\beta_1$ and $\boldsymbol{\mu}_{i,2} = \boldsymbol{x}_i'\beta_2$

- When $\alpha = 1$, we have that (see Scott (2009))

$$\begin{aligned} n^{-1}\sum_{i=1}^{n}\int f_{\boldsymbol{\theta}_m}(y|\boldsymbol{x}_i)^2 dy &= \sum_{i=1}^{n}\frac{\pi^2}{2\sqrt{\pi}\sigma_1} + \frac{(1-\pi)^2}{2\sqrt{\pi}\sigma_2} + \\ &+ 2\pi(1-\pi)\phi(0|\boldsymbol{\mu}_{i,1} - \boldsymbol{\mu}_{i,2}, \sigma_1^2 + \sigma_2^2) \end{aligned} \tag{6}$$

MDPD estimation of finite mixtures of Gaussian regression models - 2

- Unfortunately, when $\alpha < 1$, we have

$$n^{-1} \sum_{i=1}^{n} \int \left[\pi N(y, \boldsymbol{\mu}_{i,1}, \sigma_1) + (1-\pi) N(y, \boldsymbol{\mu}_{i,2}, \sigma_2)\right]^{(1+\alpha)} dy \qquad (7)$$

  which has no closed-form

- We propose to numerically integrate (7) using Gauss-Hermite quadrature

- This strategy does not need any polynomial expansion before the numerical integration and it is easily extendable to mixtures with more than 2 components

The basic `rfmm` syntax is the following

`rfmm` *depvar* [*indepvars*] [*if*] [*in*] [*weight*] [, `options`]

`pweight`, `aweight`, `iweight` and `fweight` are allowed. Factor variables are not allowed (yet)

Options:
`mixtureof`(*distribution*) specifies the parametric family for the mixture. May be *normal*, *poisson* and *negbin2*. Default is *poisson*

`components`(#) specifies the number of mixture's components. Default is 2

`alpha`(#) specifies the value for the tuning parameter which controls the trade-off between robustness and efficiency. It must be $0 < \alpha \leq 1$. Default value is 0.5

`noconstant` suppresses the constant term for each component of the mixture

`cluster`(*varname*) adjust standard errors for intragroup correlation

`checkcomponents` draws the L2e criteria profile for the (unconditional) response variable `predict [type] newvar [if] [in] [, statistic equation(component#)]`

where `statistic` includes:

`mean`, the default, calculates the predicted mean. To obtain within class means, specify the `equation`(*component#*) option

`prior` calculates the prior component probabilities. With `prior`, `equation`(*component#*) must also be specified

`posterior` computes the posterior component probabilities. With `posterior`, `equation`(*component#*) must also be specified

Simulation study - Correctly specified models

- We consider as d.g.p. the following 2-components mixture of Poisson regression models

$$f_{\theta}(Y_i|X_i) = \pi_1 f(Y_i|\beta_{01} + \beta_{11}X_i) + (1-\pi_1)f(Y_i|\beta_{02} + \beta_{12}X_i) \tag{8}$$

  where X_i is taken to be a uniform $[0, 1]$, (β_{01}, β_{11}) and (β_{02}, β_{12}) represent respectively the 1^{st} and 2^{nd} component's parameters vector, and π_1 is the 1^{st} component mixing proportion

- Simulations are conducted through an "almost-full" factorial design controlling for: *i*) sample size $(900, 3600$ and $8100)$; *ii*) 1^{st} component's mixing proportion $\pi_1 = 0.1, 0.3, 0.5, 0.7, 0.9$; *iii*) a widely differing set of components parameters' values

Simulation study - Parameters' values and component's separation

(a) $\beta_{01} = 0, \beta_{11} = 0.5$

		β_{12}		
		0.75	**1.25**	**1.75**
β_{02}	**0.25**	0.5	0.9	1.4
	0.75	1.4	2.1	3.0
	1.25	3.0	4.1	5.5

(b) $\beta_{01} = 0, \beta_{11} = 0.5$

		β_{12}		
		1	**1.5**	**2**
β_{02}	**0.5**	1.1	1.7	2.5
	1	2.5	3.5	4.8
	1.5	4.8	6.4	8.5

- By varying the 2^{nd} component's parameters, we consider fourteen levels of components separation defined as

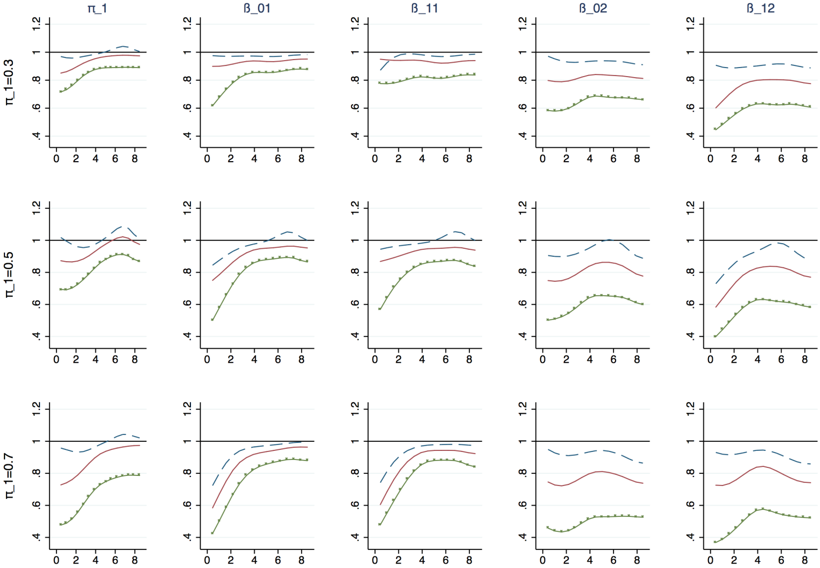
$$S = \frac{\exp\left(\beta_{02} + \beta_{12}\bar{X}_i\right) - \exp\left(\beta_{01} + \beta_{11}\bar{X}_i\right)}{\exp\left(\beta_{01} + \beta_{11}\bar{X}_i\right)}$$

  for which the simulated mixture's mean ranges between 1 and 8

- This set-up gives a total of 270 experiments. Each experiment is based on 100 converged replications

- MDPD estimates for different values of α (0.25, 0.5 and 1) are obtained through `rfmm` with BFGS algorithm and analytical gradient/hessian using ML estimates as starting values. The Newton-Rapson algorithm with analytical derivatives has been used instead for the ML estimation (using Partha Deb's `fmm`)

Relative efficiency - MSE(ML)/MSE(MDPD)



Simulation study - Contaminated models

- We assume that the data come from the following "contaminated" mixture of Poisson regression models

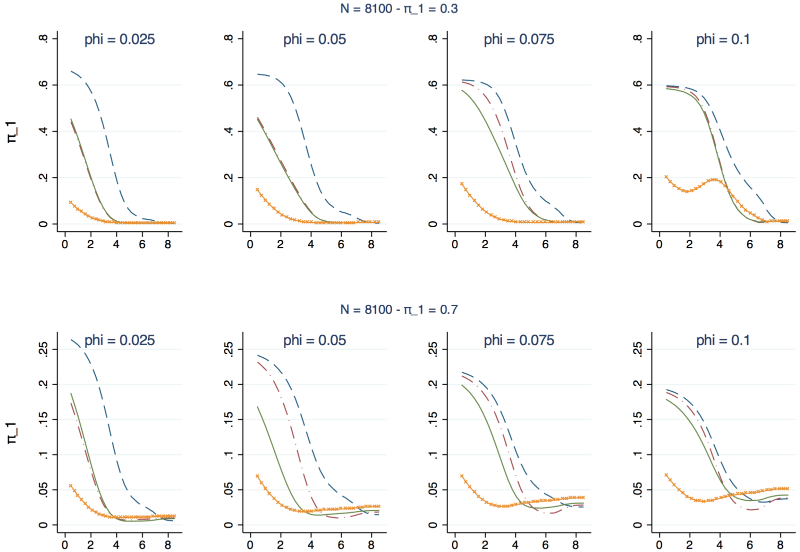
$$f_C(Y_i|X_i) = (1-\phi)f_\theta(Y_i|X_i) + \phi c(Y_i|X_i) \qquad (9)$$

  where the probability ϕ associated with the contaminant $c(Y_i|X_i)$, is chosen to cover a plausible set of contaminations ($\phi = 0.025, 0.05, 0.075, 0.1$)

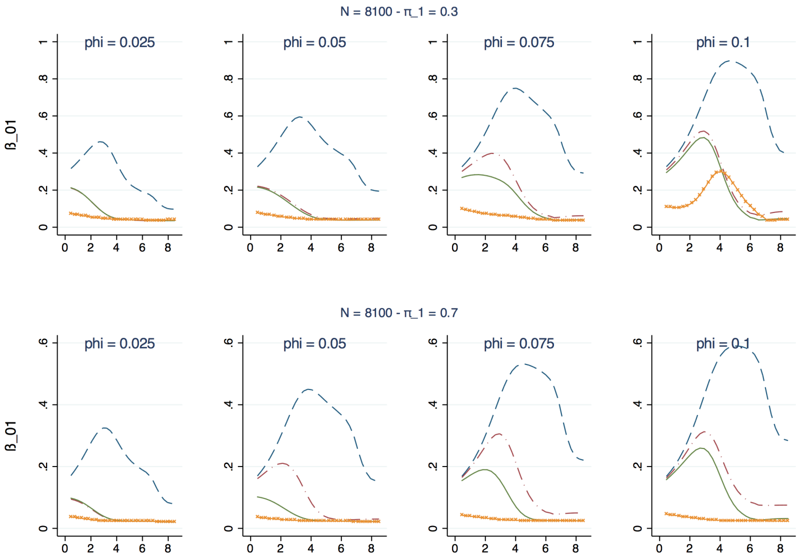
- In order to ensure a well defined notion of contamination π_1 was allowed to take only three values ($\pi_1 = 0.3, 0.5, 0.7$)

- We specify the contaminant to be Poisson distributed with a conditional mean of $g(Y_i; 3) = e^3 \simeq 20$

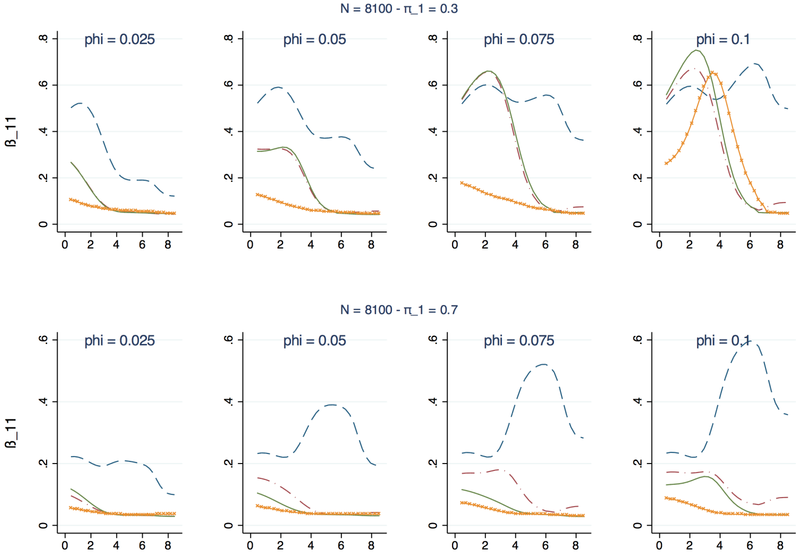
MEAD - π_1 (N = 8100)



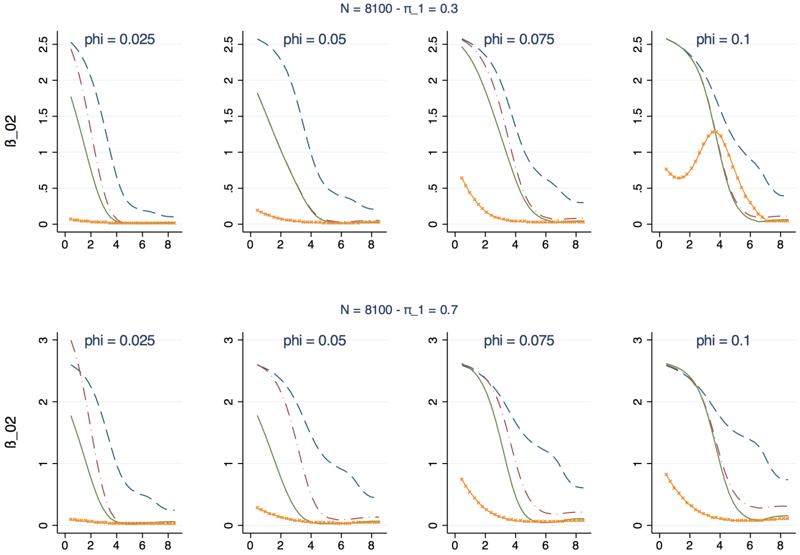
MEAD - β_{01} (N = 8100)



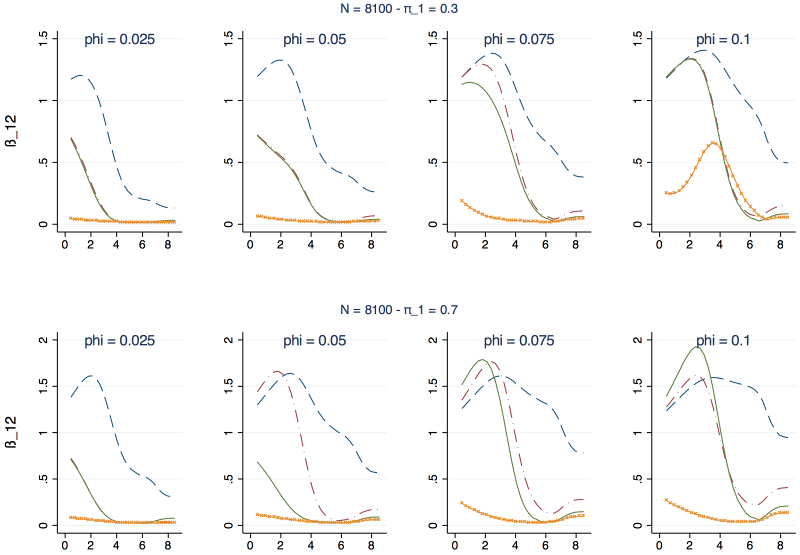
MEAD - β_{11} (N = 8100)



MEAD - β_{02} (N = 8100)



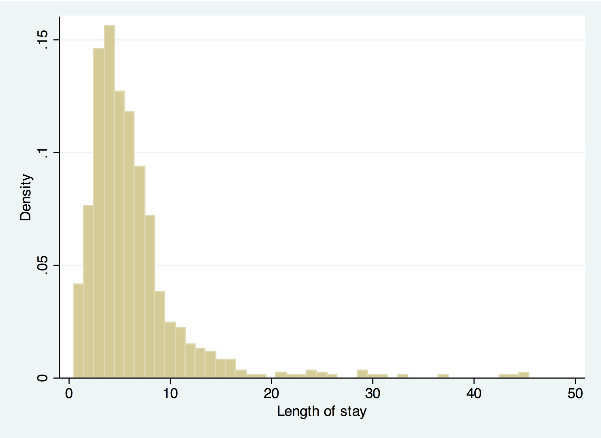
MEAD - β_{12} (N = 8100)



Empirical application - Maternity Length of Stay (LOS)

- Obstetrical LOS data drawn from the 1998/1999 Western Australia hospital morbidity database used in Lu et al. (2003) and Lee and Sriram (2012)
- The outcome variable (maternity LOS) is defined as the discrete count in days after delivery to discharge
- The majority of patients (97.5%) spent less than 20 days in hospital, with an average LOS of 6.24 (and a median of 5)

Figure: Hospital length of stay



	ML		MDPD $\alpha = 0.25$		MDPD $\alpha = 0.5$		MDPD $\alpha = 0.75$		MDPD $\alpha = 1$	
component 1										
not married	0.006		0.033		0.060		0.011		0.005	
emergency admitted	0.129	**	0.131	**	0.122	**	0.464	***	0.457	***
privately paid	0.167	*	0.227	***	0.257	***	0.225		0.224	
rural	0.108		0.129	**	0.148	**	0.189		0.215	*
employed	-0.068		-0.057		-0.066		0.074		0.057	
aboriginal	-0.018		-0.068		-0.102		0.069		0.047	
constant	1.610	***	1.534	***	1.497	***	1.442	***	1.442	***
component 2										
not married	-0.241		-0.158		0.056		0.150		0.172	
emergency admitted	0.200		0.342	**	0.276	*	-0.495		-0.519	
privately paid	-0.342		0.012		0.162		0.382	*	0.399	**
rural	0.086		-0.051		-0.058		-0.078		-0.130	
employed	-0.203		-0.020		-0.064		-0.251	*	-0.239	*
aboriginal	0.218		0.325	*	0.282		-0.367	***	-0.371	***
constant	2.971	***	2.419	***	2.202	***	1.776	***	1.766	***
π_1	0.073		0.132		0.186		0.380		0.369	
π_2	0.927		0.868		0.814		0.620		0.631	
μ_1	18.778		12.619		10.324		4.802		4.750	
μ_2	5.229		4.912		4.740		5.484		5.434	

Drop extreme values ($> 95^{th}$ or 99^{th} percentile)

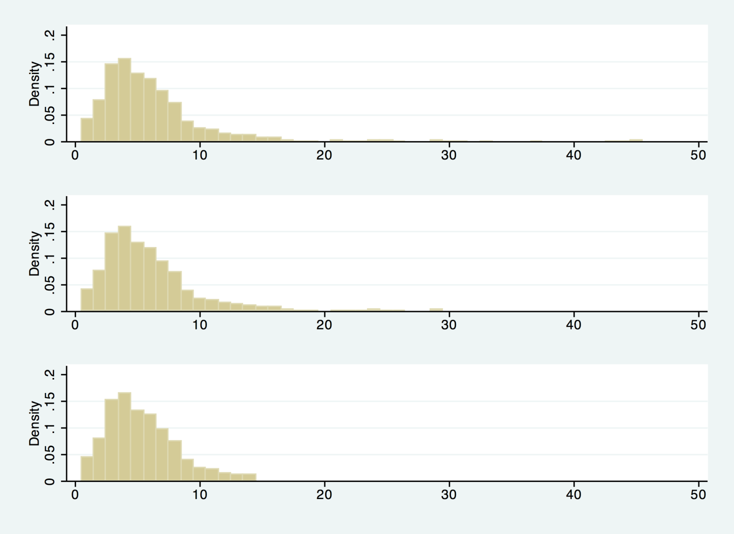
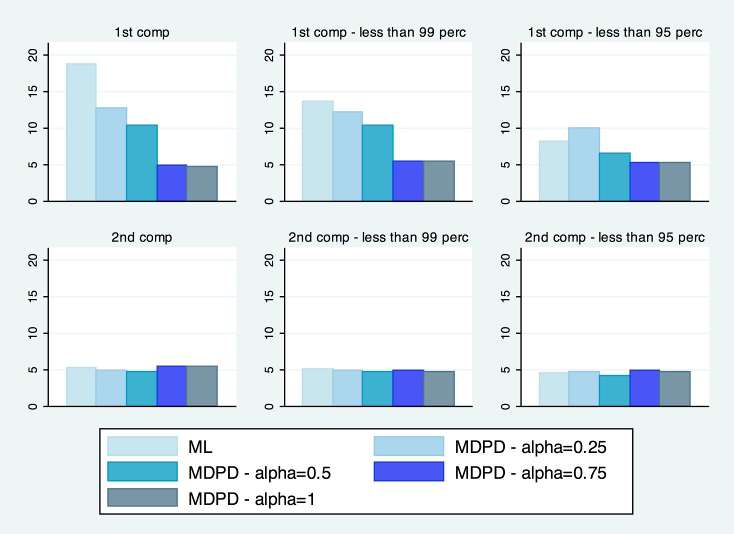


Figure: Predicted components' means



Aitkin, M. and Wilson, G. T. (1980). Mixture models, outliers, and the em alghoritm. *Technometrics*, 22:325–331.

Basu, A., Harris, I. R., Hjort, H. L., and Jones, M. C. (1998). Robust and efficient estimation by minimizing a density power divergence. *Biometrika*, 85:549–560.

Beran, R. (1977). Minimum hellingher distance estimation for parametric models. *The Annals of Statistics*, 5:445–463.

Karlis, D. and Xekalaki, E. (1998). Minimum hellingher distance estimation for poisson mixtures. *Computational Statistics and Data Analysis*, 29:81–103.

Lee, J. and Sriram, T. N. (2012). On the performance of l2e estimation in modelling heterogeneous count responses with extreme values. *Journal of Statistical Computation and Simulation*, pages 1–18.

Lu, Z., Hui, Y. V., and Lee, A. H. (2003). Minimum hellingher distance estimation for finite mixtures of poisson regression models and its applications. *Biometrics*, 59:1016–1026.

Scott, D. W. (2009). The l2e method. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1(1):45–51.

Wang, P. M., Puterman, M. L., Cockburn, I., and Le, N. (1996). Mixed poisson regression models with covariate dependent rates. *Biometrics*, 52:381–400.