Label switching in mixtures

Christophe Biernacki[a,c], Vincent Vandewalle[b,c]
[a]**Laboratoire Paul Painlevé** (Université Lille 1 - CNRS)
[b]**Équipe d'Accueil 2694** (Université Lille 2)
[c]**Équipe MODAL** (INRIA Lille Nord Europe)

The label switching problem

Mixture of g distributions

$$p(\cdot|\theta) = \sum_{k=1}^{g} \alpha_k p(\cdot|\beta_k)$$

- α_k: mixtures weights ($\alpha_k > 0$ and $\sum_k \alpha_k = 1$)
- β_k: parameters of each component distribution
- $\theta_k = (\alpha_k, \beta_k)$
- $\theta = (\theta_1, \ldots, \theta_g) \in \Theta$

Generative interpretation

$x = (x_1, \ldots, x_n)$ an n i.i.d. sample from $p(\cdot|\theta)$
$z = (z_1, \ldots, z_n) \in \mathcal{Z}$ is the *latent partition* which as been used to generate x

- $z_i \sim \mathcal{M}(1; \alpha_1, \ldots, \alpha_g)$
- $x_i \sim p(\cdot|\beta_{z_i})$

Bayesian framework

- $p(\theta)$ a prior distribution on θ
- Bayesian inference is based on the posterior distribution $p(\theta|x) \propto p(x|\theta)p(\theta)$

The problem

If $p(x|\theta)$ and $p(\theta)$ are invariant up to a mixture component renumbering then so does $p(\theta|x)$

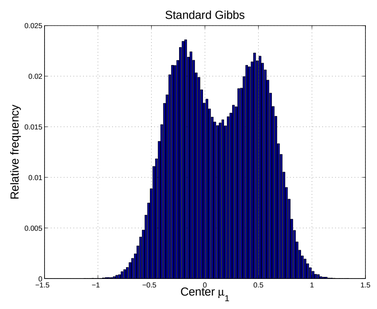
- $\mathcal{P}_g$ is the set of $\{1, \ldots, g\}$ permutations
- $\sigma(\theta) = (\theta_{\sigma(1)}, \ldots, \theta_{\sigma(g)})$ is the parameter θ permuted in index with $\sigma \in \mathcal{P}_g$

$$p(\theta|x) = p(\sigma(\theta)|x)$$

This exact symmetry is called *label switching*. It then makes meaningless direct computation of many usual punctual estimators as the posterior mean. The aim of many approaches is to remove this symmetry.

Illustration of the problem

- Two univariate components ($g = 2$)
- $p(\cdot|\beta_k) = \mathcal{N}(\beta_k, 1)$
- $\alpha_1 = \alpha_2 = 0.5$
- Proportions and variances known and fixed
- Mean $\theta_1 = \beta_1$ et $\theta_2 = \beta_2$ unknown
- Prior distributions on $\theta_k \sim \mathcal{N}(0, 1)$ with $\theta_1 \perp \theta_2$
- Posterior distributions
  - $\theta_k|z, x \sim \mathcal{N}(n_k\bar{x}_k/(n_k+1), 1/(n_k+1))$
  - $z_i|\theta, x \sim \mathcal{M}_2(1, t_{i1}(\theta), t_{i2}(\theta))$
  - $n_k = \sum_{i=1}^n \mathbb{I}_{z_i=k}$
  - $\bar{x}_k = \sum_{i=1}^n \mathbb{I}_{z_i=k} x_i/n_k$
  - $t_{ik}(\theta) = p(z_i = k|x, \theta)$
- $\theta_1 = 0$ and $\theta_2 = 0.25$



Posterior distribution $p(\theta_1|x)$.

Two modes can be seen on the posterior distribution of θ_1 when only one would be expected in absence of label switching. It is then impossible to make relevant analysis of the posterior distributions component-wise.

Standard solutions

Modified prior distribution

- Artificial identifiability constraints on the parameters (Diebolt et Robert, 1994)
- Ordering constraints : $\theta_1 < \theta_2$
- Modification of the prior distribution which becomes proportional to $p(\theta)\mathbb{I}_{\theta_1<\theta_2}$
- Not enough to solve the label switching problem (Celeux et al., 2000 ; Jasra *et al.*, 2005)

$k-means$ algorithm on the parameters space

- Relabeling algorithm of the generated parameters (Stephens, 1997 ; Celeux, 1998)
- Find the permutation for the fixed parameter which minimizes a loss function
- $k-means$ type algorithm on the parameters space
- Underestimation of the dispersion of the posterior distribution

Invariant loss function

- Loss function invariant up to the parameters permutation (Celeux et *al.*, 2000)
- Choice of a loss function adapted to the inferential problem
- Optimization of this last

Probabilistic relabeling

- Probabilistic approach (Jasra et al., 2005) to take into account the uncertainness of the attribution of the permutation to the parameters
- Model on the deswitched posterior distribution learned from an unswitched sequence
- Probability for each permutation of the parameter get by the Gibbs sampler
- Computation of quantities of interest such as the posterior mean

Bibliographic overview

- Methods allowing to partially solve the problem
- Problem when posterior distributions are poorly separated, tuning parameters to set
- The latent partition is not taken into account

The *latent partition* is now used to solve the label-switching problem

Bibliography

[1] Celeux, G., (1997) Discussion of 'On Bayesian analysis of mixtures models with an unknown number of components' (with discussion), *Journal of Royal Statistical Society: Series B*, 59, 775–776.
[2] Celeux, G., (1998) Bayesian inference for mixtures: the label-switching problem, *R. Payne & P. J. Greens, eds, COMPSTAT 98, Physica, Heidelberg*, 227–232.
[3] Celeux, G., Hurn, M. et Robert, C. P. (2000) Computational and Inferential Difficulties with Mixture Posterior Distributions, *Journal of the American Statistical Association*, 95, 451, 957–970.
[4] Diebolt, J. et Robert, C. P. (1994) Estimation of finite mixture distributions, *Journal of Royal Statistical Society: Series B*, 56, 363–375.
[5] Jasra, A., Holmes, C. C. et Stephens, D. A. (2005) Markov Chain Monte Carlo Methods and the Label Switching Problem in Bayesian Mixture Modeling, *Statistical Science*, 20, 1, 50–67.
[6] Sperrin, M. and Jaki, T. and Wit, E. (2010) Probabilistic relabelling strategies for the label switching problem in Bayesian mixture models, *Statistics and Computing*, 20, 3, 357–366.
[7] Stephens, M. (1997) Bayesian Methods fo Mixtures of Normal Distribution, D. Phil. thesis, Department of Statistics, University of Oxford.

Idea: using the numbering information

$\tilde{\mathcal{Z}} = \{\mathcal{Z}_1, \ldots, \mathcal{Z}_{g!}\}$ a partition of the set of partitions $\mathcal{Z}$. It sets a particular numbering for each partition z of the dataset x

$$\forall h, h' \in \{1, \ldots, g!\}, \exists! \sigma \in \mathcal{P}_g \text{ tq } z \in \mathcal{Z}_h \Leftrightarrow \sigma(z) \in \mathcal{Z}_{h'}$$

with $\sigma(z) = (\sigma(z_1), \ldots, \sigma(z_n))$ indicating that z is permuted in indexes for $\sigma \in \mathcal{P}_g$.
Decomposition of the usual posterior distribution as a mixture of $g!$ posterior distributions conditioned by any particular numbering $\tilde{\mathcal{Z}}$ of the partitions.

$$p(\theta|x) = \sum_{h=1}^{g!} p(\theta|x, \mathcal{Z}_h) p(\mathcal{Z}_h|x) = \frac{1}{g!} \sum_{h=1}^{g!} p(\theta|x, \mathcal{Z}_h).$$

- $\tilde{\mathcal{Z}}$ is unknown but it corresponds to a latent numbering information
- $p(\theta|x, \mathcal{Z}_h)$ not strictly invariant up to a renumbering of z
- The asymmetry depends on the choice of $\tilde{\mathcal{Z}}$
- Choose the cutting $\tilde{\mathcal{Z}}$ which separates the best the distributions $p(\theta|x, \mathcal{Z}_h)$
- Keep as new definition of the posterior distribution any of these $g!$ distributions

Choosing a $g!$ fraction $\tilde{\mathcal{Z}}$ of $\mathcal{Z}$

Choice 1: $\tilde{\mathcal{Z}}^{KL}$

$\tilde{\mathcal{Z}}^{KL}$ maximizes the Kullback-Leibler divergence between the mixture components on $\mathcal{Z}_h$ which is written

$$\tilde{\mathcal{Z}}^{KL} = \arg\max_{\tilde{\mathcal{Z}}} \min_{h=2,\ldots,g!} \int_\Theta p(\theta|x, \mathcal{Z}_1) \ln \left[\frac{p(\theta|x, \mathcal{Z}_1)}{p(\theta|x, \mathcal{Z}_h)} \right] d\theta.$$

Intractable even for very small sample sizes.

Choice 2: $\tilde{\mathcal{Z}}^{MAP}$

$$\tilde{\mathcal{Z}}^{MAP} = \arg\max_{\tilde{\mathcal{Z}}} \min_{h=2,\ldots,g!} \frac{p(\theta^{MAP}|x, \mathcal{Z}_1)}{p(\theta^{MAP}|x, \mathcal{Z}_h)}.$$

It is equivalent to find the most probable numbering unit by unit computed in θ^{MAP} :

$$\mathcal{Z}_1^{MAP} = \left\{ z \in \mathcal{Z} / Id = \arg\max_{\sigma \in \mathcal{P}_g} p(\sigma(z)|x, \theta^{MAP}) \right\},$$

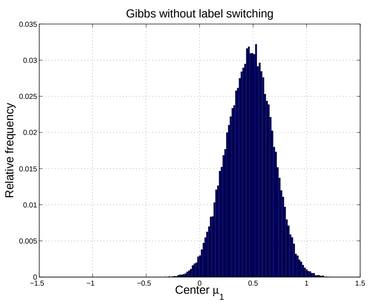
Id is the identity permutation. θ^{MAP} is the reference parameter for the numbering of the latent partition.

Proposed Gibbs algorithm

The classical Gibbs algorithm is slightly modified

- $z \sim p(\cdot|x, \theta)$,
- z permuted in order to $\sigma(z) \in {\mathcal{Z}_1}^{KL}$ or $\sigma(z) \in {\mathcal{Z}_1}^{MAP}$
- $\theta \sim p(\cdot|x, \sigma(z))$

Additional algorithmic complexity negligible for ${\mathcal{Z}_1}^{MAP}$.



New posterior distribution $p(\theta_1|x, \mathcal{Z}_1^{MAP})$.

Numerical experiments

Experiments in the Gaussian setting (running example continued)

- 100 samples x of size $n \in \{3, 10, 100\}$
- Burning sequence of 100 iterations
- $2,000$ iterations of Gibbs

Strategy	$n = 3$	$n = 10$	$n = 100$
Gibbs/k-means	0.18648 (0.10316)	0.09613 (0.09677)	0.02594 (0.04200)
KL	0.03358 (0.04357)	NA	NA
MAP	0.03372 (0.04679)	0.06135 (0.08815)	0.02364 (0.04157)

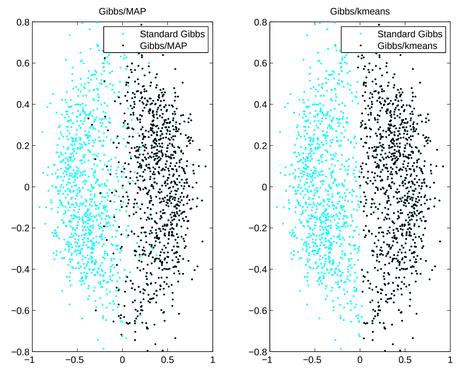
Mean (and standard deviation) of the posterior mean square error.

Experiments in the multinomial product mixture setting

- Qualitative simulated data with $g = 2$ poorly separated classes in equal proportions
- 6 variables: 4 with 3 modalities and 2 with 4 modalities
- Estimated model: mixture of $g = 2$ products of multinomial distributions, all parameters free
- 100 samples of sizes $n = 50$
- Burning sequence of $1,000$ iterations
- $10,000$ iterations of Gibbs

Strategy	$n = 50$
Gibbs/k-means	0.10949 (0.04588)
MAP	0.10627 (0.04549)

Mean (and standard deviation) of the Kullback divergence to the true distribution.



Posterior distribution according to the strategy.

Conclusion and perspectives

Conclusion

- Separation of the posterior modes without break
- Assumption free on the unswitched distribution
- Computational cost similar to standard solutions

Perspectives

- Monitor the convergence of the Gibbs algorithm
- Many application areas: hidden Markov models, Potts model, ...