Augmented Likelihood Estimators for Mixture Models

Markus Haas
Jochen Krause
Marc S. Paolella

Swiss Banking Institute, University of Zurich

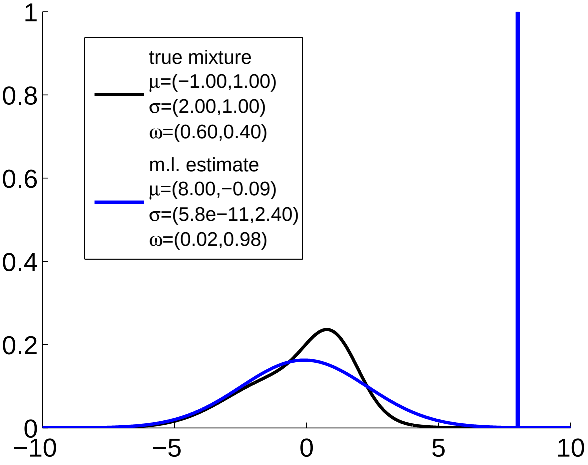
What is mixture degeneracy?

- mixtures under study are finite convex combinations of $1 \leq k < \infty$ (single-component) probability density functions

$$f_{\mathrm{MIX}}(\varepsilon; \boldsymbol{\theta}) = \sum_{i=1}^{k} \omega_i f_i(\varepsilon; \boldsymbol{\theta}_\mathbf{i})$$

- **unbounded** mixture likelihood function
- infinite likelihood values (**singularities**)
- mixture components degenerate to Dirac's delta function ▸ Delta Fun.
- maximum-likelihood estimation yields **degenerated estimates**
- set of local optima includes singularities

Why does degeneracy matter for mixture estimation?



mixture of two (e.g., normal) densities and exemplary m.l.e., $N = 100$

Selected literature on mixture estimation

- first occurrence of mixture estimation (method of moments)
  K. Pearson (1894)
- unboundedness of the likelihood function, e.g.
  J. Kiefer and J. Wolfowitz (1956); N. E. Day (1969)
- expectation maximization concepts for mixture estimation, e.g.
  V. Hasselblad (1966); R. A. Redner and H. F. Walker (1984)
- constraint maximum-likelihood approach, e.g.
  R. J. Hathaway (1985)
- penalized maximum-likelihood approach, e.g.
  J. D. Hamilton (1991); G. Ciuperca et al. (2003); K. Tanaka (2009)
- semi-parametric smoothed maximum-likelihood approach, e.g.
  B. Seo and B. G. Lindsay (2010)

Bib.

What is the contribution?

- **Fast, Consistent and General Estimation of Mixture Models**
  - fast: as fast as maximum-likelihood estimation (MLE)
  - consistent: if the true mixture is non-degenerated
  - general: likelihood-based, neither constraints nor penalties
- **Augmented Likelihood Estimation (ALE)**
  - shrinkage-like solution of the mixture degeneracy problem
  - approach copes with all kinds of local optima, not only singularities

A simple solution using the idea of shrinkage

augmented likelihood estimator: $\hat{\boldsymbol{\theta}}_{\text{ALE}} = \arg\max_{\boldsymbol{\theta}} \tilde{\ell}(\boldsymbol{\theta}; \boldsymbol{\varepsilon})$

augmented likelihood function:

$$\begin{aligned}
\tilde{\ell}(\boldsymbol{\theta}; \boldsymbol{\varepsilon}) &= \ell(\boldsymbol{\theta}; \boldsymbol{\varepsilon}) + \tau \sum_{i=1}^{k} \bar{\ell}_i(\boldsymbol{\theta}_i; \boldsymbol{\varepsilon}) \\
&= \sum_{t=1}^{T} \log \sum_{i=1}^{k} \omega_i f_i(\varepsilon_t; \boldsymbol{\theta}_i) + \underbrace{\tau \sum_{i=1}^{k} \frac{1}{T} \sum_{t=1}^{T} \log f_i(\varepsilon_t; \boldsymbol{\theta}_i)}_{\text{CLF}}
\end{aligned}$$

- number of component likelihood functions (CLF): $k \in \mathbb{N}$
- shrinkage constant: $\tau \in \mathbb{R}^+$
- geometric average of the ith likelihood function: $\bar{\ell}_i \in \mathbb{R}$

A simple solution using the idea of shrinkage

augmented likelihood estimator: $\hat{\boldsymbol{\theta}}_{\mathrm{ALE}} = \arg\max_{\boldsymbol{\theta}} \tilde{\ell}(\boldsymbol{\theta}; \boldsymbol{\varepsilon})$

augmented likelihood function:

$$\begin{aligned}
\tilde{\ell}(\boldsymbol{\theta}; \boldsymbol{\varepsilon}) &= \ell(\boldsymbol{\theta}; \boldsymbol{\varepsilon}) + \tau \sum_{i=1}^{k} \bar{\ell}_i(\boldsymbol{\theta}_i; \boldsymbol{\varepsilon}) \\
&= \sum_{t=1}^{T} \log \sum_{i=1}^{k} \omega_i f_i(\varepsilon_t; \boldsymbol{\theta}_i) + \underbrace{\tau \sum_{i=1}^{k} \frac{1}{T} \sum_{t=1}^{T} \log f_i(\varepsilon_t; \boldsymbol{\theta}_i)}_{\text{CLF}}
\end{aligned}$$

- CLF **penalizes** for small component likelihoods
- CLF **rewards** for high component likelihoods
- CLF identifies the ALE

A simple solution using the idea of shrinkage

augmented likelihood estimator: $\hat{\boldsymbol{\theta}}_{\text{ALE}} = \arg\max_{\boldsymbol{\theta}} \tilde{\ell}(\boldsymbol{\theta}; \boldsymbol{\varepsilon})$

augmented likelihood function:

$$\begin{aligned}
\tilde{\ell}(\boldsymbol{\theta}; \boldsymbol{\varepsilon}) &= \ell(\boldsymbol{\theta}; \boldsymbol{\varepsilon}) + \tau \sum_{i=1}^{k} \bar{\ell}_i(\boldsymbol{\theta}_i; \boldsymbol{\varepsilon}) \\
&= \sum_{t=1}^{T} \log \sum_{i=1}^{k} \omega_i f_i(\varepsilon_t; \boldsymbol{\theta}_i) + \underbrace{\tau \sum_{i=1}^{k} \frac{1}{T} \sum_{t=1}^{T} \log f_i(\varepsilon_t; \boldsymbol{\theta}_i)}_{\text{CLF}}
\end{aligned}$$

- **consistent** ALE as $T \to \infty$
- ALE $\to$ MLE, if $\tau \to 0$ or if $k = 1$
- separate component estimates for $\tau \to \infty$

How does the ALE work?

- assume **all mixture components** of the true underlying data generating mixture process as **non-degenerated**
- likelihood product is zero for **degenerated** components
- individual mixture components not prone to degeneracy
- prevent degeneracy by **shrinkage**
- shrink overall mixture likelihood function towards component likelihood functions

shrinkage term

$$CLF = \sum_{i=1}^{k} \tau_i \bar{\ell}_i \left(\boldsymbol{\theta}_i ; \varepsilon \right)$$

A short comparison, mixture of normals

IG p.d.f.

Penalized Maximum Likelihood Estimation, Ciuperca et al. (2003), Inverse Gamma (IG) Penalty:

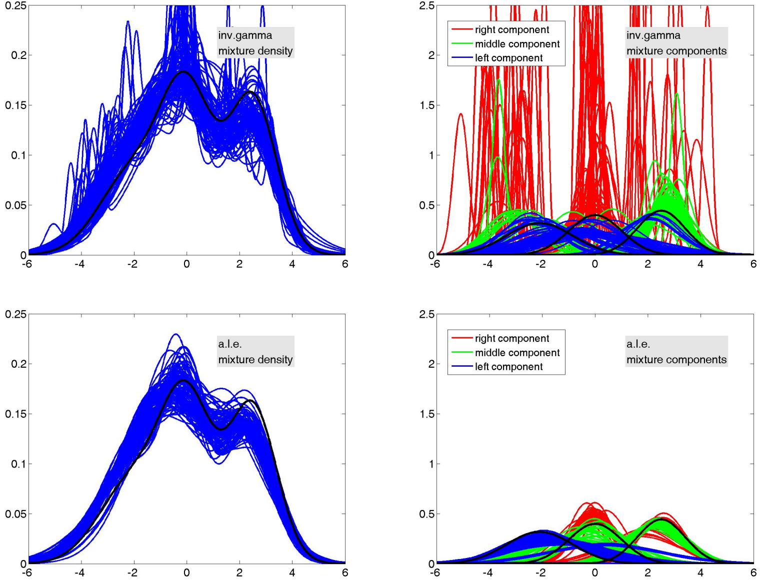
$$\ell_{\mathrm{IG}}(\boldsymbol{\theta};\boldsymbol{\varepsilon}) = \sum_{t=1}^{T} \log f_{\mathrm{MixN}}(\varepsilon;\boldsymbol{\theta}) + \sum_{i=1}^{k} \log f_{\mathrm{IG}}(\sigma_i; 0.4, 0.4)$$

Augmented Likelihood Estimator, $\tau = 1$:

$$\ell_{\mathrm{ALE}}(\boldsymbol{\theta};\boldsymbol{\varepsilon}) = \sum_{t=1}^{T} \log f_{\mathrm{MixN}}(\varepsilon;\boldsymbol{\theta}) + \sum_{i=1}^{k} \frac{1}{T} \sum_{t=1}^{T} \log f_i(\varepsilon_t;\boldsymbol{\theta}_i)$$

100 estimations, 500 simulated obs., random starts

▸ Details



Conclusion & Further Research

What is the contribution of ALE?

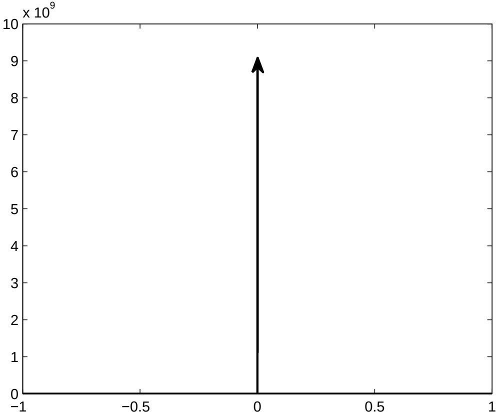
+ **solution** to the mixture degeneracy problem
+ very **simple implementation**
+ **no prior information** required, except for shrinkage constant(s)
+ purely based on likelihood values
+ applicable to mixtures of mixtures
+ gives **consistent** estimators
+ directly extendable to multivariate mixtures (e.g., for classification)
+ computationally feasible for out-of-samples exercises

- further research: trade-off between potential shrinkage bias and number of local optima as well as small sample properties

Augmented Likelihood Estimators for Mixture Models

Thank you for your attention!

What is a delta function?



probability density function with point support

Back

Bibliography I

- K. Pearson (1894)
  "Contributions to the Mathematical Theory of Evolution"
- J. Kiefer and J. Wolfowitz (1956)
  "Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters"
- V. Hasselblad (1966)
  "Estimation of Parameters for a Mixture of Normal Distributions"
- N. E. Day (1969)
  "Estimating the Components of a Mixture of Normal Distributions"
- R. A. Redner and H. F. Walker (1984)
  "Mixture Densities, Maximum Likelihood and the EM Algorithm"
- R. J. Hathaway (1985)
  "A Constrained Formulation of Maximum-Likelihood Estimation for Normal Mixture Distributions"

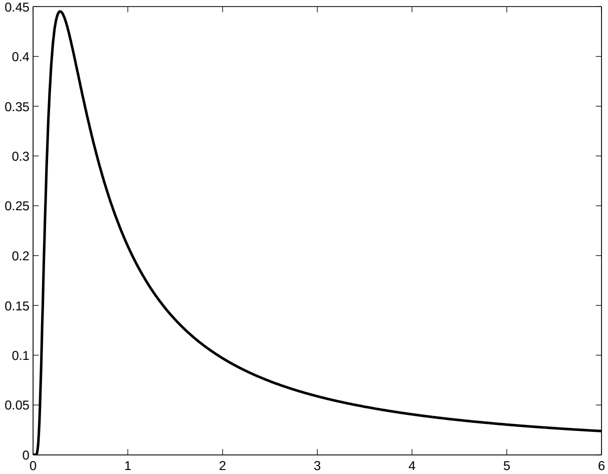
Back

Bibliography II

- J. D. Hamilton (1991)
  "A Quasi-Bayesian Approach to Estimating Parameters for Mixtures of Normal Distributions"
- G. Ciuperca, A. Ridolfi and J. Idier (2003)
  "Penalized Maximum Likelihood Estimator for Normal Mixtures"
- K. Tanaka (2009)
  "Strong Consistency of the Maximum Likelihood Estimator for Finite Mixtures of LocationScale Distributions When Penalty is Imposed on the Ratios of the Scale Parameters"
- B. Seo and B. G. Lindsay (2010)
  "A Computational Strategy for Doubly Smoothed MLE Exemplified in the Normal Mixture Model"

Back

Inverse Gamma Probability Density Function



Inverse Gamma p.d.f. as used in Ciuperca et al. (2003); $\alpha = 0.4$, $\beta = 0.4$.

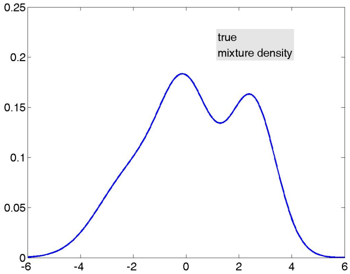
Back

Simulation Study - Details

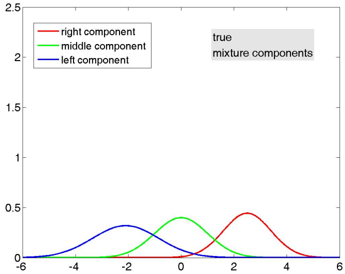
- number of simulations, 100
- initial starting values, uniformly drawn from hand-selected intervals
- hybrid optimization algorithm, BFGS, Downhill-Simplex, etc.
- maximal tolerance, 10^{-8}
- maximal number of function evaluations, $100'000$
- estimated mixture components, sorted in increasing order by σ_i

Back

Simulation Study - the true mixture density



mixture of three normals



mixture components

$$\boldsymbol{\theta}_{\text{true}} = (\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\omega}) = (2.5, 0.0, -2.1, \;\; 0.9, 1.0, 1.25, \;\; 0.35, 0.4, 0.25)$$

Back

Variance weighted extension

An extended augmented likelihood estimator:

$$\begin{aligned}
\ell_{\mathrm{ALE}}(\boldsymbol{\theta};\boldsymbol{\varepsilon}) &= \sum_{t=1}^{T} \log f_{\mathrm{MIX}}(\boldsymbol{\varepsilon};\boldsymbol{\theta}) \\
&+ \sum_{i=1}^{k} \log \left[\prod_{t=1}^{T} f_i(\varepsilon_t;\boldsymbol{\theta}_i)\right]^{\frac{1}{T}} \\
&- \sum_{i=1}^{k} \log \left[1 + \frac{1}{T}\sum_{t=1}^{T}\left(f_i(\varepsilon_t;\boldsymbol{\theta}_i) - \left[\prod_{t=1}^{T} f_i(\varepsilon_t;\boldsymbol{\theta}_i)\right]^{\frac{1}{T}}\right)^2\right]
\end{aligned}$$

This specific ALE not only enforces a meaningful (high) explanatory power for all observations, it also enforces a meaningful (small) variance of the explanatory power.