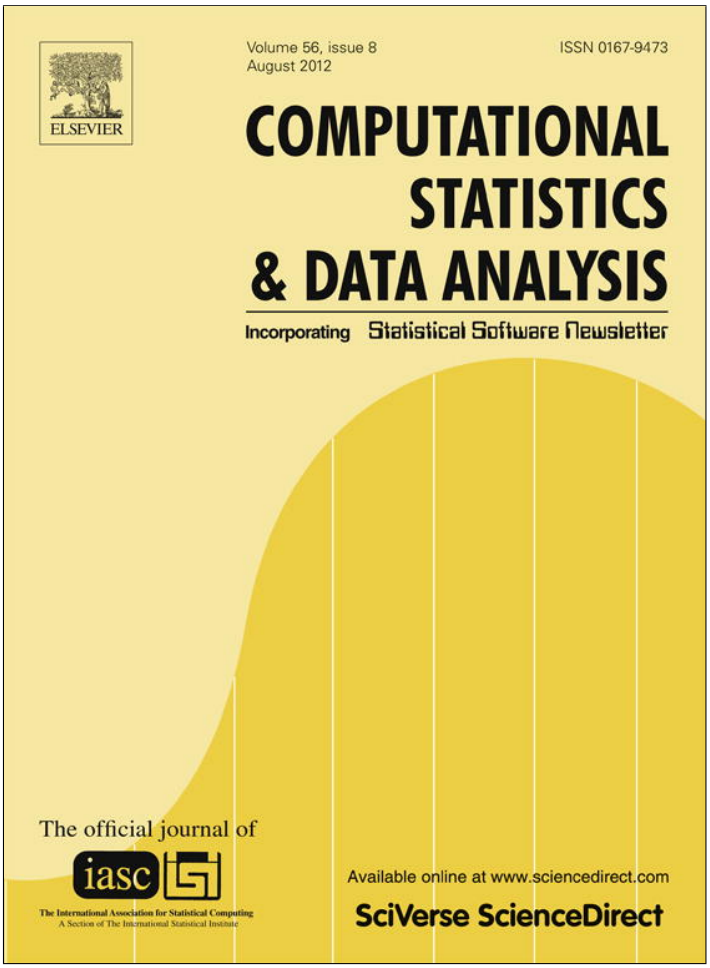
Variational Bayes approach for model aggregation in unsupervised classification with Markovian dependency

Stevenn Volant, Marie-Laure Magniette Martin-Magniette, Stephane S. Robin



Computational Statistics and Data Analysis 56 (2012) 2375–2387

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Variational Bayes approach for model aggregation in unsupervised classification with Markovian dependency

Stevenn Volant [a,b,*], Marie-Laure Martin Magniette [a,b,c,d,e], Stéphane Robin [a,b]

[a] AgroParisTech, 16 rue Claude Bernard, 75231 Paris Cedex 05, France
[b] INRA UMR MIA 518, 16 rue Claude Bernard, 75231 Paris Cedex 05, France
[c] INRA UMR 1165, URGV, 2 rue Gaston Crémieux, CP5708, 91057, Evry Cedex, France
[d] UEVE, URGV, 2 rue Gaston Crémieux, CP5708, 91057, Evry Cedex, France
[e] CNRS ERL 8196, URGV, 2 rue Gaston Crémieux, CP5708, 91057, Evry Cedex, France

ARTICLE INFO

Article history:
Received 11 April 2011
Received in revised form 23 January 2012
Accepted 23 January 2012
Available online 15 February 2012

Keywords:
Model averaging
Variational Bayes inference
Markov chain
Unsupervised classification

ABSTRACT

A binary unsupervised classification problem where each observation is associated with an unobserved label that needs to be retrieved is considered. More precisely, it is assumed that there are two groups of observation: normal and abnormal. The 'normal' observations are coming from a known distribution whereas the distribution of the 'abnormal' observations is unknown. Several models have been developed to fit this unknown distribution. An alternative based on a mixture of Gaussian distributions is proposed. The inference is performed within a variational Bayesian framework and the aim is to infer the posterior probability of belonging to the class of interest. To this end, it makes little sense to estimate the number of mixture components since each mixture model provides more or less relevant information to the posterior probability estimation. By computing a weighted average (named aggregated estimator) over the model collection, Bayesian Model Averaging (BMA) is one way of combining models in order to account for information provided by each model. An aim is then the estimation of the weights and the posterior probability for a specific model. Optimal approximations of these quantities from the variational theory are derived; other approximations of the weights are also proposed. It is assumed that the data are dependent (Markovian dependency) and hence a Hidden Markov Model is considered. A simulation study is carried out to evaluate the accuracy of the estimates in terms of classification performance. An illustration on both epidemiologic and genetic datasets is presented.

1. Introduction

Binary unsupervised classification. We consider an unsupervised classification problem where each observation is associated with an unobserved label that we want to retrieve. Such problems occur in a wide variety of domains, such as climate, epidemiology (see Sun and Cai (2009)), or genomics (see McLachlan et al. (2002)) where we want to distinguish 'normal' observations from abnormal ones or, equivalently, to distinguish pure noise from signal. In such situations, some prior information about the distribution of 'normal' observations, or about the distribution of the noise is often available and we want to take advantage of it.

* Corresponding author at: AgroParisTech, 16 rue Claude Bernard, 75231 Paris Cedex 05, France.
E-mail addresses: stevenn.volant@agroparistech.fr (S. Volant), marie_laure.martin@agroparistech.fr (M.-L. Martin Magniette), stephane.robin@agroparistech.fr (S. Robin).

0167-9473/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2012.01.027

More precisely, based on observations $X = \{X_t\}$, we want to retrieve the unknown binary labels $S = \{S_t\}$ associated with each of them. We assume that 'normal' observations (labelled 0) have distribution ϕ, whereas 'abnormal' observations (labelled 1) have distribution f. We further assume that the null distribution ϕ is known, whereas the alternative distribution f is not. From a classification perspective, we want to compute

$$T_t = \Pr\{S_t = 0|X\}. \tag{1}$$

Bayesian model averaging (BMA). The probability T_t depends on the unknown distribution f. Many models can be considered to fit this distribution and we denote $\mathcal{M} = \{f_m; m = 1, \ldots, M\}$ a finite collection of such models. As none of these models is likely to be the true one, it seems more natural to gather information provided by each of them, rather than to try to select the 'best' one. The Bayesian framework is natural for this purpose, as we have to deal with model uncertainty.

Bayesian model averaging (BMA) has been mainly developed by Hoeting et al. (1999) and provides the general framework of our work. It has been demonstrated that BMA can improve predictive performance and parameter estimation in Madigan and Raftery (1993), Madigan and Hutchinson (1995), Volinsky et al. (1997), Raftery and Zheng (2003) or Ruggieri and Lawrence (2011). Jaakkola and Jordan (1998) also demonstrated that model averaging often provides a gain in terms of classification and fitting. The determination of the weight α_m associated with each model m when averaging is a key ingredient of all these approaches.

Weight determination. As shown in Hoeting et al. (1999) the standard Bayesian reasoning leads to $\alpha_m = \Pr\{M = m|X\}$, where M stands for the model. In a classical context, the calculation of α_m requires one to integrate the joint conditional distribution $P(M, \Theta|X)$, where Θ is the vector of model parameters, and several approaches can be used. The BIC criterion (Schwarz, 1978) is based on a Laplace approximation of this integral, which is questionable for small sample sizes. One other commonly used method is MCMC (Monte Carlo Markov Chain, Andrieu (2003)) which samples the distribution and can provide an accurate estimation of the joint conditional distribution, but at the cost of huge (sometimes prohibitive) computational time.

In the unsupervised classification context, the problem is even more difficult as we need to integrate the conditional $P(M, \Theta, S|X)$ since the labels are unobserved. This distribution is generally not tractable but, for a given model, Beal and Ghahramani (2003) developed a variational Bayes strategy to approximate $P(\Theta, S|X)$. Variational techniques aim at minimising the Kullback–Leibler (KL) divergence between $P(\Theta, S|X)$ and an approximated distribution $Q_{\Theta,S}$ (Corduneanu and Bishop, 2001; Wainwright and Jordan, 2008; Ren and Hodges, 2011). Jaakkola and Jordan (1998) proved that the variational approximation can be improved by using a mixture of distributions rather than factorised distribution as the approximating distribution. A mixture distribution Q_{mix} is chosen to minimise the KL-divergence with respect to $P(\Theta, S|X)$. Unfortunately, their method averages the log of Q_{mix} over all the configurations which leads to untractable computation and a costly algorithm involving a smoothing distribution.

Our contribution. In this article, we propose variational-based weights for model averaging, in presence of a Markov dependency between the unobserved labels. We prove that these weights are optimal in terms of KL-divergence from the true conditional distribution $P(M|X)$. To this end, we optimise the KL-divergence between $P(\Theta, S, M|X)$ and an approximated distribution $Q_{\Theta,S,M}$ (Section 2). This optimisation problem differs from that of Jaakkola and Jordan (1998) (see their Eq. (14)). Based on the approximated distribution of $P(\theta, S|M, X)$, we derive other estimations of the weights.

We then reconsider the case of unsupervised classification and consider a collection $\mathcal{M}$ of mixtures of parametric exponential family distributions (Section 3). We propose a complete inference procedure that does not require any specific development in terms of an inference algorithm. In order to assess our approach, we propose a simulation study which highlights the gain of model averaging in terms of binary classification (Section 4). We also present two illustrations on epidemiologic and genomic datasets (Section 5). An R package named VBMA4HMM (Variational Bayes Models Averaging for hidden Markov models) is available on the CRAN.

2. Variational weights

The aim of model averaging is to account for the information in each model of a collection of M models. To do so, we need to calculate the weight of each model. In this section, we propose three different weights based on the variational Bayes theory.

2.1. A two-step optimisation problem

In a Bayesian Model Averaging context, we focus on averaged estimators to account for model uncertainty. It implies evaluating the conditional distribution:

$$P(M|X) = \int P(H, M|X)dH, \tag{2}$$

where H stands for all hidden variables, that is $H = (S, \Theta)$, and M denotes the model.

In order to calculate this distribution, we need to compute the joint posterior distribution of H and M. Due to the latent structure of the problem, this is not feasible. However, the mean field/variational theory allows an approximation of this

distribution to be derived. This theory has mainly been developed by Parisi (1988) and provides an alternative approach to MCMC for inference problems within a Bayesian framework. The variational approach is based on the minimisation of the KL-divergence between $P(H, M|X)$ and an approximated distribution $Q_{H,M}$. The optimisation problem can be decomposed as follows:

$$\min_{Q_{H,M}} KL(Q_{H,M}\|P(H, M|X)) = \min_{Q_M}\left[KL(Q_M\|P(M|X)) + \sum_m Q_M(m)\min_{Q_{H|m}} KL(Q_{H|m}\|P(H|X, m))\right]. \tag{3}$$

This decomposition separates Q_M and $Q_{H|M}$, and these optimisations can thus be realised independently. We are mostly interested in Q_M which provides an approximation of $P(M|X)$ given in Eq. (2). Furthermore, since the collection $\mathcal{M}$ is finite, we do not need to put any restriction on the form of Q_M and may deal with the weights $\alpha_m = Q_M(m)$ for each $m \in \mathcal{M}$. In the following, we will first minimise the KL-divergence with regard to Q_M leading to weights that depend on $Q_{H|m}$. In a second step, we will consider the approximation of $P(H|X, m)$.

2.2. Weight function of any approximation of $P(H|X, m)$

We now consider the optimisation of Q_M. Proposition 2.1 provides the optimal weights.

Proposition 2.1. *The weights that minimise $KL(Q_{H,M}\|P(H, M|X))$ with respect to Q_M, for given distributions $\{Q_{H|m}, m \in \mathcal{M}\}$, are*

$$\overline{\alpha}_m(Q_{H|m}) \propto P(m)\exp[-KL(Q_{H|m}\|P(H|X, m)) + \log P(X|m)], \tag{4}$$

with $\sum_m \overline{\alpha}_m(Q_{H|m}) = 1$.

Proof. $KL(Q_{H,M}\|P(H, M|X))$ can be rewritten as:

$$\begin{aligned}
&\sum_m \int Q_{H|m}(h)Q_M(m)\log\left[\frac{Q_{H|m}(h)Q_M(m)}{P(h, m, X)/P(X)}\right]dh\\
&= \sum_m \int Q_{H|m}(h)Q_M(m)\left[\log Q_{H|m}(h) + \log Q_M(m) + \log P(X) - \log P(h, m, X)\right]dh\\
&= \sum_m \left(\int Q_{H|m}(h)Q_M(m)\left[\log\frac{Q_{H|m}(h)}{P(h, X|m)} + \log Q_M(m) - \log P(m)\right]dh\right) + \log P(X)\\
&= \sum_m \left(Q_M(m)\left[KL(Q_{H|m}\|P(H, X|m)) + \log Q_M(m) - \log P(m)\right]\right) + \log P(X).
\end{aligned}$$

The minimisation with respect to Q_M subject to $\sum_m Q_M(m) = 1$ gives the result. □

Note that if $Q_{H|m} = P(H|X, m)$ then KL-divergence in the exponential is 0, so $\overline{\alpha}_m$ equals $P(m|X)$.

2.3. Weights based on the optimal approximation of $P(H|X, m)$

We now derive three different weights based on a variational Bayes approximation. The first one comes from the complete optimisation of the KL divergence (see Corollary 2.1). The second one is based on a plug-in approach. The third uses the variational posterior as a proposal for importance sampling.

Full variational approximation. To solve optimisation problem (3) we still need to minimise the divergence $KL(Q_{H|m}\|P(H|X, m))$ for each model m, where $H = (S, \Theta)$. The minimum is clearly reached for $Q_{H|m} = P(H|X, m)$.

Due to the latent structure, the optimisation cannot be done directly. When $P(X, S|\Theta, M)$ belongs to the exponential family and if $P(\Theta|M)$ is the conjugate prior, the Variational Bayes EM (VBEM; Beal and Ghahramani (2003)) algorithm allows us to minimise this KL-divergence within the class of factorised distributions: $\mathcal{Q}_m = \{Q_{H|m} : Q_{H|m} = Q_{S|m}Q_{\Theta|m}\}$. We approximate $P(H|X, m)$ using the $Q^{VB}_{H|m}$ defined as:

$$Q^{VB}_{H|m} = \arg\min_{Q\in\mathcal{Q}_m} KL(Q_{H|m}\|P(H|X, m)).$$

This approximation of $P(H|X, m)$ allows us to define variational weights $\widehat{\alpha}^{VB}_m$ in Corollary 2.1.

Corollary 2.1. *The weights $\widehat{\alpha}^{VB}_m$ achieving optimisation problem (3) for factorised conditional distribution $Q_{H|m}$ are derived from Eq. (4) and are defined as:*

$$\widehat{\alpha}^{VB}_m \propto P(m)\exp\left[-\min_{Q_{H|m}\in\mathcal{Q}_m} KL(Q_{H|m}\|P(H|X, m)) + \log P(X|m)\right],$$

where $\min_{Q_{H|m}\in\mathcal{Q}_m} KL(Q_{H|m}\|P(H|X, m))$ is achievable by the VBEM algorithm (see Section 3).

The weights α^{VB} are based on the KL-divergence between the distribution $Q^{VB}_{H|m}$ and $P(H|X, m)$. Using classical approaches (Plug-in and Importance Sampling), we now define two other weights based on $Q^{VB}_{H|m}$.

Plug-in weights. The weights $\alpha_m = \Pr\{M = m|X\}$ can be estimated by using a plug-in estimation based on a direct application of Bayes' theorem. The conditional probability $P(m|X)$ is proportional to $P(X|m)$ which is equal to $P(X|m, \Theta)P(\Theta|m)/P(\Theta|X, m)$ whatever the value of Θ, thus avoiding integrating over S. The distribution $Q^{VB}_{\Theta|m}$ resulting from the VBEM algorithm is an approximation of $P(\Theta|X, m)$. Setting Θ at its (approximate) posterior mean $\theta^* = \mathbb{E}_{Q^{VB}_{\Theta}}(\Theta)$, we define the following plug-in estimate

$$\widehat{\alpha}^{PE}_m \propto P(m)\frac{P(X|m, \theta^*)P(\theta^*|m)}{Q^{VB}_{\Theta|m}(\theta^*)}. \tag{5}$$

Importance sampling. The weights given in Corollary 2.1 are based on an approximation of the conditional distribution $P(H|X)$. But, the weights defined in 2 can be estimated via importance sampling (Marin and Robert, 2010). For any distribution R, we have

$$P(m|X) \propto \int P(m)\frac{P(X|h, m)P(h|m)}{R(h)}R(h)dh.$$

Importance sampling provides an unbiased estimator of $P(m|X)$. The importance function R can be chosen to minimise the variance of the estimator. The minimal variance is reached when $R(H)$ equals $P(H|X)$ (Marin and Robert, 2010). Thus, in the variational framework, the approximated posterior distribution $Q^{VB}_{H|m}$ is a natural choice for the importance function R, leading to the following weights:

$$\widehat{\alpha}^{IS}_m \propto P(m)\frac{1}{B}\sum_{b=1}^{B}\frac{P(X|H^{(b)}, m)P(H^{(b)})}{Q^{VB}_{H|m}(H^{(b)})}, \qquad \{H^{(b)}\}_{b=1,\ldots,B} \text{ i.i.d.} \sim Q^{VB}_{H|m}.$$

Although this estimate is unbiased, when the number of observations is large, it may require a long computational time to get a reasonably small variance.

3. Unsupervised classification

3.1. Binary hidden Markov model

We now return to the original binary classification problem with Markovian dependence between the labels. To this aim we consider a classical hidden Markov model (HMM). We assume that $\{S_t\}_{1\leq t\leq n}$ is a first order Markov chain with transition matrix $\Pi = \{\pi_{ij}; i, j = 0, 1\}$. The observed data $\{X_t\}_{1\leq t\leq n}$ are independent conditionally to the labels. We denote ϕ the emission distribution in state 0 ('normal') and f the emission distribution in state 1 ('abnormal'). We recall that the function ϕ is known whereas f is unknown and we consider the collection $\mathcal{M} = \{f_m; m = 1, \ldots, M\}$ where f_m is a mixture of m components:

$$f_m(x) = \sum_{k=1}^{m} p_k\phi_k(x), \quad \text{with } \sum_{k=1}^{m} p_k = 1.$$

This collection is large as it allows us to fit the data from a two-component mixture (see McLachlan et al. (2002)) to a semi-parametric kernel-based density (see Robin et al. (2007)). When f is approximated by a mixture of m components, the initial binary HMM with latent variable S can be rephrased as an $(m + 1)$-state HMM with hidden Markov chain $\{Z_t\}$ taking its values in $\{0, \ldots, m\}$ with transition matrix

$$\Omega = \begin{pmatrix} \pi_{00} & \pi_{01}p_1 & \cdots & \pi_{01}p_m \\ \pi_{10} & \pi_{11}p_1 & \cdots & \pi_{11}p_m \\ \vdots & \vdots & \vdots & \vdots \\ \pi_{10} & \pi_{11}p_1 & \cdots & \pi_{11}p_m \end{pmatrix}.$$

The observed data $\{X_t\}_{1\leq t\leq n}$ are independent conditionally on the $\{Z_t\}$ with distribution

$$X_t|Z_t \sim \phi_{Z_t},$$

where $\phi_0 = \phi$. Hence, we have two latent variables Z and S which correspond to the group within the whole mixture and to the binary classification, respectively.

3.2. Variational Bayes inference

The VBEM (Beal and Ghahramani, 2003) aims at minimising the KL-divergence in exponential family/conjugate prior context. The quality of the VBEM estimators has been studied by Wang and Titterington (2004,?, 2003) for mixture models.

Wang and Titterington (2003) have also studied the quality of variational approximation for state space models. The VBEM algorithm has been studied by McGrory and Titterington (2006) for the HMM with emission distributions belonging to the exponential family. The convergence of the variational Bayes estimator to the maximum likelihood estimator has been demonstrated at rate $\mathcal{O}(1/n)$. In Wang and Titterington (2004), it is shown that the covariance matrix of the variational Bayes estimators is underestimated compared to the one obtained for the maximum likelihood estimators.

The VBEM algorithm is an EM-like algorithm with alternate pseudo-steps E and M. At the pseudo-M, the approximate posterior distribution of the parameters is updated, based on the calculation of an expectation with respect to current approximate posterior distribution of the hidden variables. The pseudo-step E is symmetric (Beal and Ghahramani, 2003). Interestingly, in the case of HMM, this latter step can be implemented via the popular forward–backward algorithm (Baum et al., 1970).

In our case, $P(X, S|\Theta, M)$ does not belong to the exponential family whereas $P(X, Z|\Theta, M)$ does. We will therefore make the inference on the $(m+1)$-state hidden Markov model involving Z rather than the binary hidden Markov model involving S. Despite the specific form of the transition matrix Ω, it does not modify the framework of the exponential family/conjugate prior. To be specific, $\log P(X, Z|\Theta, M)$ can be decomposed as $\log P(Z|\Theta, M) + \log P(X|Z, \Theta, M)$ and only the first term involves Ω:

$$\begin{aligned}\log P(Z|\Theta, M) = &\sum_{k=1}^{m}\sum_{j=1}^{m} N_{kj}\log\pi_{11} + N_{00}\log\pi_{00} + \sum_{k=1}^{m} N_{k0}\log\pi_{10} \\ &+ \sum_{j=1}^{m} N_{0j}\log\pi_{01} + \sum_{k=1}^{m} Z_{1k}\log q_1 + Z_{10}\log q_0 + \sum_{k=0}^{m}\sum_{j=1}^{m} N_{kj}\log p_j + \sum_{k=1}^{m} Z_{1k}\log p_k, \end{aligned} \tag{6}$$

with $N_{kj} = \sum_{t\geq 2} Z_{t-1,k}Z_{tj}$ and q is the stationary distribution of Π. Since $\log P(Z|\Theta, M)$ can be written as a scalar product $\Phi \cdot u(Z)$ with Φ the vector of parameters and $u(Z)$ the vector containing the $\{N_{kj}\}_{1\leq k,j\leq m}$ and the sums over Z, it shows that $Z|\Theta, M$ belongs to the exponential family and that this specific form of Ω only affects the updating step of hyper-parameters.

3.3. Model averaging

For each model m from the collection $\mathcal{M}$, the VBEM algorithm provides the optimal distributions $Q^{VB}_{H|m}$, from which we can derive the three weights defined in Section 2: $\widehat{\alpha}^{VB}_m$, $\widehat{\alpha}^{PE}_m$ and $\widehat{\alpha}^{IS}_m$. Based on these weights, we can get an averaged estimate of the distribution f:

$$\widetilde{f}^{\mathcal{A}} = \sum \widehat{\alpha}^{\mathcal{A}}_m \widehat{f}_m,$$

where $\mathcal{A}$ corresponds to one of the proposed approaches (VB, PE or IS). Although the largest model only involves M components, the averaged distribution is a mixture with $M(M+1)/2$ components. As we are mostly interested in the estimation of the posterior probability T_t defined in (1), we similarly define its averaged estimate:

$$\widetilde{T}^{\mathcal{A}}_t = 1 - \sum_m \widehat{\alpha}^{\mathcal{A}}_m \mathbb{E}_{Q^{VB}_{Z|m}}(S_t),$$

where $\mathbb{E}_{Q^{VB}_{Z|m}}(S_t)$ corresponds to the expected value of S calculated with the optimal variational posterior distribution of Z. This expectation does not depend on $\mathcal{A}$.

In this article, we propose a variational-based approach for model averaging, the steps of the process can be summarised as follows:

Consider a dataset X and a collection of models $\mathcal{M}$.

1. For each $m \in \mathcal{M}$, run the VBEM algorithm and collect the posteriors.
2. Compute the weights $\{\widehat{\alpha}^{\mathcal{A}}\}_{m\in\mathcal{M}}$.
3. Estimate the averaged estimator $\widetilde{T}^{\mathcal{A}}$.

4. Simulation study

In this section, we study the efficiency of the estimators defined in the previous sections. First, we study the accuracy of α^{VB} and α^{PE} in terms of weight estimation. Then, we focus on the accuracy from a classification point of view. We therefore compare the averaged estimator of the posterior probability T_t to the theoretical one. We also compare the averaging approach with a classical two-state HMM and with the HMM which has the highest weight calculated with the IS approach, called throughout the remainder of the paper "selected HMM". This means that the "selected HMM" approach accounts for estimation given by $\widehat{m}$:

$$\widehat{m} = \underset{m\in\mathcal{M}}{\operatorname{argmax}}\, \widehat{\alpha}^{IS}_m. \tag{7}$$

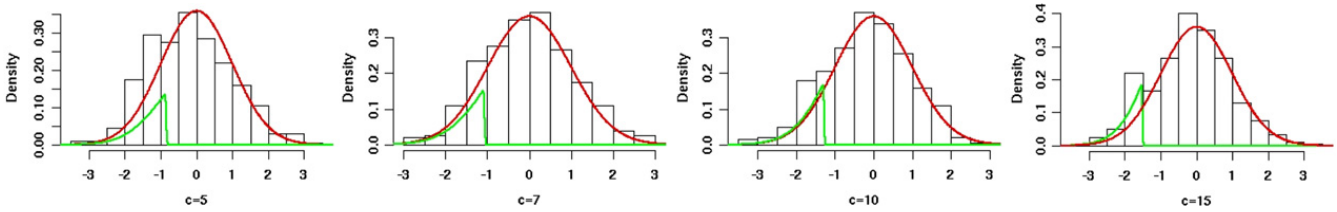


Fig. 1. Exemple of simulated datasets according to the value of c. The standard Gaussian distribution ϕ and the alternative f are represented (in red and green, respectively). Left: $c = 5$; right: $c = 15$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4.1. Simulation design

We simulate a binary HMM as described in Section 3, where f is non Gaussian and defined as the density of a random variable Y defined as follows:

$$Y = \Phi^{-1}(U), \quad \text{where } U \sim \mathcal{U}([0, 1/c]) \text{ and } c \in [5, 7, 10, 15].$$

The function Φ is the cdf of the standard Gaussian distribution. The known distribution ϕ is supposed to be distributed as a standard Gaussian ($\mathcal{N}(0, 1)$). Fig. 1 displays examples of the simulated datasets according to the value of c. We note that the overlap between the two groups is high when c is small. Therefore, the difficulty of the problem decreases with the parameter c. We also consider four different transition matrices which have the same form given by:

$$\Pi_u = \begin{pmatrix} 1 - lu & lu \\ l(1-u) & 1 - l(1-u) \end{pmatrix}, \tag{8}$$

where l is the shifting rate which varies from 0 to 1 and u corresponds to the proportion of the group of interest and is chosen within $\{0.05, 0.1, 0.2, 0.3\}$. For each of the 16 configurations we generate $P = 100$ datasets of size $n = 100$. The inference is performed for a semi-homogeneous case: for each simulation condition, we fit a 7-component Gaussian mixture with common variance σ^2 and means μ_k for the alternative. In a Bayesian context, the parameters are random variables with prior distributions. These prior distributions are chosen to be conjugate. Denoting by λ the precision parameter, $\lambda = \frac{1}{\sigma^2}$, we have:

- Transition matrix: for $j = 1, 2$, $\pi_{j\cdot} \sim \mathcal{D}(1, 1)$.
- Mixture proportions: $p \sim \mathcal{D}(1, \ldots, 1)$.
- Precision: $\lambda \sim \Gamma(0.01, 0.01)$.
- Means: $\mu_k|\lambda \sim \mathcal{N}\left(0, \frac{1}{0.01 \times \lambda}\right)$.

For calculating the weights α_m^{IS}, the numbers of draws has been fixed to $B = 5000$.

4.2. Results

We present the results for $l = 0.6$. We considered other values within the range $\{0.1, 0.2, 0.4, 0.6, 0.8, 0.9\}$ for this parameter but the performance is almost the same.

4.2.1. Accuracy of the weight

We consider importance sampling as a reference for weight estimation as it provides an unbiased estimate of the true weights whatever the approximation. We compared it to VB and PE weights by calculating the total variation distance, which quantifies the dissimilarity between two distributions α^1 and α^2:

$$\delta(\alpha^1, \alpha^2) = \frac{1}{2} \sum_x |\alpha^1(x) - \alpha^2(x)|. \tag{9}$$

The closer to 0 this distance is, the better the estimation of the weights.

Table 1 shows that VB weights are the closest to IS weights. The total variation distance $\delta(\alpha^{VB}, \alpha^{IS})$ is close to 0 whatever the simulation study. In contrast, the PE weights do not seem to be correct for approximating the true weights except when the two populations are well separated. These trends are also highlighted when we focus on the weights calculated for the P samples given a simulation condition. On average, compared to the PE approach, the VB method tends to provide weight estimations close to those of the IS approach. For instance, for $c = 7$ and $u = 0.2$, they mix three models with a huge weight ($\approx$0.70) for f_1 and weights around 0.15 for f_2 and f_3. However, the VB method has more stable estimated weights than IS. PE is the more stable approach among the three but it tends to only select the two-component model with an average weight around 0.95.

Conclusion on the weight estimation. By directly analysing the weight estimation, the similarities between the IS and the VB methods have clearly appeared. The VB method provides a good estimation of the true weights which is not the case for PE. Hence, when the computational time of the IS method becomes very high, we get a real advantage by using the VB method in terms of weight estimation.

Table 1
The total variation distance defined by (9) between the estimated weights with respect to importance sampling for each value of u and c.

c	$u = 0.05$		$u = 0.1$		$u = 0.2$		$u = 0.3$	
	PE	VB	PE	VB	PE	VB	PE	VB
5	0.419	0.069	0.370	0.101	0.453	0.120	0.456	0.069
7	0.438	0.096	0.403	0.101	0.287	0.101	0.257	0.072
10	0.386	0.092	0.271	0.180	0.232	0.115	0.107	0.092
15	0.372	0.093	0.303	0.158	0.258	0.129	0.102	0.101

4.2.2. Accuracy of the posterior probabilities

Once the weights have been estimated, the averaged estimates of the posterior probabilities T_t are computed for each approach. The aim of the VB method is to cluster the data into two populations. In many cases, these populations are difficult to distinguish but some observations are easily classifiable without any statistical approach. Hence, we put aside observations with a theoretical probability of belonging to the cluster of interest smaller than 0.2 or higher than 0.8. A classical indicator to measure the quality of a given classification is the MSE (Mean Square Error) which evaluates the difference between the averaged estimate $\widetilde{T}^{\mathcal{A}}$ of one method of $\mathcal{A}$ and the theoretical values $T^{(\mathrm{th})}$.

$$MSE^{\mathcal{A}} = \frac{1}{P}\sum_{p=1}^{P}\frac{1}{n}\sum_{t=1}^{n}(\widetilde{T}_{t,p}^{\mathcal{A}} - T_{t,p}^{(\mathrm{th})})^2. \tag{10}$$

The $MSE^{\mathcal{A}}$ estimation allows us to evaluate the quality of the estimates provided by Model m over all datasets $p = \{1, \ldots, P\}$ and one approach of $\mathcal{A}$. The smaller the MSE, the better the performances.

Since we deal with synthetic data, we can look at the best achievable MSE. This aims at minimising the MSE within the averaged estimator family to obtain an oracle weight. We denote this oracle by α^* and we have:

$$\alpha^* = \underset{\alpha}{\operatorname{argmin}} \left\| T^{(th)} - \sum_{m=1}^{M} \alpha_m \widehat{T}^{(m)} \right\|_2, \tag{11}$$

with $\sum_{m=1}^{M}\alpha_m = 1$ and $\forall m \in \{1, \ldots, M\}$, $0 \leq \alpha_m \leq 1$. The variable $\widehat{T}^{(m)}$ is the estimation of T supplied by model m. This oracle can be viewed as the weights we would choose if the theoretical posterior probability of belonging to the group of interest were known. This oracle estimator is obtained by a functional regression under non-negativity constraint and it can be written as:

$$\alpha^* = (\widehat{T}'\widehat{T})^{-1}\widehat{T}'T^{(th)} \times \gamma, \tag{12}$$

where γ is a normalising constant and $\widehat{T}$ is the matrix containing the estimates $\widehat{T}^{(m)}$ for all model m. Several algorithms allow this estimator to be calculate numerically by taking constraints into account. In this article, the optimisation has been achieved by the Newton–Raphson algorithm.

Fig. 2 displays the MSE calculated for the different methods under the various simulation conditions. First, we notice that the VB method based on the optimal variational weights provides good results in most of the cases. Moreover, we observe that an averaging approach with either the IS or VB method provides better results than the selected HMM. We observe that the PE method and the two-state-HMM provide worse estimates for many simulation conditions than do the VB and IS methods. Another comment is that there is no method which is the best whatever the simulation condition. Moreover, the estimations get closer to the oracle estimator as the problem becomes easier.

Fig. 3 shows the standard deviation of the MSE over all the simulation conditions. We notice that the VB method has one of the lowest variabilities. Once more, the two-state HMM has the worst performances.

Table 2 includes information on the misclassification for the three averaging approaches. The misclassification rate is calculated on the P samples whatever the simulation condition. The values in bold correspond to the smallest misclassification rate among the PE, VB and IS approaches. First, we note that the VB and the IS methods have very similar misclassification rates whatever the simulation condition. Moreover, this rate corresponds to the best rate of the three averaging methods. The averaged estimator supplied by the plug-in weights estimation seems to misclassify more data than the other approaches. Once again, Table 2 shows us that the VB approach provides good results when the simulation condition is complicated. In fact, when c equals either 5 or 7, the averaging method based on optimal variational weights provides the lowest misclassified rate among the three averaging approaches. Since the misclassification rate of the oracle is close to the rates obtained by VB and IS estimation, the two approaches provide good results for each value of c and u. Another comment is that the selected HMM approach always provides worse results than the IS and VB. This means that the averaging approach brings a gain to the posterior probability estimation.

Fig. 4 shows the entropy of the weights. We note that the optimal variational weights have one of the largest entropies among all the proposed weights. This means that the VB method tends to mix several models. Contrary to the other three weights, PE has a low entropy. This method seems to select only one model to infer posterior probability and does not take others into account.

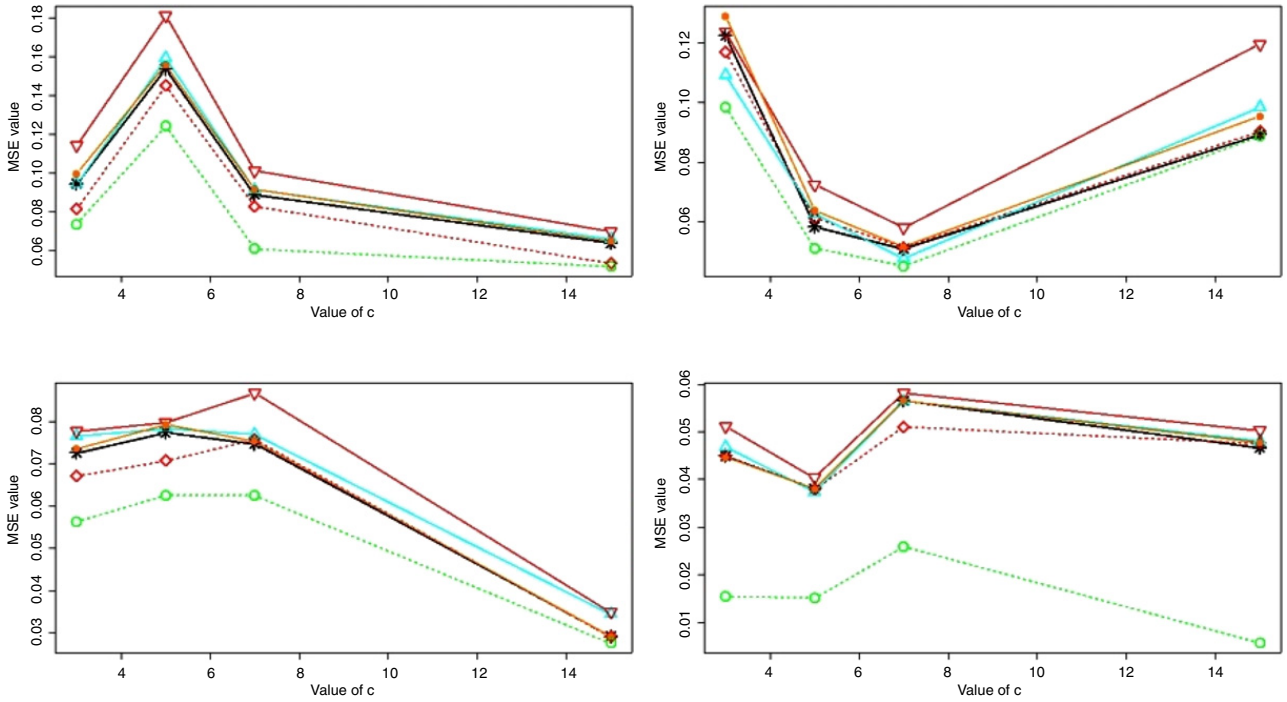


Fig. 2. Mean square error (MSE) between the true posterior probabilities and the estimates as a function of the uniform parameters calculated over the $P = 100$ datasets. Methods: "Δ": PE, "∇": two-state-HMM, "*": IS, "•": selected HMM, "$\diamond$": VB, "O" : Oracle. Top left: $\Pi_{0.05}$, top right: $\Pi_{0.1}$, bottom left: $\Pi_{0.2}$, bottom right: $\Pi_{0.3}$. VB and Oracle are in dotted lines.

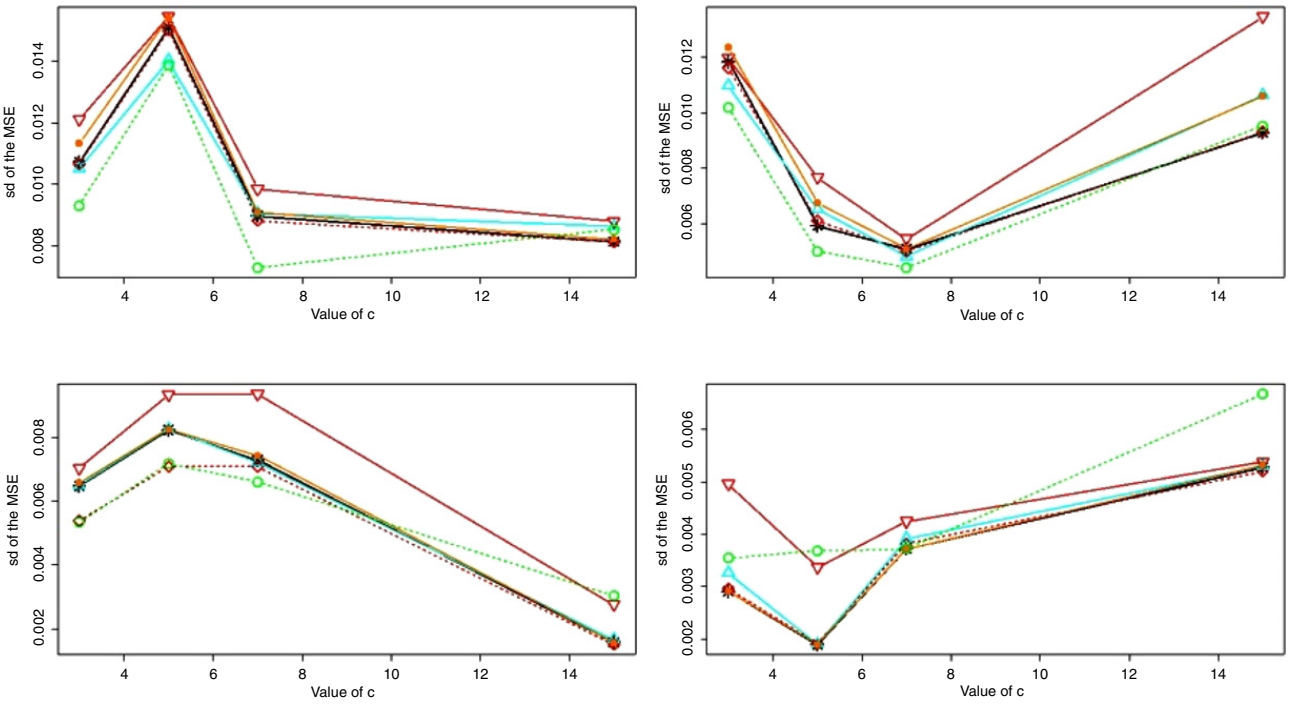


Fig. 3. Standard deviation of the MSE calculated over the $P = 100$ datasets. Methods: "Δ": PE, "∇": two-state-HMM, "*": IS, "•": selected HMM, "$\diamond$": VB, "O" : Oracle. Top left: $\Pi_{0.05}$, top right: $\Pi_{0.1}$, bottom left: $\Pi_{0.2}$, bottom right: $\Pi_{0.3}$. VB and Oracle are in dotted lines.

Conclusion on the accuracy of the estimates. Studying the MSE indicator allows us to compare the methods in terms of classification. Except for the "two-state-HMM" approach, we highlight that all the proposed methods have quite similar behaviours. However, the VB method provides better results in terms of MSE and its standard deviation than does the PE approach. These results are very close to those of IS and even often better. The focus on the misclassification rate confirmed the closeness between our approach and that of IS. These methods have a quite similar misclassification rate whatever the simulation condition. Furthermore, this rate corresponds to the best rate among the three averaging approaches. The

Table 2
Mean(sd) of the misclassification rate for the three averaging approaches.

	c	PE	VB	IS	Selected HMM	Oracle
$u = 0.05$	5	0.44 (0.04)	**0.36 (0.03)**	0.38 (0.04)	0.42 (0.03)	0.31 (0.02)
	7	0.54 (0.04)	**0.42 (0.04)**	0.43 (0.04)	0.47 (0.03)	0.34 (0.02)
	10	0.35 (0.04)	**0.30 (0.04)**	**0.30 (0.04)**	0.34 (0.04)	0.21 (0.03)
	15	0.38 (0.04)	0.34 (0.04)	**0.33 (0.04)**	0.36 (0.03)	0.23 (0.03)
$u = 0.1$	5	0.40 (0.04)	**0.37 (0.03)**	0.39 (0.03)	0.39(0.03)	0.29 (0.03)
	7	0.29 (0.03)	**0.23 (0.03)**	**0.23 (0.03)**	0.25 (0.03)	0.17 (0.02)
	10	0.28 (0.03)	0.28 (0.03)	**0.23 (0.03)**	0.28 (0.03)	0.16 (0.02)
	15	0.25 (0.04)	0.22 (0.03)	**0.20 (0.03)**	0.22 (0.03)	0.17 (0.02)
$u = 0.2$	5	0.33 (0.03)	**0.29 (0.03)**	0.30 (0.03)	0.31 (0.03)	0.19 (0.02)
	7	0.26 (0.03)	**0.23 (0.02)**	0.24 (0.02)	0.25 (0.02)	0.18 (0.02)
	10	0.23 (0.03)	0.20 (0.02)	**0.19 (0.02)**	0.23 (0.01)	0.17 (0.02)
	15	0.08 (0.01)	0.09 (0.01)	**0.07 (0.01)**	0.09 (0.01)	0.06 (0.02)
$u = 0.3$	5	0.23 (0.02)	**0.19 (0.01)**	0.20 (0.01)	0.22 (0.01)	0.16 (0.01)
	7	0.13 (0.01)	**0.11 (0.01)**	0.12 (0.01)	0.13 (0.01)	0.09 (0.01)
	10	0.17 (0.02)	0.12 (0.01)	**0.11 (0.01)**	0.18 (0.01)	0.03 (0.01)
	15	0.12 (0.01)	0.10 (0.01)	**0.09 (0.01)**	0.12 (0.01)	0.06 (0.01)

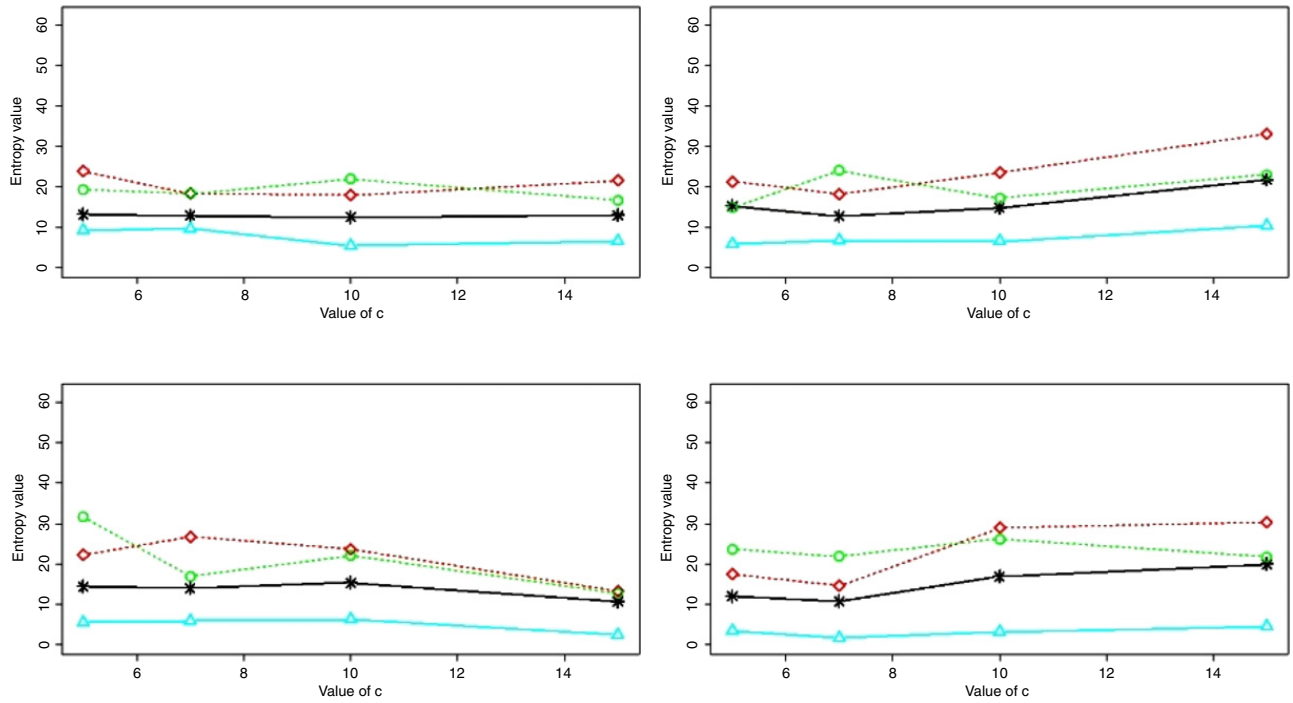


Fig. 4. Entropy of the weights calculated over the $P = 100$ datasets. Methods: "Δ": PE, "*": IS, "$\diamond$": VB, "O" : Oracle. Top left: $\Pi_{0.05}$, top right: $\Pi_{0.1}$, bottom left: $\Pi_{0.2}$, bottom right: $\Pi_{0.3}$. VB and Oracle are in dotted lines.

computational time is also a key point of these classification methods. Indeed, the VB method has a negligible computational time compared with IS. This may dramatically increase further with the size of the data.

4.3. Discussion

It is known that the true posterior distribution achieves the minimal variance of importance sampling estimates. Therefore, it may seem natural to use the variational posterior $Q_{H|M}^{VB}$ for importance sampling as its 'best' approximation. However, it is known that variational posterior often underestimates the posterior variance (Wang and Titterington, 2004). In Robert and Casella (2004), it is advisable to avoid importance functions with tails lighter than those of the distribution $P(H|X, M)$. In this case, the variances of the corresponding estimators could be infinite. In order to determine whether the underestimation of the posterior variance leads to a bad estimation of the weights we have carried out a simulation study. We performed the simulation in the most complicated configuration $u = 0.05$ and $c = 5$. In this configuration, we multiplied the variance of the variational posterior by an increasing parameter $\kappa \geq 1$. The results (not shown) show that the value of κ giving the minimal variance for the weights is 2.5. This variance is 1.6 times smaller than the one we get when using the variational posterior directly and both variances have the same order of magnitude (about 10^{-2}). Moreover, the

Table 3
Parameter estimation of the Gaussian mixture within the alternative distribution f.

m	Mean	Variance	Proportions	α^{VB}
1	4.9	1.1	1	$<10^{-4}$
2	(4.5, 5)	0.9	(0.67, 0.33)	$<10^{-4}$
3	(4, 4.2, 6)	0.3	(0.32, 0.32, 0.34)	0.34
4	(3.9, 4.1, 5, 6.3)	0.2	(0.22, 0.27, 0.26, 0.25)	0.66
5	(3.8, 4, 4.1, 5.2, 6.4)	0.18	(0.17, 0.19, 0.22, 0.22, 0.20)	$<10^{-4}$
6	(3.8, 4, 4.1, 4.8, 5.6, 6.5)	0.15	(0.14, 0.16, 0.16, 0.20, 0.16, 0.18)	$<10^{-4}$

total variation distance between the two sets of weights is 6×10^{-2}. We therefore conclude that, although it is suboptimal, this choice does not affect the evaluation of the respective model weights.

5. Illustration

5.1. Epidemiologic dataset

5.1.1. Description

The data. In this section, we focus on the analysis of a real dataset collected from public health surveillance systems. These data have also been studied in the recent paper of Sun and Cai (2009) using an FDR (False Discovery Rate) approach. The database is composed of 1216 time points. The data and log-transformation of them are shown in Fig. 5. The events described by the data can be classified into 2 groups: usual or unusual. These two groups correspond to a regular low rate and an irregular high rate respectively. Hence, the first group represents our group of interest and the other one the alternative. Moreover, it is clear that an event highly depends on the past and Le Strat and Carrat (1999) demonstrated that this kind of data can be described by using a two-state HMM. In this analysis, we thus aim at retrieving the two groups in the population and we want to estimate well the posterior probability of belonging to the group of interest.

Initialisation of the algorithm. To avoid any influence of the prior distributions, they have been chosen as described in Section 4.1. As considered in the simulation section, the alternative distribution has been fitted by a Gaussian mixture with common variance. The number of components m within the alternative distribution varies from 1 to 6 and the fixed distribution $\mathcal{N}(2.37, 0.76^2)$ has been chosen according to results of Sun and Cai (2009).

5.1.2. Results

For each number of components we infer the model parameters and estimate the weights with the VB method. The results we obtained are summarised in Table 3.

Every model presented in Table 3 has the same estimation of the transition matrix $\begin{pmatrix} 0.96 & 0.04 \\ 0.04 & 0.96 \end{pmatrix}$. In their article, Sun and Cai selected a model with two heterogeneous Gaussian distributions for the alternative. In our approach, due to the homogeneous assumption, the number of components increases and we keep two models with three and four components respectively. The other models have a low weight, smaller than 10^{-4}, and have no influence on the posterior probability estimation. We now focus on the classification provided by the averaged distribution and the 3-component model proposed by Sun and Cai (2009) and we notice that only 3 points differ between our approach and that of Cai. However if we focus on these three points, we observe that they correspond to points with a posterior probability close to 0.5. These points are on the borderline between the two classes. As our approach tends to increase the posterior probabilities (see Fig. 6), the epidemical ranges are greater with our approach. In two cases, the epidemics are declared earlier with the VB method than with that of Cai.

Fig. 6 (left) displays the averaged posterior probabilities against the estimations obtained by the model proposed by Cai. The first comment is that the two approaches provide close estimations. This is especially the case for probabilities smaller than 0.3 or greater than 0.7. These ranges correspond to low entropy areas. The main comment is that an averaging approach tends to refine posterior probabilities between 0.3 and 0.7. This high entropy area is considered as a difficult area for estimating the probabilities. In fact, it mainly corresponds to data points which are on the borderline between the two classes.

The fit of the averaged distribution is given in Fig. 6 (right). We note that this distribution provides a good fit to the data and it does not correspond to a mixture of Gaussian distributions.

5.2. Genomic dataset

5.2.1. Description

We now focus on a genomic dataset related to the model plant *Arabidopsis thaliana*. The experiment aims at identifying differentially expressed genes by comparing two experimental conditions on a specific microarray, a so-called tiling array, where probes cover the whole genome of the plant (Berard, 2011). For detecting the differentially expressed genes, we

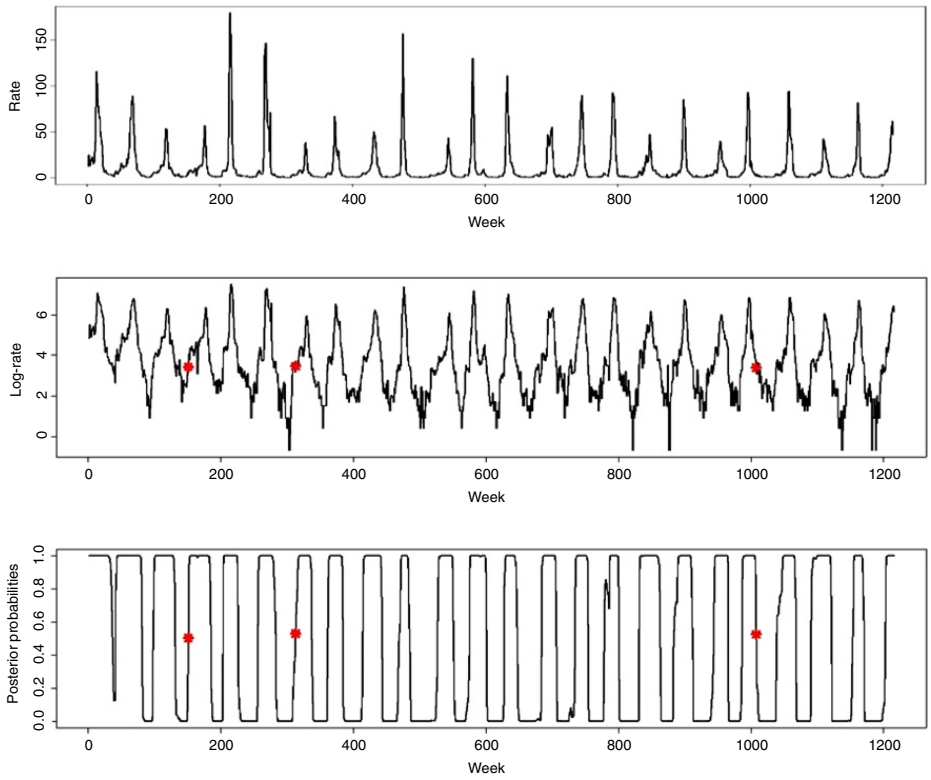


Fig. 5. Top: weekly ILI rate, middle: log-transformed weekly ILI rate, bottom: aggregated posterior probability of ILI epidemic over weeks. The three red points correspond to the points which have a different classification from one method to another.

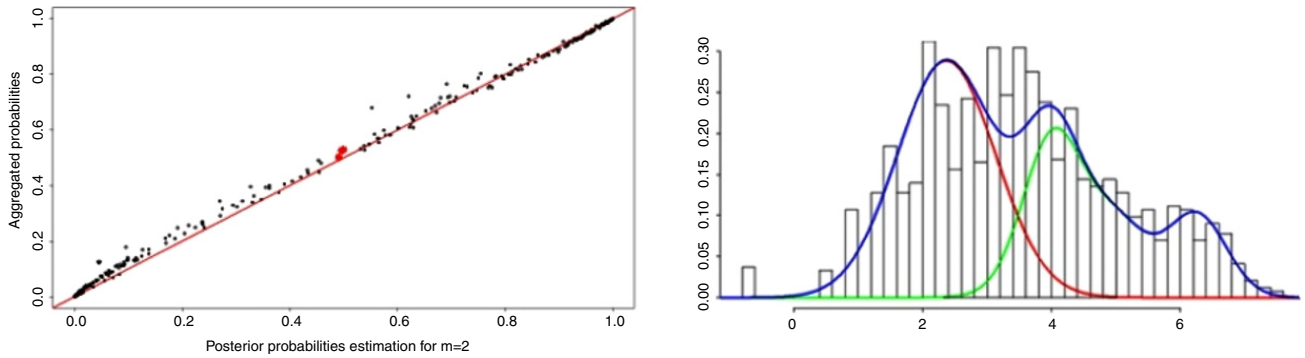


Fig. 6. Left: aggregated posterior probabilities according to the estimation of the posterior probabilities with the 2 heterogeneous component model. The three red points correspond to the points which have a different classification from one method to another. Right: fit of the known distribution ϕ, the averaged distribution and the global mixture (in red, green and blue respectively). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

applied the VARMIXT method (Delmar et al., 2005) and obtained a p-value p_i for each probe i. These p-values are distributed as a mixture of an uniform distribution on $[0;\ 1]$ ($\mathcal{U}([0;\ 1])$), referring to non-differentially expressed genes, and an unknown distribution, referring to differentially expressed genes. We applied a *probit* transformation on the p_i, so that the uniform distribution $U([0;\ 1])$ becomes a standard Gaussian ($\mathcal{N}(0,\ 1)$), while the other distribution remains unknown. We analysed a specific sample of 800 probes and we used the proposed averaging approach to detect transcriptional areas differentially expressed. We calculated the optimal variational weights for the models with 1, 2 and 3 Gaussian components for the alternative distribution f.

5.2.2. Results

Only the models with 1 and 2 Gaussians had a significant weight (see Table 4).

The model with one component has the largest weight, however model averaging leads to a different classification of the probe. These modifications are represented in Fig. 7 which displays the averaged posterior probabilities against the estimations of the model $m = 1$. We note that, once more, model averaging tends to refine posterior probabilities in high entropy areas.

Table 4
Parameter estimation of the Gaussian mixture within the alternative distribution f.

m	Mean	Variance	Proportions	α^{VB}
1	−2.32	0.16	1	0.73
2	(−2.32, −0.71)	0.15	(0.70, 0.30)	0.27

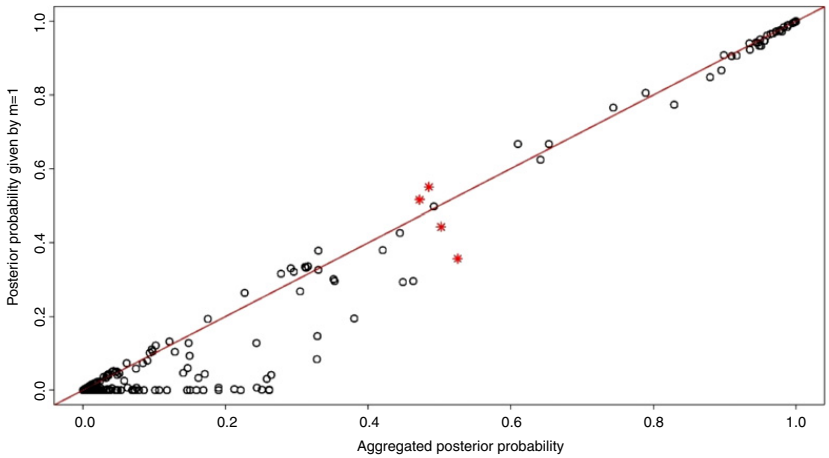


Fig. 7. Aggregated posterior probabilities according to the estimation of the posterior probabilities with the model $m = 2$. The four red points correspond to the points whose classification is modified.

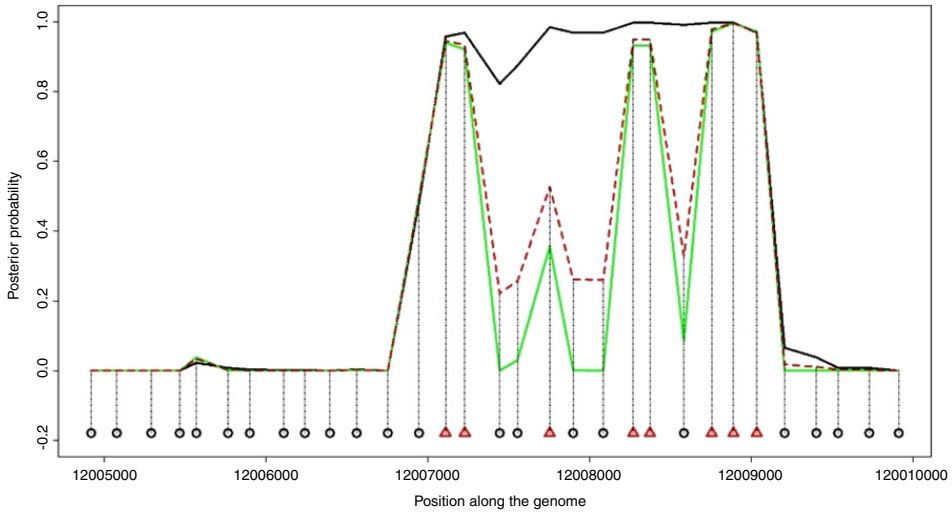


Fig. 8. Posterior probabilities of the two considered models (Green: $m = 1$ and Black: $m = 2$) and the averaged one (Dashed red line) for the region containing the studied gene. The classification achieved by the averaging approach is also represented: "Λ": differentially expressed and "O": non-differentially expressed. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

To study the changes in the classification, we focus on a 30-probe region containing gene AT4G22850, which is made up of four exons and three introns and covered by 11 probes. From a biological point of view, only exons are expected to be (differentially) expressed. Fig. 8 displays the posterior probability estimations supplied by models $m = 1$ and $m = 2$ and the averaged estimator. We note that the HMM with one or two Gaussians for the alternative distribution tends to smooth the posteriors.

For model $m = 2$, the 11 probes of the gene are declared as differentially expressed, and thus the introns are not detected. As for model $m = 1$, five successive probes are declared non-differentially expressed so the exon covered by only one probe is not detected. When considering an averaged estimator, this exon is detected so it seems that the smoothing effect due to the Markovian dependency is attenuated. The other points whose classification is modified mainly correspond to probes located at the beginning or the end of genes, where the annotation is known to be questionable.

6. Conclusion

We proposed a method for binary classification problems based on averaged estimators within a variational Bayesian framework. This approach allows us to avoid model selection and take model uncertainty into account. It can theoretically be proved that using an averaged estimator provides a gain in terms of MSE and increases the lower bound of the log-likelihood. We proposed a method based on optimal variational weights which derive from a modification of the classical lower bound of the log-likelihood. Our method does not require more computational time than the more commonly used selection approach. For studying performance, the method has been used on both synthetic and real data.

The results we obtained on synthetic data showed that our method enhances the estimator in terms of MSE in many simulation conditions. We also highlighted that the averaging approach improves the posterior probability estimation provided by the classical selection approach. Moreover, we showed that optimal variational weights are closer to importance sampling than the plug-in estimates. Since the importance sampling coped with computational time problems for high dimensional datasets, our method is of significant interest in this case.

An illustration on epidemiologic and transcriptomic datasets has been carried out to highlight the performance of the method we proposed on real datasets. In this context, the aggregation model still refines the estimation of posterior probabilities. We note in particular that the classification is different in cases where the probability is close to 0.5, i.e. when the classification is difficult. Model averaging allows us to refine the start of the epidemic period and to better determine the gene along the genome.

Acknowledgments

The authors thank the reviewers for their helpful comments which have enabled us to improve the presentation of our work.

References

Andrieu, C., 2003. An introduction to MCMC for machine learning. Machine Learning.
Baum, L., Petrie, T., Soules, G., Weiss, N., 1970. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. The Annals of Mathematical Statistics 41, 164–171.
Beal, M.J., Ghahramani, Z., 2003. The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures. Bayesian Statistics.
Berard, C., Martin-Magniette, M.L., Brunaud, V., Aubourg, S., Robin, S., 2011. Unsupervised classification for tiling arrays: chip–chip and transcriptome. Statistical Applications in Genetics and Molecular Biology.
Corduneanu, A., Bishop, C.M., 2001. Variational Bayesian model selection for mixture distributions. Statistics in Medicine 18, 3463–3478.
Delmar, P., Robin, S., Daudin, J.J., 2005. Varmixt: efficient variance modelling for the differential analysis of replicated gene expression data. Bioinformatics 21, 502–508.
Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T., 1999. Bayesian model averaging: a tutorial. Statistical Science 14, 382–417.
Jaakkola, T.S., Jordan, M.I., 1998, Improving the mean field approximation via the use of mixture distributions, pp. 163–173.
Le Strat, Y., Carrat, F., 1999. Monitoring epidemiologic surveillance data using hidden Markov models. Statistics in Medicine 18, 3463–3478.
Madigan, D., Hutchinson, F., 1995. Enhancing the predictive performance of Bayesian graphical models. Communications in Statistics: Theory and Methods 24.
Madigan, D., Raftery, A.E., 1993. Model selection and accounting for model uncertainty in graphical models using Occam's window. Journal of the American Statistical Association 1535–1546.
Marin, J., Robert, C.P., 2010, Importance sampling methods for Bayesian discrimination between embedded models, 0910.2325.
McGrory, C.A., Titterington, D.M., 2006. Variational Bayesian analysis for hidden Markov models. Australian & New Zealand Journal of Statistics 51, 227–244.
McLachlan, G.J., Bean, R.W., Peel, D., 2002. A mixture model-based approach to the clustering of microarray expression data. Bioinformatics 18, 413–422.
Parisi, G., 1988. Statistical Field Theory. Addison Wesley.
Raftery, A., Zheng, Y., 2003. Long-run performance of Bayesian model averaging. Journal of the American Statistical Association 98, 931–938.
Ren Q., B.S.F.A., Hodges, J., 2011. Variational Bayesian methods for spatial data analysis. Computational Statistics and Data Analysis 55, 3197–3217.
Robert, C.P., Casella,, 2004. Monte Carlo Statistical Methods, second ed. Springer, New York, volume.
Robin, S., Bar-Hen, A., Daudin, J., Pierre, L., 2007. A semi-parametric approach for mixture models: application to local false discovery rate estimation. Computational Statistics and Data Analysis 51, 5483–5493.
Ruggieri, E., Lawrence, C.E., 2011. On efficient calculations for Bayesian variable selection. Computational Statistics and Data Analysis.
Schwarz, G., 1978. Estimating the dimension of a model. The Annals of Statistics 6, 461–464.
Sun, W., Cai, T., 2009. Large-scale multiple testing under dependence. Journal of the Royal Statistical Society 71, 393–424.
Volinsky, C.T., Madigan, D., Raftery, A.E., Kronmal, R.A., 1997. Bayesian model averaging in proportional hazard models: assessing the risk of a stroke. Applied Statistics 433–448.
Wainwright, M.J., Jordan, M.I., 2008. Graphical Models, Exponential Families, and Variational Inference. Hanover, MA, USA.
Wang, B., Titterington, D.M., 2003a. Lack of consistency of mean field and variational Bayes approximations for state space models. Neural Processing Letters 20, 151–170.
Wang, B., Titterington, D.M., 2003b, Local convergence of variational Bayes estimators for mixing coefficients, Technical report.
Wang, B., Titterington, D.M., 2004a. Convergence and Asymptotic Normality of Variational Bayesian Approximations for Exponential Family Models with Missing Values. AUAI Press.
Wang, B., Titterington, D.M., 2004b, Inadequacy of interval estimates corresponding to variational Bayesian approximations.