Variational bayesian approach for model aggregation in non-supervised classification

Stevenn Volant

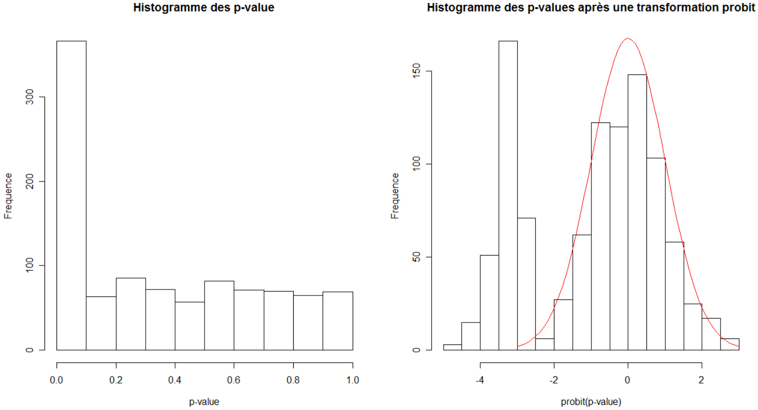
AgroParisTech, UMR 518 MIA

June 1, 2010

Table of contents

Multiple testing: local FDR

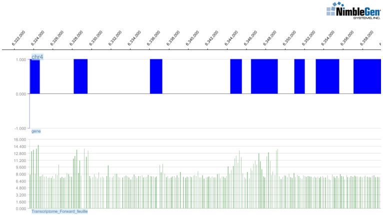
Question: Is gene G differentially expressed under specific conditions?
$\Rightarrow$ Multiple testing, local FDR(False Discovery Rate).



$\hookrightarrow$ Mixture models

Tiling array

1. Provides a mesure of the expression for each probe
2. Dimension: 10^4 to 10^6

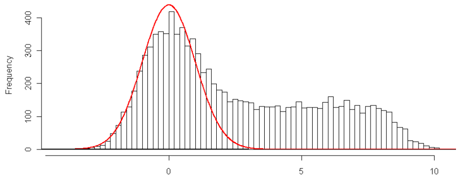


2 kind of probes:

- Expressed: High or middle intensity
- Non expressed: Low intensity near from 0 $\Rightarrow$ easily recognisable

Spatial dependance: A adjacent probe of an expressed probe is more likely to be expressed(and vice versa). $\Rightarrow$ HMM

Binary classification problem



We consider a mixture between two populations :

$$g(x) = Sf(x) + (1 - S)\phi(x) \quad (1)$$

where S equals 1 if x belongs to f and 0 otherwise. The density function ϕ is known and f must be ajusted.
The variable S is either distributed as a **Multinomial** or a first order **Markov Chain**.

Model averaging (1/2)

We are interested in the posterior distribution of S:

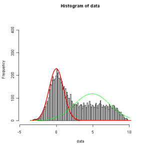
$$P(S|X) = \int P(S, \Theta|X) d\Theta,$$

where Θ is the vector of parameters.
Many models can be considered for the estimation of $P(S|X)$. However,

- Each model brings some information
- Select one specific model is not judicious

$\Rightarrow$ Use an aggregated estimator which combines all the models.

Model averaging (2/2)

	Estimators	Weights	Aggregated estimator
	$\hat{P}^{(1)}(S\|X)$	$\tilde{\alpha}_1$	$\tilde{P}(S\|X) = \sum_m \tilde{\alpha}_m \hat{P}^{(m)}(S\|X)$
	$\hat{P}^{(2)}(S\|X)$	$\tilde{\alpha}_2$	
	$\hat{P}^{(3)}(S\|X)$	$\tilde{\alpha}_3$	
...	$\hat{P}^{(m)}(S\|X)$	$\tilde{\alpha}_m$	

Objective: Estimate α_m.

Model averaging

Jaakkola and Jordan (1998) prouved that combining models provided better results than selecting only one model in $\mathcal{J}$.

$$\min KL(Q_{aggre}(S)||P(S|X)) \leq \min KL(Q_m(S)||P(S|X))$$

Problem:

The quantity $\min KL(Q_{aggre}(S)||P(S|X))$ is hard to calculate.

Shift the original problem:

Instead of minimising $KL(Q(S)||P(S|X))$, we focus on the minimisation of

$$KL(Q(S, M)||P(S, M|X))$$

Table of contents

Minimisation of the Kullback-Leibler divergence (1/2)

In the bayesian framework, the natural weights are based on:

$$P(M|X) = \int P(S, \Theta, M|X) dSd\Theta$$

with M the model.

Theorem

Let M be a random variable, distributed as a multinomiale with parameter r, it yields:

$$M \sim \mathcal{M}(1, r) \quad \textit{with} \quad P(M = m) = r_m.$$

We denote by $\tilde{\alpha}_m$ the posterior distribution of the variable M obtained by the minimisation of $KL(Q(S, \Theta, M|X)||P(S, \Theta, M|X))$. Hence,

$$\tilde{\alpha}_m = \int Q(S, \Theta, M|X) dSd\Theta \propto r_m e^{-KL(Q(S,\Theta|m)||P(S,\Theta,X|m))}.$$

Minimisation of the Kullback-Leibler divergence (2/2)

Proof.

$$\begin{aligned}
KL(Q(H,M)||P(H,M|X)) &= \iint Q(H,M)\log\frac{Q(H,M)}{P(H,M|X)}dHdM \\
&= \iint Q(H|M)Q(M)\log\frac{Q(H|M)Q(M)P(X)}{P(H,M,X)}dHdM \\
&= \int KL(Q(H|M)||P(H,X|M))Q(M)dM - \mathcal{E}(Q(M)) \\
&\quad + \log P(X) - \int \log P(M)Q(M)dM \\
&= \log P(X) + \sum_m Q(m)\,[KL(Q(H|M)||P(H,X|M)) \\
&\quad + \log Q(m) - \log P(M)]
\end{aligned}$$

where, $\mathcal{E}(X) = -\int X \log X dX$.
The minimum is obtained with Lagrange multipliers, i.e. we minimize the functional

$$KL\left(Q(H,M)\|P(H,M|X)\right) - \lambda(\sum_m Q(m) - 1)$$

Interpretation of the theorem

Other writing of the theorem

$$
\begin{aligned}
\tilde{\alpha}_m &\propto r_m e^{-KL(Q(S,\Theta|X,m)||P(S,\Theta,X|m))} \\
&\propto r_m e^{-KL(Q(S,\Theta|X,m)||P(S,\Theta|X,m))+\log P(X|m)}
\end{aligned}
$$

True weights

If $KL(Q(S,\Theta|X,m)||P(S,\Theta|X,m)) = 0$, then $\tilde{\alpha}_m = P(m|X)$

Consequence

We want to minimise $KL(Q(S,\Theta|X,m)||P(S,\Theta|X,m))$

- VBEM algorithm for the bayesian case
- EM algorithm for the frequentist case

Table of contents

Sampling

The true distribution is given by :

$$\begin{aligned} P(M|X) &= \frac{P(M)}{P(X)} P(X|M) \\ &\propto \int P(X,\Theta|M)P(\Theta)d\Theta. \end{aligned}$$

The integral is then estimate by:

$$\hat{\alpha}_m \propto \frac{1}{B} \sum_{b=1}^{B} P(X,\Theta^{(b)}|M=m), \tag{2}$$

avec $\Theta^{(b)} \sim_{iid} P(\Theta)$.

Problem: The variance is very high

Solution: Modified the distribution $P(H)$ in order to speed up the convergence and reduce the variance $\Rightarrow$ Importance Sampling

Importance sampling

We have:

$$P(M|X) \quad \propto \quad \int \frac{P(X,\Theta|M)P(\Theta)}{\mathcal{G}(\Theta)}\mathcal{G}(\Theta)d\Theta. \qquad (3)$$

The function $\mathcal{G}(\Theta)$ represents the importance function.

$$\hat{\alpha}_m \propto \frac{1}{B}\sum_{b=1}^{B}\frac{P(X,\Theta^{(b)}|M=m)P(\Theta^{(b)})}{\mathcal{G}(\Theta^{(b)})}, \qquad (4)$$

with $\Theta^{(b)} \sim_{iid} \mathcal{G}(\Theta)$

Remarks

- Provides an estimation of the posterior distribution $P(M|X)$.
- Reduces the variance of $\hat{\alpha}_m$ for a good choice of $\mathcal{G}$.
- The higher B, the more accurate the estimation is.

Choice of the importance function

Two natural choices of function for $H = \Theta$.

- The posterior distribution of the VBEM algorithm $Q_V(\Theta)$
- The asymptotic normal distribution of the parameters with mean $\hat{\Theta}$ and the variance-covariance calculated from the Fisher information matrix $\mathcal{N}(\hat{\Theta}, \mathcal{I}^{-1})$

$$\mathcal{I}(\Theta, x) = \mathbb{E}\left[-\frac{\partial^2}{\partial\Theta\partial\Theta^T}\mathcal{L}(X,\Theta)\right]. \tag{5}$$

- Is there an optimal choice of $\mathcal{G}$?

Chibs' weights

Chib's method is a direct application of the Bayes theorem, we have:

$$\forall\theta, P(X|M) = \frac{P(X|M,\theta)P(\theta|M)}{P(\theta|X,M)}, \tag{6}$$

- We choose θ as the posterior mean of Θ, $\theta^* = \mathbb{E}(\Theta|X)$.
- $P(\theta^*|X,M)$ is approximated by the distribution $Q(\theta^*|X,M)$

Table of contents

f_m as a mixture density

We consider f_m as a mixture of K_m Gaussian distributions.

$$f_m(x) = \sum_{k=1}^{K_m} p_k \phi_k(x) \quad \text{with} \quad \sum_k p_k = 1$$

Hence,

$$g_m(x) = \sum_{k=0}^{K_m} \pi_k \phi_k(x) \quad \text{with} \quad \sum_k \pi_k = 1$$

Then, we denote by Z_i the label of observation i:

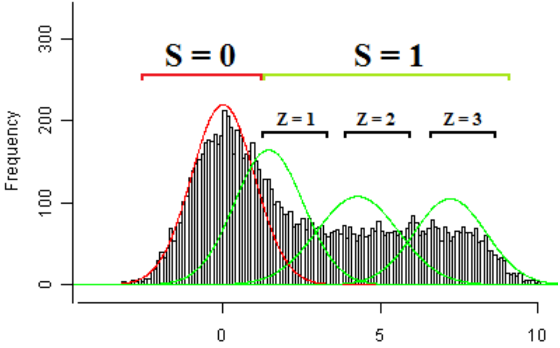
$$Z_i = k \quad \text{if} \quad i \in k$$

These variable are distributed as:

- multinomial, it is a classical mixture model with independant latent variables.
- Markov chain, it is a HMM(Hidden Markov Model) with spatially dependant latent variables(with a specific transition matrix).

Exemple

- The class of interest is modelised by a standard gaussian distribution.
- The alternative is fitted by a 3-components gaussian mixture.



Divide the problem: proposition

Proposition

Minimise $KL(Q(S,\Theta|X,M)||P(S,\Theta|X,M))$
is equivalent to
Minimise $KL(Q(Z,\Theta|X,M)||P(Z,\Theta|M,X))$.

Interpretation

- We can devide the problem into easier sub-problems.
- It is more convenient to use Z rather than S.

Objective: Minimise $KL(Q(Z,\Theta|X,M)||P(Z,\Theta|M,X))$

Variational approximation

We want:

$$\operatorname*{Argmin}_{\Theta} KL(Q(\Theta, Z)||P(\Theta, Z|X))$$

The minimum is obtained for $Q(\Theta, Z) = P(\Theta, Z|X)$.

Problem : We must know the marginal likelihood to calculate $P(\Theta, Z|X)$.
$\Rightarrow$ We consider a distribution Q_V define by:

$$Q_V(\Theta, Z) = Q_\Theta(\Theta) \times Q_Z(Z).$$

Prior distributions: Normal inverse-gamma model

Hypothesis on latent variables

We suppose that latent variables are independant and:

$$Z_i \sim \mathcal{M}(1; \pi),$$

and,

$$Q_Z(Z) = \prod_i P(Z_i)$$

Prior distributions

Data are distributed as a mixture of $\mathcal{N}(\mu_k, \frac{1}{\lambda_k})$.
We consider a particular class models , which are called conjugate-exponential (CE) models (Beal et Gharahmani (2003)) :

- $\pi \sim \mathcal{D}(p_0, \ldots, p_{K-1}) \Rightarrow$ Proportions
- $\mu_k \sim \mathcal{N}(m, \frac{1}{t \times \lambda_k}) \Rightarrow$ Means
- $\lambda_k \sim \Gamma(a, b) \Rightarrow$ Precision

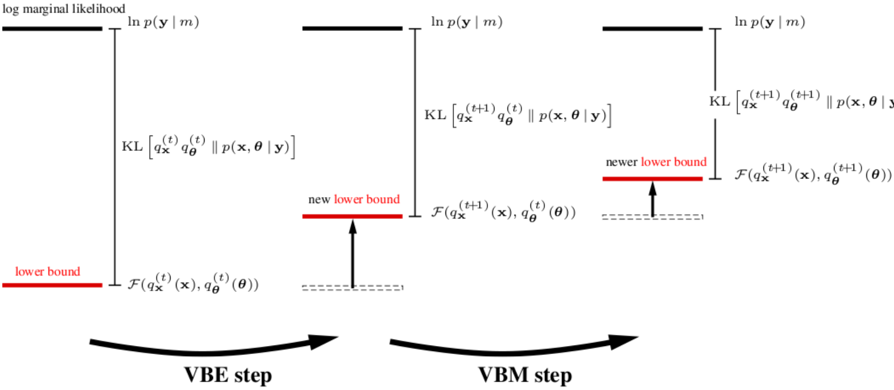


Figure: The VBEM steps

source: Beal thesis

Prior distributions

Hypothesis on latent variables

We suppose that:

$$Q_Z(Z) = \prod_i P(Z_i | Z_{i-1})$$

Prior distributions

We denote by $\Pi = \{\pi_{kj}\}_{k=0\ldots K-1, j=0\ldots K-1}$ the transition matrix:

$$\pi_{kj} = P(Z_{t+1} = j | Z_t = k).$$

- $\pi_{k.} \sim \mathcal{D}(p_1^{(k)}, \ldots, p_K^{(k)}) \Rightarrow$ Transition matrix
- $\mu_k | \lambda_k \sim \mathcal{N}\left(m, \frac{1}{t \times \lambda_k}\right) \Rightarrow$ Mean
- $\lambda_k \sim \Gamma(a, b) \Rightarrow$ Precision

VB-HMM algorithm

"Step E:" $Q_Z(Z)$

$Q_Z(Z)$ is obtained via a forward-bakward algorithm.

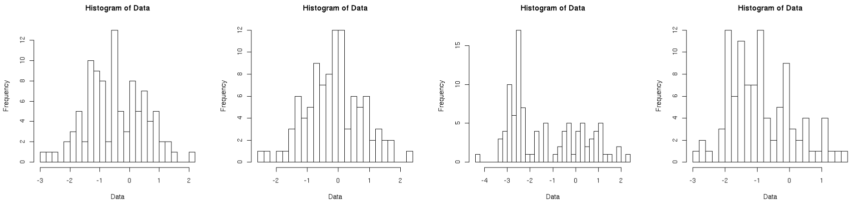
"Step M:"$Q_\Theta(\Theta)$

It is the same M step for the VBEM and the VB-HMM algorithms.

Table of contents

Design

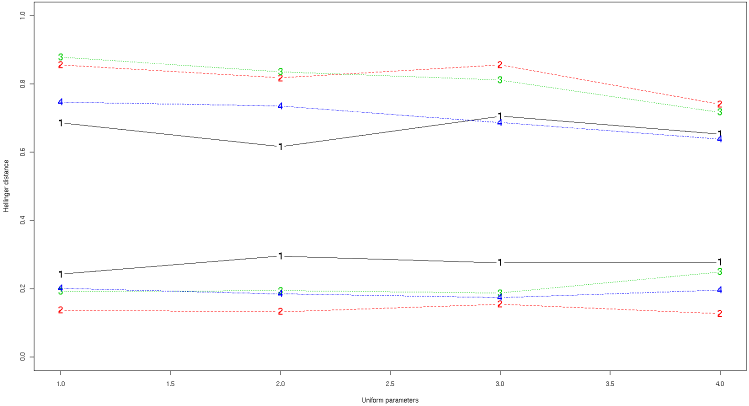
- Simulate a mixture of a $\mathcal{U}[0,1]$ and a β-distribution or a $\mathcal{U}$-distribution.
- Apply a probit transformation.
- Choose 4 transition matrix and 4 different parameters for the alternative distribution.



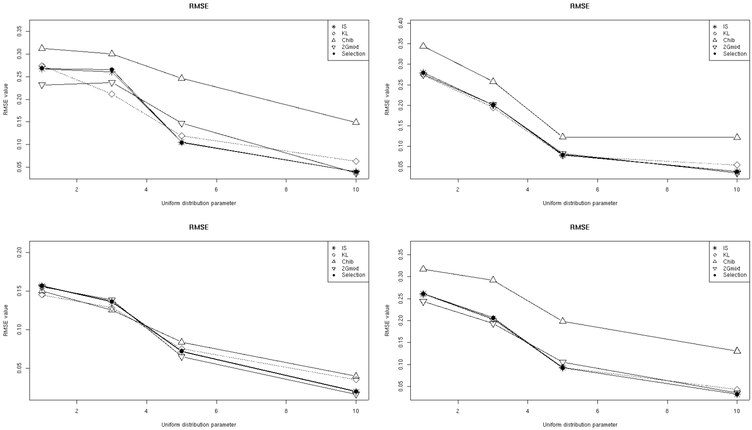
- Simulate $S = 100$ datasets of size 100
- We compare the optimal variational weights and Chib's weights to the IS approach.
- We calculate the RMSE between the theoretical distribution of S and its estimation.

Hellinger distance

The next figure displays the Hellinger distance between the estimated weights (Uniform simulation case).



RMSE


RMSE
RMSE value
Uniform distribution parameter
IS
KL
Chib
2Gmixt
Selection

Conclusion

Conclusion

- The optimal weights are closer to the true weights than Chib's ones.
- The RMSE highlights promising results for model averaging.

Perspectives

- Application of the method on transcriptional dataset.
- Extension to HMRF (Hidden Markov Random Fields)