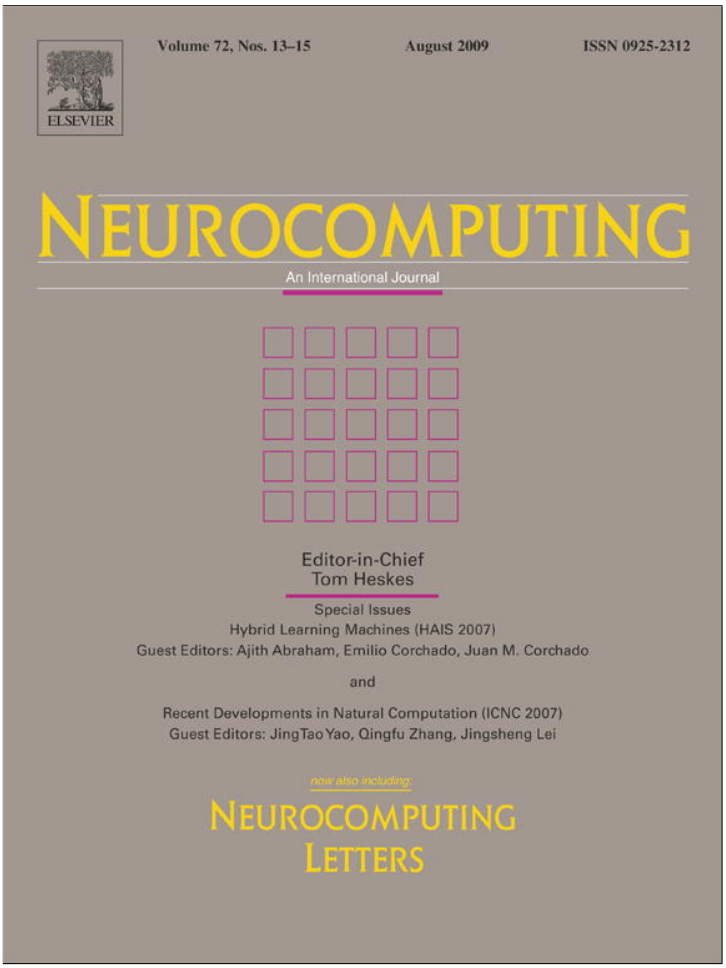

Neurocomputing 72 (2009) 3366–3369

Contents lists available at ScienceDirect

Neurocomputing

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

Letters

A novel view of the variational Bayesian clustering

Takashi Takekawa [a,*], Tomoki Fukai [a,b]

[a] Laboratory for Neural Circuit Theory, RIKEN Brain Science Institute (BSI), 2-1 Hirosawa, Wako, Saitama 351-0198, Japan
[b] Department of Complexity Science and Engineering, University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8561, Japan

ARTICLE INFO

Article history:
Received 30 June 2008
Accepted 16 April 2009
Communicated by L.C. Jain
Available online 10 May 2009

Keywords:
Unsupervised learning
Bayesian estimation
Variational approximation
Model selection
Robust variational Bayes

ABSTRACT

We prove that the evaluation function of variational Bayesian (VB) clustering algorithms can be described as the log likelihood of given data minus the Kullback–Leibler (KL) divergence between the prior and the posterior of model parameters. In this novel formalism of VB, the evaluation functions can be explicitly interpreted as information criteria for model selection and the KL divergence imposes a heavy penalty on the posterior far from the prior. We derive the update process of the variational Bayesian clustering with finite mixture Student's *t*-distribution, taking the penalty term for the degree of freedoms into account.

1. Introduction

Data observed in experiments or statistical researches can often be categorized into clusters of data points with similar features. Such a clustering, however, is not trivially easy for various reasons: the features used for it may not be optimal, or the data set may include noisy elements. Here, we reformulate the variational Bayesian (VB) method to solve difficult clustering problems.

Let m be the number of clusters and $\theta = \{\alpha_1, \alpha_2, \ldots, \alpha_m, \beta_1, \beta_2, \ldots, \beta_m\}$ be the set of parameters, where α and β represent the size and other parameters of cluster k, respectively. When m and θ are given, we can define a finite mixture model by the following conditional probability $p(x_n, z_n = k|\theta, m)$ that the n-th data takes a value of $x_n \in \mathscr{R}^D$ and belongs to the k-th cluster with probability α_k:

$$p(x_n, z_n = k|\theta, m) = \alpha_k \mathscr{A}(x_n|\beta_k). \tag{1}$$

Here, $\mathscr{A}(x_n|\beta_k)$ is some probability distribution to generate x_n, and variable $z_n \in \{1, 2, \ldots, m\}$ specifies the cluster from which x_n is taken.

Conversely, we want to estimate the optimal values of m, θ and $\mathbf{z} = \{z_1, z_2, \ldots, z_N\}$ when data points are given as $\mathbf{x} = \{x_1, x_2, \ldots, x_N\}$. If the prior distribution is set as $p(\theta, m) = p(m)p(\theta|m)$, we can in principle estimate those quantities by calculating the marginal likelihood

$$p(\mathbf{x}) = \sum_m p(m) \int p(\theta|m) \sum_n \sum_k p(x_n, z_n = k|\theta, m)\, d\theta \tag{2}$$

and the posterior distributions:

$$p(m|\mathbf{x}) = \frac{p(m) \int p(\theta|m) \sum_n \sum_k p(x_n, z_n = k|\theta, m)\, d\theta}{p(\mathbf{x})}, \tag{3}$$

$$p(z_n = k|\mathbf{x}, m) = \frac{\int p(\theta|m) p(x_n, z_n = k|\theta, m)\, d\theta}{p(\mathbf{x})}. \tag{4}$$

In practice, however, the above high-dimensional integrals are difficult to calculate. To overcome this difficulty, we may employ Markov Chain Monte Carlo (MCMC) methods for performing the above integrations (see [6]). In MCMC, however, the convergence to a correct solution is generally slow and is ensured only on a strict condition.

VB gives a deterministic algorithm to find approximate solutions to the posterior distributions [3]. The algorithm of VB is fast and resembles expectation–maximization (EM) algorithm for the maximum likelihood estimation [4]. In particular, optimizing the resultant clusters does not use information criteria, such as AIC [1] or BIC [7]. Rather, the model selection in VB is automatically accomplished by maximizing an estimation function. VB often shows a better generalization ability than EM when the number of data points is small. Clustering by VB can also be improved by using mixtures of principal component analyzers [5]. However, the formalism of VB is not transparent since the estimation function does not provide a simple information-theoretic interpretation.

* Corresponding author.
E-mail address: takekawa@riken.jp (T. Takekawa).

0925-2312/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2009.04.003

In this report, we show that the estimation function can be expressed as a sum of the log likelihood and a penalty term that has a simple geometrical meaning. Namely, the penalty term can be expressed as the Kullback–Leibler divergence of the prior and the posterior of model parameters. This result achieves a deeper insight into the theoretical framework of VB. Moreover, the result provides a more sophisticated software implementation than the previous ones. To show this explicitly, we derive a VB algorithm by using a mixture of Student's t-distributions. Such an algorithm is known to be more robust against noisy data points (i.e., outliers) than the algorithm with Gaussian mixtures [2,8].

The previous algorithm of VB with Student's t-mixtures assumed a uniform distribution for the priori of the degree of freedom v [2]. This prior, however, makes the calculations of the penalty term of v difficult, so the term was simply neglected in the formula. Here, we derive a natural expression of the penalty term by employing an exponential distribution for the prior of v. Preliminary results obtained in the analysis of multi-unit recording data [9] revealed that our algorithm significantly improves the speed and precision of spike sorting. The results of spike sorting will be reported elsewhere.

2. Preparations

Below we explain our framework of variational Bayesian formulated with a mixture of Student's t-distributions.

2.1. Approximation

The Student's t-distribution can be described as follows:

$$\mathscr{T}(x|\nu,\mu,\Sigma)=\int_0^\infty \mathscr{N}(x|\mu,u^{-1}\Sigma)\,\mathscr{G}\Big(u\Big|\frac{\nu}{2},\frac{\nu}{2}\Big)\,d\mathbf{u}, \tag{5}$$

where the normal distribution $\mathscr{N}$ and the gamma distribution $\mathscr{G}$ are given as

$$\mathscr{N}(x|\mu,\Sigma)=\frac{1}{|2\pi\Sigma|^{1/2}}\exp\Big\{-\frac{1}{2}\operatorname{tr}\Sigma^{-1}(x-\mu)(x-\mu)^{\mathrm{T}}\Big\}, \tag{6}$$

$$\mathscr{G}(u|a,b)=\frac{b^a}{\Gamma(a)}u^{a-1}\exp(-bu). \tag{7}$$

We define a probability distribution involving unobserved latent variable u_{nk} as

$$p(x_n,z_n=k,u_{nk}|\theta,m)=\alpha_k\mathscr{N}(x_n|\mu_k,u_{nk}^{-1}\Sigma_k)\,\mathscr{G}\Big(u_{nk}\Big|\frac{\nu_k}{2},\frac{\nu_k}{2}\Big), \tag{8}$$

where ν_k, μ_k and Σ_k represent the degree of freedom, mean and variance matrix of cluster k, respectively. Then, we introduce a functional of test function $q(\mathbf{z},\mathbf{u},\theta,m)$, which should approximate the posterior distribution $p(\mathbf{z},\mathbf{u},\theta,m|\mathbf{x})$, as

$$\begin{aligned}F[q(\mathbf{z},\mathbf{u},\theta,m)]&=\sum_m\sum_{\mathbf{z}}\int\int q(\mathbf{z},\mathbf{u},\theta,m)\log\frac{p(\mathbf{x},\mathbf{z},\mathbf{u},\theta,m)}{q(\mathbf{z},\mathbf{u},\theta,m)}d\theta\,d\mathbf{u}\\&=\log p(\mathbf{x})-\mathrm{KL}[q(\mathbf{z},\mathbf{u},\theta,m),p(\mathbf{z},\mathbf{u},\theta,m;\mathbf{x})].\end{aligned} \tag{9}$$

Here, Kullback–Leibler divergence

$$\mathrm{KL}[q(\mathbf{z},\mathbf{u},\theta,m),p(\mathbf{z},\mathbf{u},\theta,m|\mathbf{x})]=\sum_m\sum_{\mathbf{z}}\int\int q(\mathbf{z},\mathbf{u},\theta,m)\log\frac{q(\mathbf{z},\mathbf{u},\theta,m)}{p(\mathbf{z},\mathbf{u},\theta,m|\mathbf{x})}d\theta\,d\mathbf{u} \tag{10}$$

vanishes when the test distribution coincides with the posterior. Since $p(\mathbf{x})$ takes a constant for given $\mathbf{x}$, minimizing $\mathrm{KL}[q(\mathbf{z},\mathbf{u},\theta,m),p(\mathbf{z},\mathbf{u},\theta,m|\mathbf{x})]$ is equivalent to maximizing F and hence allows $q(\mathbf{z},\mathbf{u},\theta,m)$ to approach $p(\mathbf{z},\mathbf{u},\theta,m|\mathbf{x})$.

2.2. Model selection

To maximize F, we introduce the following function:

$$F_m[q(\mathbf{z},\mathbf{u},\theta|m)]=\sum_{\mathbf{z}}\int\int q(\mathbf{z},\mathbf{u},\theta|m)\log\frac{p(\mathbf{x},\mathbf{u},\mathbf{z},\theta|m)}{q(\mathbf{z},\mathbf{u},\theta|m)}d\theta\,d\mathbf{u}. \tag{11}$$

By using the above function and $q(\mathbf{z},\mathbf{u},\theta,m)=q(m)q(\mathbf{z},\mathbf{u},\theta|m)$, we can rewrite F as

$$F=\sum_m\Big\{q(m)F_m-q(m)\log\frac{q(m)}{p(m)}\Big\}. \tag{12}$$

By maximizing F with respect to $q(m)$, we obtain

$$q(m)\propto p(m)\exp F_m. \tag{13}$$

Since no prior information is available for m, we may regard $p(m)$ as constant. Therefore, we should select the value of m that maximizes F_m as the optimal number of clusters. Thus, we have to solve the maximization problem of F_m.

3. Implementations

Hereafter, we explain our algorithm to maximize F_m. For brevity, we do not explicitly show the dependence of m in the argument and write $q(a|m),p(a|m)$ as $q(a),p(a)$ in the abbreviated form. In addition, we assume that the following factorization holds: $q(z,u,\theta)=q(z,u)q(\theta)$. With this assumption, Eq. (11) can be transformed into

$$F_m[q(\mathbf{z},\mathbf{u}),q(\theta)]=\sum_{\mathbf{z}}\int\int q(\mathbf{z},\mathbf{u})q(\theta)\log\frac{p(\mathbf{x},\mathbf{z},\mathbf{u}|\theta)p(\theta)}{q(\mathbf{z},\mathbf{u})q(\theta)}d\theta\,d\mathbf{u}. \tag{14}$$

3.1. Prior

The prior of the model parameters can be decomposed as $p(\theta)=p(\boldsymbol{\alpha})\prod_k\{p(\nu_k)p(\Sigma_k)p(\mu_k|\Sigma_k)\}$. We use the following distributions in this study:

$$p(\boldsymbol{\alpha})=\mathscr{D}(\{\alpha_1,\alpha_2,\ldots,\alpha_m\}|\{\kappa_0,\kappa_0,\ldots,\kappa_0\}), \tag{15}$$

$$p(\nu_k)=\xi_0\exp(-\xi_0\nu_k), \tag{16}$$

$$p(\Sigma_k)=\mathscr{W}^{-1}(\Sigma_k|\gamma_0,\gamma_0\Sigma_0), \tag{17}$$

$$p(\mu_k|\Sigma_k)=\mathscr{N}(\mu_k|\mu_0,\eta_0^{-1}\Sigma_k), \tag{18}$$

where the Dirichlet distribution $\mathscr{D}$ and the inverse Wishart distribution $\mathscr{W}^{-1}$ are given as

$$\mathscr{D}(\{\alpha_1,\ldots,\alpha_m\}|\{\kappa_1,\ldots,\kappa_m\})=\frac{\Gamma(\sum_k\kappa_k)}{\prod_k\Gamma(\kappa_k)}\prod_k\alpha_k^{\kappa_k-1}, \tag{19}$$

$$\mathscr{W}^{-1}(\Sigma|\gamma,\Delta)=\frac{|\frac{1}{2}\Delta|^{\gamma/2}\exp(-\frac{1}{2}\operatorname{tr}\Delta\Sigma^{-1})}{\Gamma_D(\gamma/2)|\Sigma|^{1/2(\gamma+D+1)}} \tag{20}$$

and Γ_D is the multivariate gamma function with dimension D. The conventional distributions are used for $\boldsymbol{\alpha}$, μ_k and Σ_k. As mentioned before, we adopted an exponential distribution for ν_k rather than the uniform distribution chosen previously [2].

3.2. Initial conditions

VB algorithm alternately performs an E step and an M step. An E step computes the expectation of the likelihood as if the latent variables were observable, and an M step computes the maximum likelihood estimates of the parameters by maximizing the expected likelihood found on the preceding E step. The parameters found on the M step are then used to begin another E step.

The value of optimization function F_m in Eq. (14) is evaluated after every E step, and the process is repeated until the value of F_m converges to a steady value.

The test distributions can be obtained in the derivation of the M step. Below, hyper-parameters of the test distribution are shown with tilde. As in the case of the prior of the model parameters $p(\theta)$, the test distribution of the model parameters can also be decomposed as $q(\theta)=q(\boldsymbol{\alpha})\prod_k\{q(v_k)q(\Sigma_k)q(\mu_k|\Sigma_k)\}$. We use the following distributions in this study:

$$q(\boldsymbol{\alpha})=\mathscr{D}(\{\alpha_1,\alpha_2,\ldots,\alpha_m\}|\{\tilde{\kappa}_1,\tilde{\kappa}_2,\ldots,\tilde{\kappa}_m\}), \tag{21}$$

$$q(v_k)=\frac{(v_k/2)^{v_k/2}\exp(-\tilde{\xi}_k v_k)}{C_v(\tilde{\xi}_k)\Gamma(v_k/2)}, \tag{22}$$

$$q(\Sigma_k)=\mathscr{W}^{-1}(\Sigma_k|\tilde{\gamma}_k,\tilde{\gamma}_k\tilde{\Sigma}_k), \tag{23}$$

$$q(\mu_k|\Sigma_k)=\mathscr{N}(\mu_k|\tilde{\mu}_k,\tilde{\eta}_k^{-1}\Sigma_k), \tag{24}$$

where Γ is the gamma function. For the convenience of practical computations, we may replace $q(v_k)$ with $\delta(v_k-\tilde{v}_k)$ by MAP approximation, where $\tilde{v}_k$ is the value of v that maximizes $q(v_k)$. Thus, $\tilde{v}_k$ is a solution to $dq(v_k)/dv_k=0$. Normalization constant $C_\mu(\xi)$ is determined as

$$C_\mu(\xi)=\int\frac{(v/2)^{v/2}\exp(-\xi v)}{\Gamma(v/2)}\,dv. \tag{25}$$

3.3. E step

On an E step, $F_m[q(\mathbf{z},\mathbf{u}),q(\theta)]$ is maximized with respect to $q(\mathbf{z},\mathbf{u})$ while $q(\theta)$ is fixed. We can find such a $q(\mathbf{z},\mathbf{u})$ as

$$q(z_n=k,u_{nk})=\frac{\zeta_{nk}}{\sum_{k'}\rho_{nk'}} \tag{26}$$

by using the Lagrange multiplier method, where

$$\begin{aligned}\zeta_{nk}&=\exp\int q(\theta)\log p(x_n,z_n=k,u_{nk}|\theta)\,d\theta\\&=\frac{\hat{\alpha}_k(\tilde{v}_k/2)^{\tilde{v}_k/2}}{(2\pi)^{D/2}\hat{\Sigma}^{1/2}\Gamma(\tilde{v}_k/2)}u_{nk}^{(a_{nk}-1)}\exp(-b_{nk}u_{nk}),\end{aligned} \tag{27}$$

$$\rho_{nk}=\int\zeta_{nk}\,du_{nk}=\frac{\hat{\alpha}_k(\tilde{v}_k/2)^{\tilde{v}_k/2}}{(2\pi)^{D/2}\hat{\Sigma}^{1/2}\Gamma(\tilde{v}_k/2)}\Gamma(a_{nk})b_{nk}^{-a_{nk}}. \tag{28}$$

The parameters appearing in the above equations are defined in Appendix A. Then, using

$$q(u_{nk}|z_n=k)=\frac{\zeta_{nk}}{\rho_{nk}}=\mathscr{G}(u_{nk}|a_{nk},b_{nk}) \tag{29}$$

we can derive the following quantities:

$$\bar{z}_{nk}=q(z_n=k)=\frac{\rho_{nk}}{\sum_{k'}\rho_{nk'}}, \tag{30}$$

$$\bar{u}_{nk}=\int q(u_{nk}|z_n=k)u_{nk}\,du_{nk}=\frac{a_{nk}}{b_{nk}}, \tag{31}$$

$$\begin{aligned}\log\hat{u}_{nk}&=\int q(u_{nk}|z_n=k)\log u_{nk}\,du_{nk}\\&=\Psi(a_{nk})-\log b_{nk},\end{aligned} \tag{32}$$

which are used on the succeeding M step, where Ψ is the digamma function.

3.4. Evaluation

The output of each E step is evaluated by the value of $F_m[q(\mathbf{z},\mathbf{u}),q(\theta)]$ before a succeeding M step is processed. We first rewrite Eq. (14) as

$$\begin{aligned}F_m=&\sum_{\mathbf{z}}\int\int q(\mathbf{z},\mathbf{u})q(\theta)\log\frac{p(\mathbf{x},\mathbf{z},\mathbf{u}|\theta)}{q(\mathbf{z},\mathbf{u})}\,d\theta\,d\mathbf{u}\\&-\sum_{\mathbf{z}}\int q(\mathbf{z},\mathbf{u})\,d\mathbf{u}\int q(\theta)\log\frac{q(\theta)}{p(\theta)}\,d\theta.\end{aligned} \tag{33}$$

Since $\sum_{\mathbf{z}}\int q(\mathbf{z},\mathbf{u})\,d\mathbf{u}=1$ holds, the second term coincides with $\mathrm{KL}[q(\theta),p(\theta)]$.

Then, using (27), we can transform the first term into

$$\sum_n\sum_k\int q(z_n=k,u_{nk})\{\log\zeta_{nk}-\log q(z_n=k,u_{nk})\}\,du_{nk}, \tag{34}$$

which we can further rewrite as

$$\begin{aligned}&\sum_n\sum_k\int q(z_n=k,u_{nk})\left\{\log\zeta_{nk}-\log\frac{\zeta_{nk}}{\sum_{k'}\rho_{nk'}}\right\}du_{nk}\\&=\sum_n\sum_k\int q(z_n=k,u_{nk})\,du_{nk}\log\sum_{k'}\rho_{nk'}\\&=\sum_n\log\sum_{k'}\rho_{nk'}\end{aligned} \tag{35}$$

by using (26) and $\sum_k\int q(z_n=k,u_{nk})\,du_{nk}=1$. Thus, we finally obtain the following expression of the estimation function:

$$F_m=\sum_n\log\sum_k\rho_{nk}-\mathrm{KL}[q(\theta),p(\theta)]. \tag{36}$$

The above results reveal that the estimation function consists of the log likelihood given by $\sum_k\rho_{nk}$, which has already been calculated at the preceding E step, and the penalty term calculated from the test function $q(\theta)$. Thus, VB method maximizes the log likelihood calculated from the test function with a penalty given as the deviation of the test function from the prior. Note that our evaluation of F_m no longer requires heavy calculations of the multi-dimensional integrals in Eq. (33), whereas the conventional evaluation step calculates the integrals directly.

3.5. M step

An M step completes one cycle of EM algorithm. M step is complementary to E step, that is, $F_m[q(\mathbf{z},\mathbf{u}),q(\theta)]$ is maximized with respect to $q(\theta)$ while $q(\mathbf{z},\mathbf{u})$ is fixed. Such a $q(\theta)$ can be expressed in a fashion similar to Eqs. (21)–(24), and the hyper-parameters are derived as

$$\tilde{\kappa}_k=\kappa_0+\bar{N}_k, \tag{37}$$

$$\tilde{\gamma}_k=\gamma_0+\bar{N}_k, \tag{38}$$

$$\tilde{\eta}_k=\eta_0+\bar{M}_k, \tag{39}$$

$$\tilde{\xi}_k=\xi_0+\frac{\bar{M}_k-\hat{M}_k}{2\bar{N}}, \tag{40}$$

$$\tilde{\mu}_k=\frac{\eta_0\mu_0+\bar{M}_k\bar{\mu}_k}{\eta_0+\bar{M}_k}, \tag{41}$$

$$\tilde{\Sigma}_k=\frac{1}{\gamma_0+\bar{N}_k}\left\{\gamma_0\Gamma_0+\bar{M}_k\bar{\Sigma}_k+\frac{\eta_0\bar{M}_k}{\eta_0+\bar{M}_k}(\bar{\mu}_k-\mu_0)(\bar{\mu}_k-\mu_0)^{\mathrm{T}}\right\}. \tag{42}$$

The parameters appearing in the above equations are defined in Appendix A.

4. Discussion

If we take the limit of vanishing ξ_0 in Eq. (40), the prior approaches a uniform distribution and the renewal rule coincides

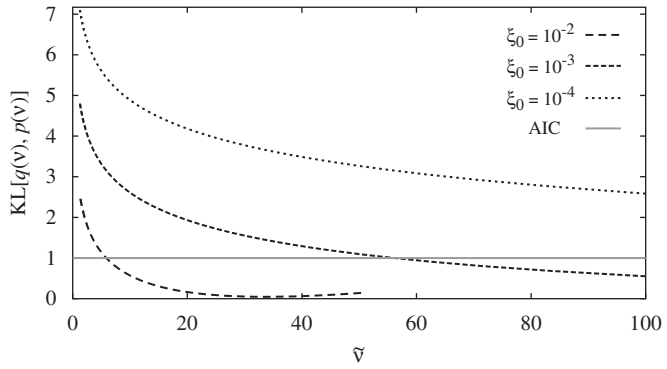

Fig. 1. Penalty terms of ν for various values of ξ_0.

with the previous one [2], in which the penalty term diverges, as shown below. By contrast, too large ξ_0 implies too strong influences of the prior. Thus, an appropriate value range of ξ_0 should exist. This point was studied in an example.

In Fig. 1, we plot the values of the penalty term of ν (i.e., $\mathrm{KL}[q(\nu), p(\nu)]$) for the estimated values of $\tilde{\nu}$. Three values of ξ_0 were tested. As a reference, we calculated the penalty term that should be added to AIC, which is 1, by the inclusion of ν. Note that at $\xi_0 = 10^{-2}$ the estimated value of $\tilde{\nu}$ cannot be larger than 50. When ξ_0 is smaller than 10^{-4}, the prior distribution is very close to a uniform distribution. However, the penalty term is rapidly increased for such a small value of ξ_0. We suggest that an appropriate value of the hyper-parameter of the prior distribution is $\xi_0 \approx 10^{-3}$.

5. Conclusions

In this letter, we have derived a simple expression for the estimation function of VB method. This expression is mathematically transparent and is easy to calculate numerically. We applied the obtained framework to derive the penalty term and the renewal rule for the degrees of freedom ν of the robust VB method.

Acknowledgments

This research was partially supported by RIKEN Special Postdoctoral Researchers Program (to T. Takekawa) and MEXT Grant-in-Aid for Scientific Research on Priority Areas 17022036 (to T. Fukai).

Appendix A

The parameters used in Eqs. (27)–(28) are defined as follows:

$$\log \hat{\alpha}_k = \int q(\boldsymbol{\alpha}) \log \alpha_k \, d\boldsymbol{\alpha} = \Psi(\tilde{\kappa}_k) - \Psi\left(\sum_{k'} \tilde{\kappa}_{k'}\right), \tag{43}$$

$$\log \hat{\Sigma}_k = \int q(\Sigma_k) \log |\Sigma_k| \, d\Sigma_k = \log \left| \frac{\tilde{\gamma}_k}{2} \tilde{\Sigma}_k \right| - \sum_{i=0}^{D-1} \Psi\left(\frac{\tilde{\gamma}_k - i}{2}\right), \tag{44}$$

$$a_{nk} = \tfrac{1}{2}(\tilde{\nu}_k + D), \tag{45}$$

$$b_{nk} = \tfrac{1}{2}\{\tilde{\nu}_k + D/\tilde{\eta}_k + \mathrm{tr}\, \tilde{\Sigma}_k^{-1}(x_n - \tilde{\mu}_k)(x_n - \tilde{\mu}_k)^{\mathrm{T}}\}. \tag{46}$$

The definitions of the parameters used in Eqs. (37)–(42) are listed as follows:

$$\bar{N}_k = \sum_n \bar{z}_{nk}, \tag{47}$$

$$\bar{M}_k = \sum_n \bar{z}_{nk} \bar{u}_{nk}, \tag{48}$$

$$\hat{M}_k = \sum_n \bar{z}_{nk} \log \hat{u}_{nk}, \tag{49}$$

$$\bar{\mu}_k = \frac{1}{\bar{M}_k} \sum_n \bar{z}_{nk} \bar{u}_{nk} x_n, \tag{50}$$

$$\bar{\Sigma}_k = \frac{1}{\bar{M}_k} \sum_n \bar{z}_{nk} \bar{u}_{nk} (x_n - \bar{\mu}_k)(x_n - \bar{\mu}_k)^{\mathrm{T}}. \tag{51}$$

References

[1] H. Akaike, A new look at the statistical model identification, IEEE Transactions on Automatic Control 19 (1974) 716–723.
[2] A. Archambeau, M. Verleysen, Robust Bayesian clustering, Neural Networks 20 (2007) 129–138.
[3] H. Attias, Inferring parameters and structure of latent variable models by variational Bayes, in: Proceeding of the 15th Conference on Uncertainty in Artificial Intelligence, 1999, pp. 21–30.
[4] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum-likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistics Society Series B 39 (1977) 1–38.
[5] S. Oba, M. Sato, S. Ishii, Variational Bayes method for mixture of principal component analyzers, in: Proceeding for 7th International Conference on Neural Information Processing, vol. 2, 2000, pp. 1416–1421.
[6] S. Richardson, P. Green, On Bayesian analysis of mixtures with unknown number of components, Journal of the Royal Statistical Society Series B 59 (1997) 731–792.
[7] G. Schwarz, Estimating the dimension of a model, Annals of Statistics 6 (1978) 461–464.
[8] M. Svensén, C.M. Bishop, Robust Bayesian mixture modeling, Neurocomputing 64 (2005) 235–252.
[9] T. Takekawa, S. Kang, Y. Isomura, T. Fukai, Robust and accurate spike sorting with matching pursuit and variational Bayesian clustering, in: Abstract of the 37th Annual Meeting of the Society for Neuroscience, 2007, 790.7.

Takashi Takekawa received the Ph.D. degree in Informatics from Kyoto University, Kyoto, in 2005. Currently, he is a Special Postdoctoral Researcher in RIKEN Brain Science Institute (BSI). His research interests include computational functions of neuronal dynamics. Since 2007, he has developed an accurate spike sorting method for more detailed analysis of experimental data.

Tomoki Fukai received a Ph.D degree in Physics from Waseda University, Tokyo, in 1985. Since 2005, he has been serving as the director of Theoretical Neuroscience Group of RIKEN Brain Science Institute. His main interests are in modeling dynamics and computational functions of neuronal networks.