Variational Bayes methods and algorithms Part II

Christine Keribin

Laboratoire de Mathématique d'Orsay, Université Paris Sud and INRIA - Saclay Ile de France
Université Paris-Saclay

CIRM - March 2016

Outline

1. Some theoretical results
2. Spatial mixture models for segmentation
3. Latent Block Model
4. References

Outline

1. Some theoretical results
2. Spatial mixture models for segmentation
3. Latent Block Model
4. References

Introduction

Previously on Variational Bayes methods...

- In case of non tractable posterior, marginal likelihood...
- Replace an integration by an optimization over a set of functions where the computation is easy.

$$\log p(\mathbf{x}) = \underbrace{\mathcal{F}(q)}_{\text{functional}} + \underbrace{KL(q,p)}_{\text{Kullback divergence}} \geq \underbrace{\mathcal{F}(q)}_{\text{lower bound}}$$

$$p(.|\mathbf{x}) \simeq q^*(.|x) = \arg\min_{q \in \mathcal{Q}} KL(q,p) = \arg\max_{q \in \mathcal{Q}} \mathcal{F}(q)$$

- Define a specific (parametric) form of one component, factorize the posterior $q(\theta, \mathbf{z}) = q(\theta)q(\mathbf{z})$, mean field $q(\mathbf{z}) = \prod_i q(z_i)$;
- Cycling algorithm

$$q_\ell^* = \arg\max_{q_\ell} \mathcal{F}(q_1, \ldots, q_n) = \frac{\exp \mathbb{E}_{i \neq \ell}(\log p(\mathbf{x}, \mathbf{z}))}{\int_{z_\ell} \exp \mathbb{E}_{i \neq \ell}(\log p(\mathbf{x}, \mathbf{z}))\ dz_\ell}$$

- Fast but distorsion on p

Questions

- localization of the mode
- value at the mode
- convergence

Gaussian mixture models [Wang and Titterington 2003, 2004, 2005]

Factorized variational distribution $q_\theta q_{\mathbf{z}}$:

- the variational simplification comes from the fact that the variational posterior is a single member of the corresponding conjugate family, whereas the true posterior is a complicated mixture of large number of such conjugate distribution

Theorem (Convergence of $\widehat{\theta}_{VB}$ in case of Gausian mixtures)

- *The coupled equations of the VBEM iterating algorithm leads to a VB estimator $\widehat{\theta}_{VB} = \mathbb{E}_{q^*}(\theta)$ that converges locally to the true value θ^* with probability 1 and when the starting values are sufficiently closed to θ^**
- *$\widehat{\theta}_{VB}$ converges locally to the maximum likelihood estimator at a rate $O(1/n)$ in the large sample limit*

- VB converges to different limits if different starting values are chosen

Mixture models [Wang and Titterington 2003, 2004, 2005]

Moreover,

- Covariance matrices from the VB approximation are in general "to small" compared with those for the MLE
  $\hookrightarrow$ especially if the components of the mixture model are not well separed
- extension to exponential family models with missing values
- at n fixed, approximated and exact posteriors are different by nature

Bayesian probit model with latent variables [Consonni and Marin 2007]

n latent variables $z_i|\theta \sim \mathcal{N}(v_i\theta, 1)$ are observed through n binary variables x_i

$$x_i = 1 \quad \text{si } x_i > 0, \quad x_i = 0 \quad \text{sinon}$$

A Gaussian prior is defined on θ

- Although the posterior is intractable, it is possible to compute the posterior variance for θ

$$\text{var}(\theta|x) = (I_p + V'V)^{-1} + \text{var}((I_p + V'V)^{-1} V'z).$$

Bayesian probit model with latent variables [Consonni and Marin 2007]

n latent variables $z_i|\theta \sim \mathcal{N}(v_i\theta, 1)$ are observed through n binary variables x_i

$$x_i = 1 \quad \text{si } x_i > 0, \quad x_i = 0 \quad \text{sinon}$$

A Gaussian prior is defined on θ

- Although the posterior is intractable, it is possible to compute the posterior variance for θ

$$\mathrm{var}(\theta|x) = (I_p + V'V)^{-1} + \mathrm{var}((I_p + V'V)^{-1}V'z).$$

- Variational Bayes EM
  - VBEM algorithm is clearly faster than a Gibbs sampler
  - VBEM always underestimates the exact posterior variance (variational variance: $(I_p + V'V)^{-1}$)
  - for small sample sizes, VBEM approximation to the posterior location could be poor, but it becomes better with more observations

Markovian models with missing values

Hall, Humphreys and Titterington 2002

A Gaussian markovian process $y = (x, z)$ is partially observed : x are the accessible observations, z the missing ones. If the z_i are sufficiently far from each others

$$p(z|x) = p(z|x^z) = \prod_{i \in \mathcal{H}} p(z_i|x^i) = \prod_{i \in \mathcal{H}} q_{z_i}(z_i).$$

If not, define a variational posterior with a mean field approximation

$$p(z|x) \simeq q_z(z|x) = \prod_{i \in \mathcal{H}} p(z_i|\widehat{x}^{0i}) = \prod_{i \in \mathcal{H}} q_{z_i}(z_i).$$

- asymptotic: When the m missing sites define a little number of well separated groups such that $m/n \to 0$, then $\widehat{\theta}_{VB} - \widehat{\theta}_{MV} = O(m/n)$ with the same asymptotical variance
- non asymptotic: likelihoods have identical forms but offseted
- fast compared to an exact EM

Result in case of a AR(1) process (m=36, n=64)

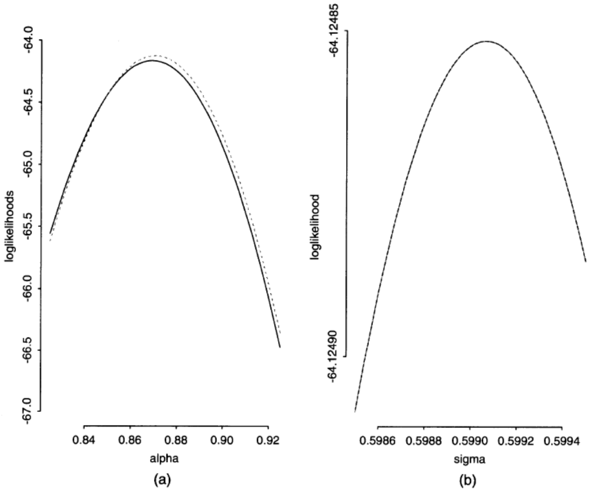


Fig. 2. Log-likelihood surfaces for a single realization ($m = 36$), plotted as one-dimensional sections, (a) with respect to α, with σ fixed at its maximum likelihood estimate, and (b) with respect to σ, with α fixed at its maximum likelihood estimate (see the text for details of the offsets to the approximate log-likelihood surfaces): ········, exact log-likelihood; ———, mean field approximation (offset)

State space model

Wang and Titterington 2004

- Model:

$$\begin{aligned} X_{i+1} &= \theta X_i + \sigma_w W_i,\ X_1 \sim \mathcal{N}(\mu_0, \sigma^2) \\ Y_i &= \alpha X_i + \sigma_v V_i; \quad \sigma_w = \sigma_v = \sigma \end{aligned}$$

- EM with mean field approximation:
  $q_x(x) = \prod_{k=1}^{n} q_i(x_i)$ with $q_i(x_i) \sim \mathcal{N}(\mu_i, \sigma_i)$.
- Kullback dissemblance $D(q_x(x)||p(x|y,\theta))$
  - does not tend to 0 when $n \to \infty$, except if θ tends to 0
  - does not depends on σ, so that it does not tend to 0, no matter how small the noise variance
- VB (and VBEM) are consistant if the noise variance tends to 0, non consistant otherwise

Summary

Easier to catch the localization of the maximum than the value at the maximum

- the mode can be quite well estimated when there is not too much missing data
- But the value of the functional at the mode is often not recovered, even when the mode is correct

$$\log \frac{p(y|m)}{p(y|m')} = \mathcal{F}(m) - \mathcal{F}(m') + D(q'(\theta)||p(\theta|y,m) - D(q(\theta)||p(\theta|y,m')$$

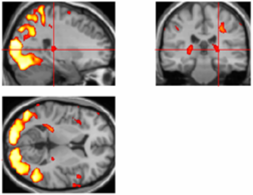
$\hookrightarrow$ need to be cautious when using the difference $\mathcal{F}(m) - \mathcal{F}(m')$ for model selection ...

Outline

1. Some theoretical results
2. Spatial mixture models for segmentation
3. Latent Block Model
4. References

Neuroimaging activation map

In functional brain imaging, the observations can be Statistical Parametric Map: brain images representing a BOLD signal during a cognitive task



- regular lattice of observations **y** where y_i is the observation at spacial location (voxel) i
- the task is to classify areas in the brain: activated, deactivacted and neutral
- encode the prior belief that neighboring voxels are likely to come from the same class

Mixture model on a regular lattice

Woolrich et al 2006

- observations $\mathbf{y}$
- mixture with $K = 3$ components, discrete labels $\mathbf{z}$. Under the assumption of the conditional independance of the likelihood

$$p(\mathbf{z} = \kappa, \theta, \phi_z | \mathbf{y}) \propto \prod_i^n \{p(y_i | z_i = \kappa_i, \theta_{k_i})\} \times p(\mathbf{z} = \kappa | \phi_z) p(\phi_z) p(\theta)$$

- Spatial prior on $\mathbf{z}$: markov random field

$$p(\mathbf{z} = \kappa | \phi_x) \propto f(\phi_z) \exp(-\frac{\phi_z}{4} \sum_i \sum_{j \in \mathcal{N}_i} \mathbb{1}[x_i \neq z_j])$$

The best value of ϕ_z will depend on the topography of the classes (control parameter) with prior

$$p(\phi_z | a, b) = \mathrm{Ga}(a, b)$$

Mixture model with continuous weights

- $f(\phi_z)$ cannot be calculated analytically and computation is very difficult $\hookrightarrow$ continuous weights

$$\prod_i^n \{p(y_i|z_i = \kappa_i, \theta_{\kappa_i})\} \hookrightarrow \prod_i^n \sum_{k=1}^K \{w_{ik} p(y_i|z_i = k, \theta_k)\}$$

with

$$w_{ik} = \frac{\exp(\tilde{w}_{ik}/\gamma)}{\sum_{l=1}^K \exp(\tilde{w}_{il}/\gamma)}$$

$$p(\mathbf{z} = \kappa, \theta, \phi_z|\mathbf{y}) \propto \prod_i^n \sum_{k=1}^K \{w_{ik} p(y_i|z_i = k, \theta_k)\} \times p(\tilde{w}|\phi_z) p(\phi_z) p(\theta)$$

- prior $p(\tilde{w}|\phi_{\tilde{w}}) = \prod_k p(\tilde{w}_k|\phi_{\tilde{w}})$ with

$$p(\tilde{w}_k|\phi_{\tilde{w}}) \sim \mathcal{N}_n(0, (I - C)^{-1}/\phi_{\tilde{w}})$$

- densities for each class: Gaussian (neutral), Gamma (activated $Ga(y_i; a_k, b_k)$ and deactivated $Ga(-y_i; c_k, d_k)$)

Mixture model with continuous weights

VB inference

posterior $p(\tilde{w}, \phi_{\tilde{w}}|y) \propto p(y|\tilde{w})p(\tilde{w}|\phi_{\tilde{w}})p(\phi_{\tilde{w}})$, approximated by $q(\phi_{\tilde{w}}, \tilde{w}|y) = q_{\phi_{\tilde{w}}}(\phi_{\tilde{w}}) \prod_i q(\tilde{w}_i|y)$

- VBE: update $q(\tilde{w}_i|y)$

$$q_w(\tilde{w}_i|y) \propto \exp(\mathbb{E}_{q_{\tilde{w}_{-i}} q_{\phi_{\tilde{w}}}}[\log p(\tilde{w}, \phi_{\tilde{w}}|y)])$$

- VBM: update $q(\phi_{\tilde{w}}|y)$

$$q_{\phi_{\tilde{w}}}(\phi_{\tilde{w}}|y) \propto \exp(\mathbb{E}_{q_{\tilde{w}}}[\log p(\tilde{w}, \phi_{\tilde{w}}|y)])$$

- needs to compute an integral of non linear components (weights)
- the likelihood $\log p(y_i|\tilde{w}_i)$ is approximated by Laplace approximation

Mixture model with continuous weights

Comparison with MCMC inference

Result: build spatial activation map from the a posteriori mean of the weights w_{ik}, for the activated and de-activated classes
With simulated data:

- the need to set adaptatively the spacial control parameter $\phi_{\tilde{w}}$.
- only little difference between VB and MCMC
  - slight advantage for MCMC with regards for the classification error
  - real improvment of the computation time for the VB method (ratio 15 for 10 000 voxels)

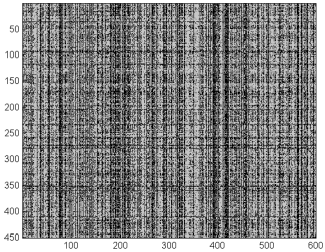
These results are also observed on real data sets.

Outline

1. Some theoretical results
2. Spatial mixture models for segmentation
3. Latent Block Model
4. References

Unsupervised block clustering framework

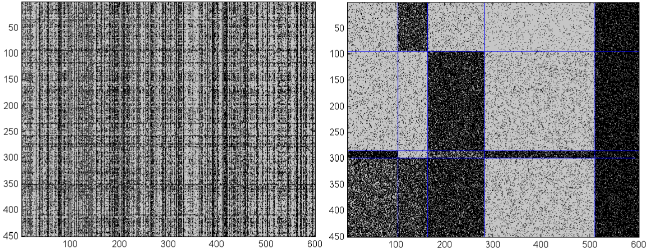
- **Data:** Let $\mathbf{x} = \{(x_{ij}; i = 1, \ldots, n; j = 1, \ldots, d)\}$ be a $n \times d$ matrix



- **Aim:** to find a block clustering structure simultaneously on rows and columns leading to a dramatically parsimonious representation : co-clustering
- **Application:** huge data sets arising in recommendation systems, genomic data analysis, text mining,...

Unsupervised block clustering framework

- Data: Let $\mathbf{x} = \{(x_{ij}; i = 1, \ldots, n; j = 1, \ldots, d)\}$ be a $n \times d$ matrix



- Aim: to find a block clustering structure simultaneously on rows and columns leading to a dramatically parsimonious representation : co-clustering
- Application: huge data sets arising in recommendation systems, genomic data analysis, text mining,...

Latent Block Model: a mixture model

Assume:

- blocks define a 'checker board'

$$p(x;\boldsymbol{\theta}) = \sum_{(z,w)\in\mathcal{Z}\times\mathcal{W}} p(z,w;\theta) \quad p(x|z,w;\theta)$$

- g row clusters: $\mathbf{z} = (z_{ik})$ where $z_{ik} = \mathbb{1}_{i\in C_k}$
- m column clusters: $\mathbf{w} = (w_{j\ell})$ where $w_{j\ell} = \mathbb{1}_{j\in C^\ell}$

Latent Block Model: a mixture model

Assume:

- blocks define a 'checker board' and row and column labels are independently assigned : $z_i \sim \mathcal{M}(1, \boldsymbol{\pi})$, $w_j \sim \mathcal{M}(1, \boldsymbol{\rho})$

$$p(x;\boldsymbol{\theta}) = \sum_{(z,w)\in\mathcal{Z}\times\mathcal{W}} \prod_{i,k} \pi_k^{z_{ik}} \prod_{j,\ell} \rho_\ell^{w_{j\ell}} \quad p(x|z,w;\theta)$$

- g row clusters: $\mathbf{z} = (z_{ik})$ where $z_{ik} = \mathbb{1}_{i\in C_k}$
- m column clusters: $\mathbf{w} = (w_{j\ell})$ where $w_{j\ell} = \mathbb{1}_{j\in C^\ell}$
- $\boldsymbol{\pi} = (\pi_1, \dots, \pi_g)$: the mixing proportions for the rows
- $\boldsymbol{\rho} = (\rho_1, \dots, \rho_m)$: the mixing proportions for the columns

Latent Block Model: a mixture model

Assume:

- blocks define a 'checker board' and row and column labels are independently assigned : $z_i \sim \mathcal{M}(1, \boldsymbol{\pi})$, $w_j \sim \mathcal{M}(1, \boldsymbol{\rho})$
- the $n \times d$ variables x_{ij} are conditionally independent given **z** and **w** and follow the same distribution which parameter only depends on the block: $x_{ij} | z_{ik} w_{j\ell} \sim \varphi(x_{ij}; \alpha_{k\ell})$

$$p(x; \boldsymbol{\theta}) = \sum_{(z,w) \in \mathcal{Z} \times \mathcal{W}} \prod_{i,k} \pi_k^{z_{ik}} \prod_{j,\ell} \rho_\ell^{w_{j\ell}} \prod_{i,j,k,\ell} \varphi(x_{ij}; \alpha_{k\ell})^{z_{ik} w_{j\ell}}$$

- g row clusters: $\mathbf{z} = (z_{ik})$ where $z_{ik} = \mathbb{1}_{i \in C_k}$
- m column clusters: $\mathbf{w} = (w_{j\ell})$ where $w_{j\ell} = \mathbb{1}_{j \in C^\ell}$
- $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_g)$: the mixing proportions for the rows
- $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_m)$: the mixing proportions for the columns

Latent Block Model: a mixture model

Observed Loglikelihood

$$\mathcal{L}(\theta) = \log p(x;\theta) = \log \left(\sum_{(z,w)\in\mathcal{Z}\times\mathcal{W}} \prod_{i,k} \pi_k^{z_{ik}} \prod_{j,\ell} \rho_\ell^{w_{j\ell}} \prod_{i,j,k,\ell} \varphi(x_{ij};\alpha_{k\ell})^{z_{ik}w_{j\ell}} \right)$$

- parameter to estimate: $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\rho}, \boldsymbol{\alpha}) \in \boldsymbol{\Theta}$
- parsimononious representation
- generic identifiability [K et al 2014]

- likelihood for 2×2 blocks and 20×20 matrix: $\approx 2^{20} \times 2^{20} \approx 10^{12}$ terms $\hookrightarrow$: intractable
- Estimation with EM?

ML estimation with the EM algorithm

- **E** step: computation of the expectation of the complete likelihood conditionally to the observations

$$Q(\theta|\theta^{(c)}) = \sum_{i,k} s_{ik}^{(c)} \log \pi_k + \sum_{j,\ell} t_{j\ell}^{(c)} \log \rho_\ell + \sum_{i,j,k,\ell} e_{i,j,k,\ell}^{(c)} \log \varphi(x_{ij}; \alpha_{k\ell})$$

where

$$s_{ik}^{(c)} = P(z_{ik} = 1|\theta^{(c)}, \mathbf{X} = \mathbf{x}), \quad t_{j\ell}^{(c)} = P(w_{j\ell} = 1|\theta^{(c)}, \mathbf{X} = \mathbf{x})$$

and

$$e_{i,j,k,\ell}^{(c)} = P(z_{ik} w_{j\ell} = 1|\theta^{(c)}, \mathbf{X} = \mathbf{x}).$$

- Intractable due to the dependence structure among the rows and columns
- **M** step: $\theta^{(c+1)} = \arg\max_\theta Q(\theta|\theta^{(c)})$, no problem !

Variational EM [Govaert and Nadif 2008]

Principle:

$$\begin{aligned}\mathcal{L}(\theta) &= \mathbb{E}_{q_{zw}}\left[\log \frac{p(\mathbf{x},\mathbf{z},\mathbf{w}|\theta)}{q_{zw}(\mathbf{z},\mathbf{w})}\right] + KL(q_{zw}||p(\mathbf{z},\mathbf{w}|\mathbf{x};\theta)) \\ &= \mathcal{F}(q_{zw},\theta) + KL(q_{zw}||p(\mathbf{z},\mathbf{w}|\mathbf{x};\theta))\end{aligned}$$

$p(\mathbf{z},\mathbf{w}|\boldsymbol{\theta}^{(c)},\mathbf{x})$ is approximated by a distribution which considers $\mathbf{z}$ and $\mathbf{w}$ conditionally independent

$$p(\mathbf{z},\mathbf{w}|\boldsymbol{\theta}^{(c)},\mathbf{x}) \simeq q_z(\mathbf{z}|\boldsymbol{\theta}^{(c)},\mathbf{x})q_w(\mathbf{w}|\boldsymbol{\theta}^{(c)},\mathbf{x})$$

$$\widehat{\boldsymbol{\theta}}_{VAR} = \arg\max_{\boldsymbol{\theta},q_z,q_w} \mathcal{F}(q_z,q_w,\boldsymbol{\theta})$$

Alternate optimization

Thanks to the factorization $q_{zw} = q_z q_w$ the computation of $s_{ik}^{(c)}$ and $t_{j\ell}^{(c)}$ is straightforward

$$s_{ik}^{(c)} = q_z(Z_{ik} = 1; \theta^{(c)}), t_{j\ell}^{(c)} = q_w(W_{j\ell} = 1; \theta^{(c)})$$

$$e_{i,j,k,\ell}^{(c)} = s_{ik}^{(c)} t_{j\ell}^{(c)}$$

Govaert et Nadif 2008

1. **VE step** : Maximize $\mathcal{F}(q_z, q_w, \theta)$ wrt q_z and q_w until convergence
   - 1.1 compute s_{ik} with fixed t_{jl} and $\theta^{(c)}$
   - 1.2 compute t_{jl} with fixed s_{ik} and $\theta^{(c)}$
     $\hookrightarrow s^{(c+1)}$ et $t^{(c+1)}$
2. **M step** : Maximize $\mathcal{F}(q_z^{c+1}, q_w^{c+1}, \theta)$ wrt θ: $\hookrightarrow \theta^{(c+1)}$

VEM properties

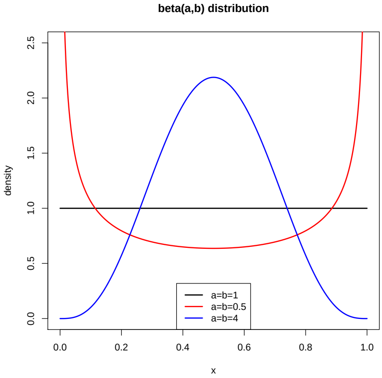
Properties:

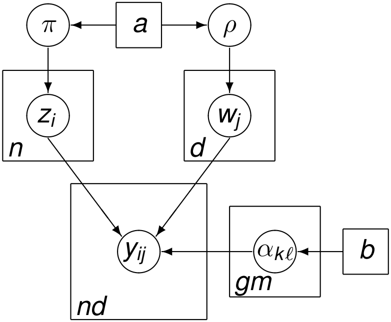
- the parameter estimates could be expected to be a good approximation of the maximum likelihood estimator
- provides a lower bound of the observed loglikelihood
- sensitive to the starting values
- replace the E step by a SE step (Gibbs sampling needed to simulate $(\mathbf{z}, \mathbf{w})$) $\hookrightarrow$ SEM-Gibbs:
  - do not increase the likelihood at each step
  - but generates a irreductible MC with a unique stationary distribution expected to be concentrated around the ML parameter estimate
  - far less sensitive to initial values
  - marked tendency to provide solutions with empty clusters:
- use Bayesian priors on ρ and π to regularize

Bayesian LBM on categorical data (K. et al 2014)

Define priors on the parameters

$$\boldsymbol{\pi} \sim \mathcal{D}(a, \ldots, a), \quad \boldsymbol{\rho} \sim \mathcal{D}(a, \ldots, a), \; \alpha_{k\ell} \sim \mathcal{D}(b, \ldots, b),$$





Bayesian LBM

Model parameter can be estimated by maximising the posterior density $p(\boldsymbol{\theta}|\mathbf{y})$, $\hookrightarrow$ Maximum A Posteriori estimate

$$\widehat{\boldsymbol{\theta}}_{MAP} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathbf{y}).$$

Use the Bayes formula

$$\log p(\boldsymbol{\theta}|\mathbf{y}) = \log p(\mathbf{y}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) - \log p(\mathbf{y})$$

to define an EM algorithm for the computation of the MAP estimate:

VBayes

- **E-V step** : same as VE step
- **M-Bayes step** : maximization of a slighly different objective function [McLachlan and Krishnan 2008]

$$\theta^{(c+1)} = \arg\max_{\theta} \left(Q(\theta, \theta^{(c)}) + \log p(\theta) \right).$$

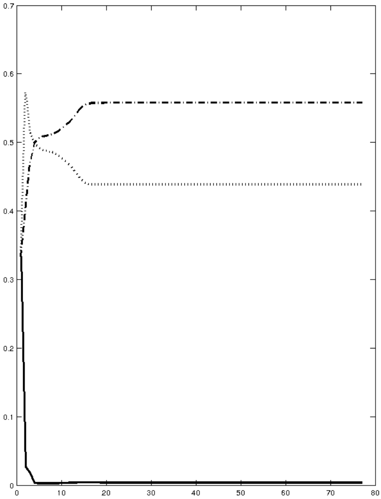
VBayes M-Step

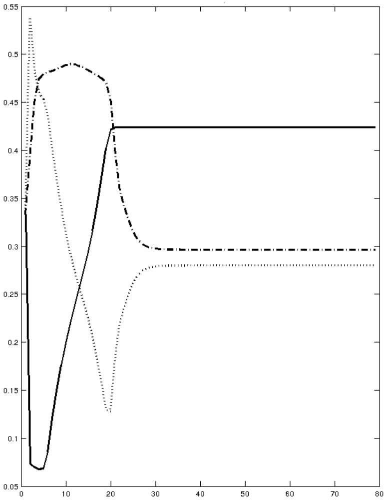
- M Bayes step : update $\theta^{(c+1)}$ with

$$\pi_k^{(c+1)} = \frac{a-1+\sum_i s_{ik}^{(c+1)}}{g(a-1)+n}, \quad \rho_\ell^{(c+1)} = \frac{a-1+\sum_j t_{j\ell}^{(c+1)}}{m(a-1)+d}$$

$$\alpha_{k\ell}^{h\;(c+1)} = \frac{b-1+\sum_{ij} s_{ik}^{(c+1)} t_{j\ell}^{(c+1)} v_{ijh}}{r(b-1)+\sum_{ij} s_{ik}^{(c+1)} t}.$$

VBayes does the job!





But depending on the initial values

Gibbs sampler

Full Bayesian settings: full conditional posterior distributions of the LBM parameters are closed form with Dirichlet prior distributions

Repeat

1. Draw $\mathbf{z}^{(c+1)}$ according to $p(\mathbf{z}|\mathbf{x}, \mathbf{w}^{(c)}, \theta^{(c)})$
2. Draw $\mathbf{w}^{(c+1)}$ according to $p(\mathbf{w}|\mathbf{x}, \mathbf{z}^{(c+1)}, \theta^{(c)})$
3. Draw $\pi^{(c+1)}$ according to $p(\pi|\mathbf{x}, \mathbf{z}^{(c+1)}, \mathbf{w}^{(c+1)}, \rho^{(c)}, \alpha^{(c)})$
4. Draw $\rho^{(c+1)}$ according to $p(\rho|\mathbf{x}, \mathbf{z}^{(c+1)}, \mathbf{w}^{(c+1)}, \pi^{(c+1)}, \alpha^{(c)})$
5. For $k = 1, \ldots, g; \ell = 1, \ldots, m$, draw $\alpha_{k\ell}^{(c+1)}$ according to $p(\alpha|\mathbf{x}, \mathbf{z}^{(c+1)}, \mathbf{w}^{(c+1)}, \pi^{(c+1)}, \rho^{(c+1)})$

$\hookrightarrow$ the stationary distribution of the Markov chain is $p(\mathbf{z}, \mathbf{w}, \pi, \rho, \alpha|\mathbf{x})$

$\hookrightarrow$ $\widehat{\theta}_{gibbs}$ is the mean of $\theta^{(c)}$ after a burn-in period.

$\hookrightarrow$ labels are defined by assignment to the majority class

Algorithmes [K et al 2014]

- *EM*
  $\hookrightarrow$ intractable
- Algorithme *VEM* [Govaert and Nadif 2008]
  $\hookrightarrow$ difficult initialisation
- *SEM-Gibbs*
  $\hookrightarrow$ absorbing states
- *V-Bayes*
  $\hookrightarrow$ rapidly leads to reasonable parameter estimates with a good initialisation
- Gibbs sampling
  $\hookrightarrow$ essentially unsensitive to starting values but fluctuating and tricky stopping criteria

Recommandation: Gibbs sampling as initialization, followed by VBayes

Model selection

Aim: choosing a relevant number of clusters

- A couple (g,m) to select instead of a single number
- Standard penalized likelihood criteria such as BIC need the computation of the loglikelihood which is not tractable

$$\begin{aligned} \mathrm{BIC}(g,m) &= \int p(\mathbf{x}|\theta;g,m)p(\theta;g,m)d\theta \\ &\simeq \max_{\theta} \log(p(\mathbf{x};\theta)) - \frac{D}{2}\log(n) \end{aligned}$$

The good news is that the integrated completed likelihood (ICL) can be derived straightforwardly

Integrated Completed Likelihood criterion

- Bayesian setting: ICL is the logarithm of the integrated completed likelihood

$$p(\mathbf{x}, \mathbf{z}, \mathbf{w} | g, m) = \int p(\mathbf{x}, \mathbf{z}, \mathbf{w} | \theta; g, m) p(\theta; g, m) d\theta$$

where the missing data are replaced by their most probable inferred values $\widehat{\mathbf{z}}$, $\widehat{\mathbf{w}}$ [Biernacki et al (2000)]

$\hookrightarrow$ ICL is focussing on the clustering view of the model

- Proper non informative conjugate priors are available for multinomial LBM :
  - $\hookrightarrow$ Dirichlet distribution $\mathcal{D}(a, \ldots, a)$ for π and ρ
  - $\hookrightarrow$ Dirichlet distribution $\mathcal{D}(b, \ldots, b)$ for α_{kl}

ICL is closed form

Using the conjugacy properties of the prior distributions we get

$$
\begin{aligned}
\log p(x,z,w) \quad = \quad & \log\Gamma(ga) + \log\Gamma(ma) - (m+g)\log\Gamma(a) + mg(\log\Gamma(rb) - r\log\Gamma(b)) \\
& - \log\Gamma(n+ga) - \log\Gamma(da) + \sum_{k=1}^{g}\log\Gamma(z_{.k}+a) + \sum_{l=1}^{m}\log\Gamma(w_{.\ell}+a) \\
& + \sum_{k,l}\left[\left(\sum_{h=1}^{r}\log\Gamma\left(N^h_{k\ell;z,w}+b\right)\right) - \log\Gamma(z_{.k}w_{.\ell}+rb)\right]
\end{aligned}
$$

where

- $z_{.k}$ is the number of rows in cluster k
- $w_{.\ell}$ is the number of columns in cluster ℓ
- $N^h_{k\ell;z,w}$ is the number of h in the block (k,ℓ)

computed from the missing labels replaced by

$$
(\hat{\mathbf{z}},\hat{\mathbf{w}}) = \arg\max_{(\mathbf{z},\mathbf{w})} p(\mathbf{z},\mathbf{w}|\mathbf{x};\hat{\theta}),
$$

Collapsed sampler [Wyse and Friel; 2012]

- Marginalisation over the model parameters with uniform prior to compute the distribution of the most visited models and the maximum a posteriori cluster membership
- Analogous to maximising ICL

$$\begin{aligned}\log p(\mathbf{z}, \mathbf{w}, g, m|\mathbf{x}) &= \mathrm{ICL}(\mathbf{z}, \mathbf{w}, g, m) \\ &\quad + \log p(g, m) - \log p(\mathbf{x}).\end{aligned}$$

- ICL appears to be efficient to find $(\mathbf{z}, \mathbf{w}, g, m)$ and less computationally demanding, but is unable to recover the uncertainty

  $\hookrightarrow$ further work: analyze the variability of ICL

Thank you for your attention!

Outline

1. Some theoretical results
2. Spatial mixture models for segmentation
3. Latent Block Model
4. References

References

- Wang B., Titterington M. (2006)
  Convergence properties of a general algorithm for calculating variational Bayesian estimates for a normal mixture model. *Bayesian Analysis, 1(3) 625–650*
- Consonni G., Marin J.-M. (2007)
  Mean-field variational approximate Bayesian inference for latent variable models. *Computational Statistics & Data Analysis, (52) 790–798*
- Wang B., Titterington M. (2004)
  Lack of consistency of mean field and variational Bayes approximations for state space models. *Neural Proceesing Letters, 20(3) 151–170*
- Hall P., Humphreys K., Titterington M. (2002)
  On the adequacy of variational lower bound functions for likelihood-based inference in Markovian models with missing values. *J. R. Statist. Soc B, 64(3) 549–564*

References

- Keribin C. (2010)
  Méthodes bayésiennes variationnelles : concepts et applications en neuroimagerie. *Journal de la Société Française de Statistiques, vol 151, N° 2*
- Woolrich M.W., Behrens T. E. (2006) Variational Bayes inference for spatial mixture models for segmentation. *IEEE Trans. Med. Imag.*
- Govaert G., Nadif M. (2008)
  Block clustering with Bernoulli mixture models: Comparison of different approaches. *Computational Statistics & Data Analysis, 52, 3233–3245*
- Keribin C., Brault V., Celeux G., Govaert G. (2014)
  Estimation and selection for the latent block model on categorical data. *STCO, DOI 10.1007/s11222-014-9472-2*
- Wyse J., Friel N. (2012) Block clustering with collapsed latent block models. *Statistics and Computing, 22:415–428*