JOHANNES KEPLER
UNIVERSITY LINZ

# Prior Specifications for Finite Bayesian Mixture Models

Discussant: Bettina Grün

JSM 2018

# Finite Bayesian mixture models

- Types of applications include
  - Semi-parametric density approximation.
  - Model-based clustering.
- Areas of applications are numerous.
- Many possible extensions and variants possible taking specific data structures into account.
- The finite mixture model is given by

$$\boldsymbol{y}_i \sim \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{y}_i|\boldsymbol{\theta}_k).$$

# Prior distributions

- A prior needs to be specified for the full set of parameters consisting of:
  - the component weights $\pi_k$, $k = 1, \ldots .K$;
  - the component-specific parameters $\boldsymbol{\theta}_k$, $k = 1, \ldots, K$;
  - the parameters for hyperpriors $\boldsymbol{\vartheta}$.
- The prior is given by

$$p(\boldsymbol{\pi}, \boldsymbol{\Theta}, \boldsymbol{\vartheta}),$$

with $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K\}$.

# Prior characteristics

- No conjugate prior for mixture models is available.
- In general proper priors are required to obtain proper posteriors.
- Assume prior independence between component weights and component-specific parameters.
- Use exchangeable or conditionally exchangeable prior for component weights and / or component-specific parameters.
- Use conditionally conjugate priors given component memberships.

# Informative versus non-informative priors

- The mixture likelihood is known to be prone to have
  - Multiple (spurious) modes and
  - Be unbounded at the boundary of the parameter space.
- Model-based clustering is an ill-posed problem:
  - The data might be compatible with several different cluster structures.
  - Certain cluster structures might not be of interest.
- For model-based clustering (weakly) informative priors are often advocated because they allow to
  - Include prior knowledge about cluster structure;
  - Regularize the likelihood.

# Kiefer-Wolfowitz example

- We consider the following mixture of two normal distributions:

$$p(y|\eta_2, \mu, \sigma_2^2) = (1 - \eta_2) f_{\mathcal{N}}(y_i|\mu, 1) + \eta_2 f_{\mathcal{N}}(y|\mu, \sigma_2^2).$$

- $\eta_2$ is assumed fixed with

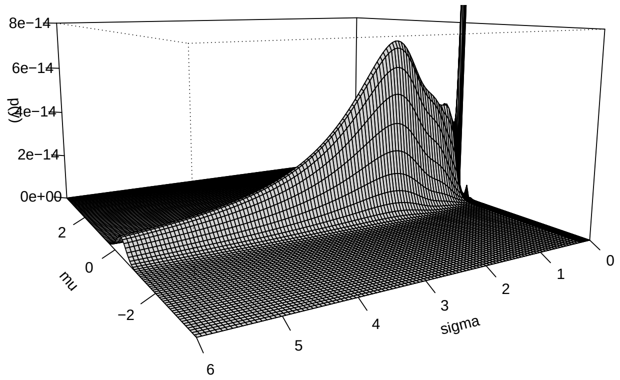
$$\eta_2 = 0.2$$

  and $\mu$ and $\sigma_2^2$ are unknown.
- Kiefer and Wolfowitz (1956) used that as an example to show that each observation in an arbitrary data set of arbitrary size $N$ generates a singularity in the mixture likelihood function.
- We simulated $N = 20$ observations from the model with

$$\mu = 0 \qquad\qquad \sigma_2^2 = 4.$$

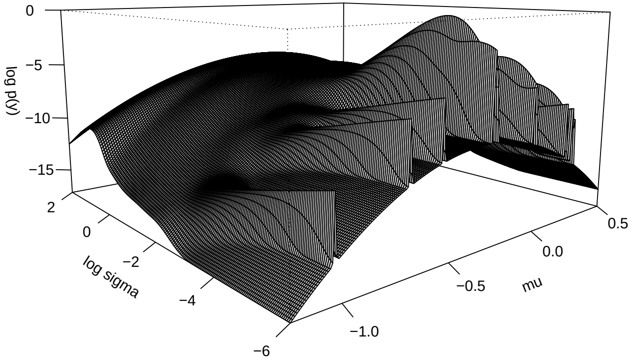
# Kiefer-Wolfowitz example / 2

Likelihood of the data:



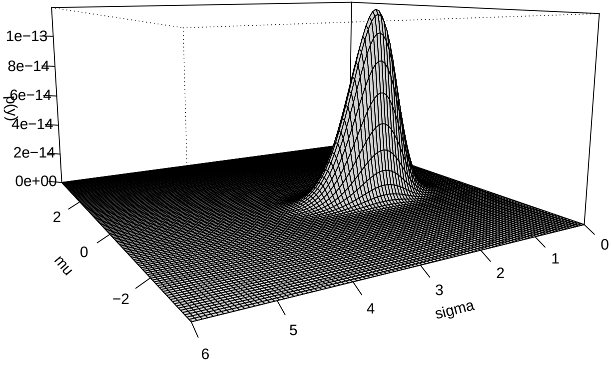
# Kiefer-Wolfowitz example / 3

Log-likelihood of the data with respect to $\log(\sigma)$:



# Kiefer-Wolfowitz example / 4

Non-normalized posterior for prior $\sigma^2 \sim \mathcal{G}^{-1}(2.5, 1)$:



# Genuine multimodality

- We consider the following mixture of two normal distributions:

$$p(y|\eta_2, \mu_1, \mu_2) = (1 - \eta_2) f_{\mathcal{N}}(y_i|\mu_1, 1) + \eta_2 f_{\mathcal{N}}(y|\mu_2, 1).$$

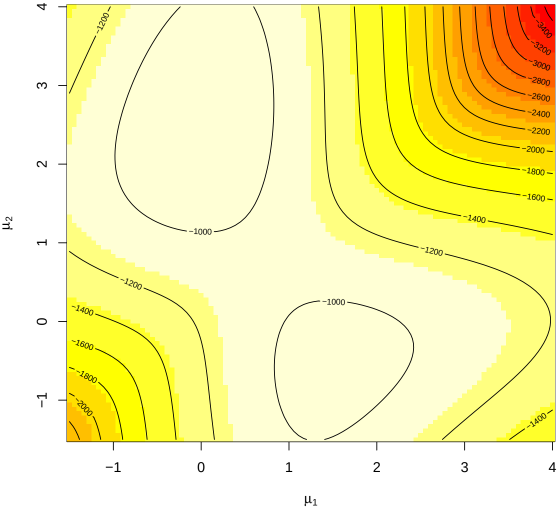
- $\eta_2$ is assumed fixed with

$$\eta_2 \neq 0.5.$$

- Marin et al. (2005) used that as an example to indicate genuine multimodality. Label switching is not an issue for $\eta_2 \neq 0.5$ known.
- We simulated $N = 500$ observations from the model with

$$\mu_1 = 0, \qquad \mu_2 = 2.5, \qquad \eta_2 = 0.3.$$

# Genuine multimodality / 2



# Non-informative priors for univariate mixtures

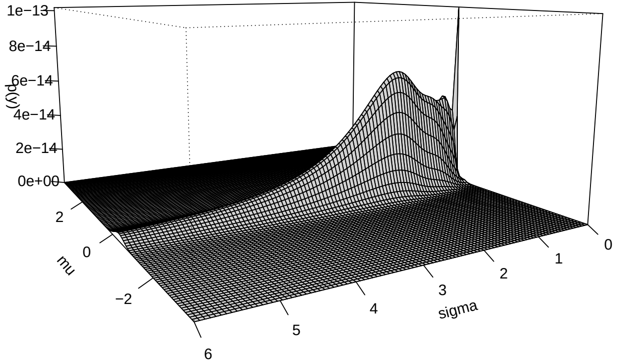
- Jeffreys priors for mixtures most often lead to improper posteriors.
- Jeffreys priors for the weights conditionally on the parameter mixture components are derived.
    - Assumption that priors on weights and component parameters are independent is not supported.
    - Richardson and Green (1997) use RJMCMC for univariate Gaussian mixtures. They note that:
        - A prior allowing for higher component specific variances leads to more weight on more components a-posteriori.
        - Reducing the variance of a flat prior on the component means leads first to an increase of the number of components, but then decreases their number.

# Non-informative priors for univariate mixtures / 2

- Reparametrization of the Gaussian mixture models allows for
  - Improper priors for the overall mean and variance.
  - Restricts the parameters defining the components to a compact sense.
- The Gaussian mixture likelihood is unbounded for zero component specific variances with mean equal to an observation and contains spurious modes:
  - Standard inverse gamma priors on the component specific variances eliminates these spurious modes.
  - Uniform priors for the reparametrization might not have this effect.

# Non-informative priors for univariate mixtures / 3

Reparametrized prior with $\mu, \sigma \propto 1/\sigma$ and uniform priors on the component-specific parameters:



# Anchored Bayesian Gaussian mixture models

- Fixing the component memberships of data points to solve the label switching problems has been previously proprosed and used also in a post-processing step.
- The proposed approach allows to select these points in an automatic and principled way not relying on manual selection.
- Selecting the number of anchor points used seems to require additional investigations with preliminary results available.
- Anchoring requires the number of components to be known a-priori.
- Diebolt and Robert (1994) indicated that improper priors in the mixture case might only be used in case only partitions are considered with a sufficient number of observations assigned to them. Anchoring points ensures a certain minimum number of observations.

# Anchored Bayesian Gaussian mixture models / 2

- Fixing component memberships assumes that the prior and posterior probabilities of some observations to be in the same component is deterministically either zero or one.
  - Resolving label switching is needed in model-based clustering applications where grouping structure is implicitly assumed to be present.
  - If the influence of this prior anchoring is strong in addition to remove label switching, no clear grouping structure might be present.
- Multi-modality of the mixture likelihood after resolving label switching is known to be an issue. These “genuine” different modes might correspond to different grouping structures requiring different anchoring points.

# Heterogeneous reciprocal graphical models

- Model developed for a specific application with known groups is adapted to be used with latent groups.
- Component model is specific to the application and uses specifically designed priors to induce sparsity.
- Non-local priors with thresholding are used to induce sparsity:
  - How is the threshold selected?
  - How do these priors perform compared to other sparsity priors?
- In principle the same model and priors are uesd regardless of if the groups are manifest or latent.
  - Difference in assumptions on dependency structure between groups.
  - This indicates more structure required to identify latent groups.

# Heterogeneous reciprocal graphical models / 2

- In the finite mixture case a Dirichlet distribution for the weights is combined with a geometric distribution for the number of components $K$:
  - How are the parameters for the Dirichlet selected?
  - What is the influence of the distribution used for the number of components?
- Finite as well as infinite mixtures are considered:
  - How can their performance be compared?
  - Could priors be selected to match their behavior?

# Summary

- Choice of priors depends on application type.
- Models derived for heterogeneous populations with known groups can be used assuming unobserved heterogeneity. Slight adaptations to the priors needed to induce identifiability.
- Standard approaches assume independent priors between component weights and component specific distributions.
- Theoretical and empirical results suggest that there is dependency between these sets of parameters.
- Rousseau and Mengersen (2011) suggest that the parameter for the prior on the weights needs to be selected depending on the dimensionality of the component specific parameters / the data. More work needed to identify how parameter choices depend on the dimensionality of the data for other priors.
- More insights needed to the impact of prior choices on results and to guide their choice.

# References


J. Diebolt and C. P. Robert. Estimation of finite mixture distributions through Bayesian sampling. **Journal of the Royal Statistical Society B**, 56: 363–375, 1994.

J. Kiefer and J. Wolfowitz. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. **The Annals of Mathematical Statistics**, 27(4):887–906, 1956.

J.-M. Marin, K. Mengersen, and C. P. Robert. Bayesian modelling and inference on mixtures of distributions. In D. Dey and C. Rao, editors, **Bayesian Thinking: Modeling and Computation**, volume 25 of **Handbook of Statistics**, chapter 16, pages 459–507. North–Holland, Amsterdam, 2005.

S. Richardson and P. J. Green. On Bayesian analysis of mixtures with an unknown number of components. **Journal of the Royal Statistical Society B**, 59(4):731–792, 1997.

J. Rousseau and K. Mengersen. Asymptotic behaviour of the posterior distribution in overfitted mixture models. **Journal of the Royal Statistical Society B**, 73(5):689–710, 2011.