# MINIMUM VARIATIONAL STOCHASTIC COMPLEXITY AND AVERAGE GENERALIZATION ERROR IN LATENT VARIABLE MODELS

*Kazuho Watanabe*

Graduate School of Information Science, Nara Institute of Science and Technology,
8916-5, Takayama-cho, Ikoma, Nara, 630-0192, JAPAN, wkazuho@is.naist.jp

## ABSTRACT

Bayesian learning is often accomplished with approximation schemes because it requires intractable computation of the posterior distributions. In this paper, focusing on the approximation scheme, variational Bayes method, we investigate the relationship between the asymptotic behavior of variational stochastic complexity or free energy, which is the objective function to be minimized by variational Bayes, and the generalization ability of the variational Bayes approach. We show an inequality which implies a relationship between the minimum variational stochastic complexity and the generalization error of the approximate predictive distribution. This relationship is also examined by a numerical experiment.

## 1. INTRODUCTION

Bayesian estimation provides a powerful framework for learning from data. Recently, its asymptotic theory has been established, which supports its effectiveness for latent variable models such as the Gaussian mixture model (GMM) and hidden Markov model (HMM). More specifically, a formula for evaluating asymptotic forms of Bayesian mixture-type stochastic complexity or free energy was obtained and the generalization errors of statistical models have been intensively analyzed [1, 2].

Practically, however, Bayesian estimation requires some approximation method since computing the Bayesian posterior distribution is intractable in general. In this study, we focus on the approximation method, variational Bayes for Bayesian estimation. This method has been successfully applied to latent variable models such as mixture models and HMMs [3, 4]. Furthermore, its asymptotic analysis has progressed in several statistical models [5, 6]. More specifically, a formula for evaluating the asymptotic form of the minimum variational free energy was derived [6]. The variational free energy, what we call variational stochastic complexity in this paper, is the objective function to be minimized by variational Bayes and provides an upper bound for the Bayesian mixture-type stochastic complexity.

In this paper, we provide as a byproduct of this analysis, a quantity which is related to the generalization ability of the variational Bayesian approach. Analysis of generalization ability of a learning machine when it is used with the variational Bayesian approximation has been successful in quite limited cases [7]. We show an inequality which implies a relationship between the minimum variational stochastic complexity of latent variable models and the generalization error of the approximate predictive distribution. This relationship is also examined by a numerical experiment of the variational Bayesian learning of the GMM.

## 2. VARIATIONAL BAYES FOR LATENT VARIABLE MODELS

Let $\boldsymbol{y}_1^n = \{y_i\}_{i=1}^n$ be the latent (unobserved) variables corresponding to the i.i.d. observations $\boldsymbol{x}_1^n = \{x_i\}_{i=1}^n$ and consider the latent variable model with parameter $\boldsymbol{w}$,

$$p(\boldsymbol{x}_1^n|\boldsymbol{w}) = \sum_{\boldsymbol{y}_1^n} p(\boldsymbol{x}_1^n, \boldsymbol{y}_1^n|\boldsymbol{w}) = \prod_{i=1}^n \sum_{y_i} p(x_i, y_i|\boldsymbol{w}),$$

where $\sum_{\boldsymbol{y}_1^n}$ denotes the summation over all possible realizations of the latent variables.

The Bayesian posterior distribution of the latent variables and parameter $\boldsymbol{w}$ is

$$p(\boldsymbol{y}_1^n, \boldsymbol{w}|\boldsymbol{x}_1^n) = \frac{p(\boldsymbol{x}_1^n, \boldsymbol{y}_1^n|\boldsymbol{w})p_0(\boldsymbol{w})}{\sum_{\boldsymbol{y}_1^n} \int p(\boldsymbol{x}_1^n, \boldsymbol{y}_1^n|\boldsymbol{w})p_0(\boldsymbol{w})d\boldsymbol{w}},$$

where $p_0(\boldsymbol{w})$ is the prior distribution of $\boldsymbol{w}$. The posterior distribution is intractable when the marginal likelihood $Z(\boldsymbol{x}_1^n) = \sum_{\boldsymbol{y}_1^n} \int p(\boldsymbol{x}_1^n, \boldsymbol{y}_1^n|\boldsymbol{w})p_0(\boldsymbol{w})d\boldsymbol{w}$ requires the sum over exponentially many terms as in the Gaussian mixture model (GMM) and the hidden Markov model (HMM). In this article,

$$F(\boldsymbol{x}_1^n) = -\log Z(\boldsymbol{x}_1^n)$$

is referred to as the Bayesian mixture-type stochastic complexity [8].

The variational Bayesian framework approximates the Bayesian posterior distribution $p(\boldsymbol{y}_1^n, \boldsymbol{w}|\boldsymbol{x}_1^n)$ of the hidden variables and the parameters by the variational posterior distribution $q(\boldsymbol{y}_1^n, \boldsymbol{w}|\boldsymbol{x}_1^n)$, which factorizes as

$$q(\boldsymbol{y}_1^n, \boldsymbol{w}|\boldsymbol{x}_1^n) = q(\boldsymbol{y}_1^n|\boldsymbol{x}_1^n)q(\boldsymbol{w}|\boldsymbol{x}_1^n), \tag{1}$$

where $q(\boldsymbol{y}_1^n|\boldsymbol{x}_1^n)$ and $q(\boldsymbol{w}|\boldsymbol{x}_1^n)$ are probability distributions on the hidden variables and the parameters respectively. The variational posterior $q(\boldsymbol{y}_1^n, \boldsymbol{w}|\boldsymbol{x}_1^n)$ is chosen so that it minimizes the functional $\overline{F}[q]$, referred to as the variational stochastic complexity or variational free energy,

$$\overline{F}[q] = F(\boldsymbol{x}_1^n) + K(q(\boldsymbol{y}_1^n, \boldsymbol{w}|\boldsymbol{x}_1^n)||p(\boldsymbol{y}_1^n, \boldsymbol{w}|\boldsymbol{x}_1^n)), \quad (2)$$

where $K(q(\boldsymbol{y}_1^n, \boldsymbol{w}|\boldsymbol{x}_1^n)||p(\boldsymbol{y}_1^n, \boldsymbol{w}|\boldsymbol{x}_1^n))$ is the Kullback information from the variational posterior $q(\boldsymbol{y}_1^n, \boldsymbol{w}|\boldsymbol{x}_1^n)$ to the Bayesian posterior $p(\boldsymbol{y}_1^n, \boldsymbol{w}|\boldsymbol{x}_1^n)$. This reduces to the following alternating optimization of $q(\boldsymbol{y}_1^n|\boldsymbol{x}_1^n)$ and $q(\boldsymbol{w}|\boldsymbol{x}_1^n)$,

$$q(\boldsymbol{w}|\boldsymbol{x}_1^n) \propto p_0(\boldsymbol{w}) \exp \left\langle \log p(\boldsymbol{x}_1^n, \boldsymbol{y}_1^n|\boldsymbol{w}) \right\rangle_{q(\boldsymbol{y}_1^n|\boldsymbol{x}_1^n)}, \quad (3)$$

and

$$q(\boldsymbol{y}_1^n|\boldsymbol{x}_1^n) \propto \exp \left\langle \log p(\boldsymbol{x}_1^n, \boldsymbol{y}_1^n|\boldsymbol{w}) \right\rangle_{q(\boldsymbol{w}|\boldsymbol{x}_1^n)}. \quad (4)$$

where $\langle \cdot \rangle_p$ denotes the expectation with respect to $p$ [3, 4].

Let

$$\overline{F}_{\min}(\boldsymbol{x}_1^n) = \min_{q(\boldsymbol{y}_1^n|\boldsymbol{x}_1^n)q(\boldsymbol{w}|\boldsymbol{x}_1^n)} \overline{F}[q]$$

be the minimum variational stochastic complexity. We assume that $p(x|\boldsymbol{w}^*)$ with the parameter $\boldsymbol{w}^*$ is the underlying distribution generating the data $\boldsymbol{x}_1^n$ independently and identically. Because of the non-identifiability of the latent variable model, the set of true parameters

$$W^* \equiv \{\tilde{\boldsymbol{w}}| \sum_y p(x, y|\tilde{\boldsymbol{w}}) = p(x|\boldsymbol{w}^*)\},$$

is not generally a point but can be a union of several manifolds with singularities [1].

For arbitrary $\tilde{\boldsymbol{w}}^* \in W^*$,

$$\overline{F}^*(n) \equiv \langle \overline{F}_{\min}(\boldsymbol{x}_1^n) + \log p(\boldsymbol{x}_1^n|\boldsymbol{w}^*) \rangle_{p(\boldsymbol{x}_1^n|\boldsymbol{w}^*)}$$

is bounded from above by

$$U^*(n) = \langle U^*(\boldsymbol{x}_1^n) \rangle_{p(\boldsymbol{x}_1^n|\boldsymbol{w}^*)}, \quad (5)$$

where $U^*(\boldsymbol{x}_1^n)$ is given by

$$-\log \int \exp\{ \left\langle \log \frac{p(\boldsymbol{x}_1^n, \boldsymbol{y}_1^n|\boldsymbol{w})}{p(\boldsymbol{x}_1^n, \boldsymbol{y}_1^n|\tilde{\boldsymbol{w}}^*)} \right\rangle_{p(\boldsymbol{y}_1^n|\boldsymbol{x}_1^n, \tilde{\boldsymbol{w}}^*)} \} p_0(\boldsymbol{w}) d\boldsymbol{w}.$$

Asymptotic evaluation of $U^*(n)$ is elaborated in [6] with an alternative view of variational Bayes as a local variational approximation [9].

## 3. VARIATIONAL STOCHASTIC COMPLEXITY AND GENERALIZATION ERROR

Let $p(x, y|\tilde{\boldsymbol{w}}^*)$ be the true distribution of the observed variable $x$ and the latent variable $y$ which has the marginal distribution $p(x|\boldsymbol{w}^*)$. We define by

$$\overline{G}^*(\boldsymbol{x}_1^n) = K(p(x, y|\tilde{\boldsymbol{w}}^*)||\tilde{p}^*(x, y|\boldsymbol{x}_1^n)), \quad (6)$$

the generalization error of the predictive distribution,

$$\begin{aligned} \tilde{p}^*(x, y|\boldsymbol{x}_1^n) &= \langle p(x, y|\boldsymbol{w}) \rangle_{q^*(\boldsymbol{w}|\boldsymbol{x}_1^n)} \\ &= \int p(x, y|\boldsymbol{w}) q^*(\boldsymbol{w}|\boldsymbol{x}_1^n) d\boldsymbol{w}, \quad (7) \end{aligned}$$

where $q^*(\boldsymbol{w}|\boldsymbol{x}_1^n)$ is the optimal approximating posterior distribution (3) for $q(\boldsymbol{y}_1^n|\boldsymbol{x}_1^n) = p(\boldsymbol{y}_1^n|\boldsymbol{x}_1^n, \tilde{\boldsymbol{w}}^*)$. We denote its mean by

$$\overline{G}^*(n) = \left\langle \overline{G}^*(\boldsymbol{x}_1^n) \right\rangle_{\prod_{i=1}^n p(x_i|\boldsymbol{w}^*)}.$$

Then, the following inequality holds,

$$U^*(n+1) - U^*(n) \geq \overline{G}^*(n), \quad (8)$$

where $U^*(n)$ is the upper bound (5) of the minimum variational stochastic complexity.

**(Proof of the inequality (8))**

Let $p_i^*(y) = p(y|x_i, \tilde{\boldsymbol{w}}^*)$. Then it follows that

$$\begin{aligned} & U^*(\boldsymbol{x}_1^{n+1}) - U^*(\boldsymbol{x}_1^n) \\ &= -\log \frac{\int \prod_{i=1}^{n+1} \exp\{ \left\langle \log \frac{p(x_i, y|\boldsymbol{w})}{p(x_i, y|\tilde{\boldsymbol{w}}^*)} \right\rangle_{p_i^*(y)} \} p_0(\boldsymbol{w}) d\boldsymbol{w}}{\int \prod_{i=1}^{n} \exp\{ \left\langle \log \frac{p(x_i, y|\boldsymbol{w})}{p(x_i, y|\tilde{\boldsymbol{w}}^*)} \right\rangle_{p_i^*(y)} \} p_0(\boldsymbol{w}) d\boldsymbol{w}} \\ &= -\log \left\langle \exp\{ \left\langle \log \frac{p(x_{n+1}, y|\boldsymbol{w})}{p(x_{n+1}, y|\tilde{\boldsymbol{w}}^*)} \right\rangle_{p_{n+1}^*(y)} \} \right\rangle_{q^*(\boldsymbol{w}|\boldsymbol{x}_1^n)} \\ &= \langle \log p(x_{n+1}, y|\tilde{\boldsymbol{w}}^*) \rangle_{p_{n+1}^*(y)} \\ & \quad - \log \left\langle \exp\{ \langle \log p(x_{n+1}, y|\boldsymbol{w}) \rangle_{p_{n+1}^*(y)} \} \right\rangle_{q^*(\boldsymbol{w}|\boldsymbol{x}_1^n)} \quad (9) \\ &\geq \sum_y p(y|x_{n+1}, \tilde{\boldsymbol{w}}^*) \log \frac{p(x_{n+1}, y|\tilde{\boldsymbol{w}}^*)}{\langle p(x_{n+1}, y|\boldsymbol{w}) \rangle_{q^*(\boldsymbol{w}|\boldsymbol{x}_1^n)}}. \quad (10) \end{aligned}$$

In the last inequality, we have applied Jensen's inequality due to the convexity of the function $\log \int \exp(\cdot) p(\boldsymbol{w}) d\boldsymbol{w}$. Taking expectation with respect to $\prod_{i=1}^{n+1} p(x_i|\boldsymbol{w}^*)$ in both sides of the above inequality yields the inequality (8). **(Q.E.D)**

The inequality (8) is analogous to the equality,

$$F^*(n+1) - F^*(n) = G(n),$$

which holds for the average mixture-type stochastic complexity,

$$F^*(n) = \langle F(\boldsymbol{x}_1^n) + \log p(\boldsymbol{x}_1^n|\boldsymbol{w}^*) \rangle_{p(\boldsymbol{x}_1^n|\boldsymbol{w}^*)},$$

and the generalization error of the Bayesian predictive distribution,

$$G(n) = \langle K(p(x|\boldsymbol{w}^*)||p(x|\boldsymbol{x}_1^n)) \rangle_{\prod_{i=1}^n p(x_i|\boldsymbol{w}^*)},$$

where $p(x|\boldsymbol{x}_1^n) = \langle p(x|\boldsymbol{w}) \rangle_{p(\boldsymbol{w}|\boldsymbol{x}_1^n)}$.

If $U^*(n)$ has the asymptotic form $U^*(n) \simeq \overline{\lambda} \log n + O(1)$ as in eq.(14), the inequality (8) suggests that

$$\overline{G}^*(n) \leq \frac{\overline{\lambda}}{n} + o\left(\frac{1}{n}\right). \quad (11)$$

This means that the coefficient $\overline{\lambda}$ of the leading term of $U^*(n)$ is directly related to the generalization error of the variational Bayes approach measured by eq.(6).

By applying Jensen's inequality with respect to $\langle \cdot \rangle_{q^*(\boldsymbol{w}|\boldsymbol{x}_1^n)}$ and the convexity of the negative logarithmic function in eq.(9), we further obtain,

$$U^*(n+1) - U^*(n) \leq \tilde{G}^*(n),$$

where $\tilde{G}^*(n)$ is the expectation of the Gibbs generalization error,

$$\langle K(p(x,y|\tilde{\boldsymbol{w}}^*)||p(x,y|\boldsymbol{w}))\rangle_{q^*(\boldsymbol{w}|\boldsymbol{x}_1^n)}.$$

## 4. GAUSSIAN MIXTURE MODEL

Let $g(x|\mu) = \frac{1}{\sqrt{2\pi}^M}\exp\{-\frac{||x-\mu||^2}{2}\}$ be the $M$-dimensional Gaussian density and consider the GMM with $K$ components,

$$p(x|\boldsymbol{w}) = \sum_y p(x,y|\boldsymbol{w}),$$

where

$$p(x,y|\boldsymbol{w}) = \prod_{k=1}^{K}\{a_k g(x|\mu_k)\}^{y^{(k)}}. \qquad (12)$$

where $x \in R^M$ and the parameter vector $\boldsymbol{w}$ consists of the mean vectors $\{\mu_k\}_{k=1}^K$ and the mixing proportions $\boldsymbol{a} = \{a_k\}_{k=1}^K$ that satisfy $0 \le a_k \le 1$ for $k = 1, \cdots, K$ and $\sum_{k=1}^K a_k = 1$. The latent variable $y = (y^{(1)}, y^{(2)}, \cdots, y^{(K)})$ indicates the component from which the datum $x$ is generated, that is, $y^{(k)} = 1$ if $x$ is from the $k$th component and $y^{(k)} = 0$ otherwise. The variational Bayes framework is successfully applied to this model using the prior distribution,

$$p_0(\boldsymbol{w}) = p_0(\boldsymbol{a})\prod_{k=1}^{K}p_0(\mu_k), \qquad (13)$$

where

$$p_0(\boldsymbol{a}) = \frac{\Gamma(K\alpha_0)}{\Gamma(\alpha_0)^K}\prod_{k=1}^{K}a_k^{\alpha_0-1}$$

is the Dirichlet distribution with hyperparameter $\alpha_0 > 0$ and

$$p_0(\mu_k) = \sqrt{\frac{\beta_0}{2\pi}}^M \exp\{-\frac{\beta_0||\mu_k-\nu_0||^2}{2}\}$$

is the Gaussian distribution with hyperparameters $\beta_0 > 0$ and $\nu_0 \in R^M$. They are the conjugate prior distributions for the mixing proportions and each mean vector respectively.

Let the true distribution $p(x|\boldsymbol{w}^*)$ be the GMM with $K_0(\le K)$ components, that is, realizable by the model. Then it can be shown that the upper bound of the minimum variational stochastic complexity is asymptotically bounded as

$$U^*(n) \le \overline{\lambda}\log n + O(1), \qquad (14)$$

where

$$\overline{\lambda} = \begin{cases} (K-K_0)\alpha_0 + \frac{MK_0+K_0-1}{2} & (\alpha_0 \le \frac{M+1}{2}), \\ \frac{MK+K-1}{2} & (\alpha_0 > \frac{M+1}{2}). \end{cases}$$

The proof is given in [5, 6].

## 5. NUMERICAL EXPERIMENT

We implemented the variational Bayesian learning of the GMM with $K$ components (12). For simplicity, we chose the true distribution to be the standard normal distribution in $R^2$, $g(x|(0,0)^T)$. According to the choice of $\tilde{\boldsymbol{w}}^*$ for evaluating $\overline{\lambda}$ in eq.(14) [6], we consider this distribution as the choice, $\tilde{\boldsymbol{w}}^* = \{\{\tilde{a}_k^*\}, \{\tilde{\mu}_k^*\}\}_{k=1}^K$, where $\tilde{a}_1^* = 1$, $\tilde{a}_k^* = 0$ for $k = 2, \cdots, K$, $\tilde{\mu}_k^* = (0,0)^T$ for $k = 1, 2, \cdots, K$ and focus on the case where $\alpha_0 < (M+1)/2 = 1.5$.

Samples of the size $n = 100$ were generated by the true distribution. The variational Bayes algorithm was executed 21 times with 20 different random initializations and the one from the true parameter $\tilde{\boldsymbol{w}}^*$. We adopted the estimate $\hat{q}(\boldsymbol{w}|\boldsymbol{x}_1^n)$ that attained the minimum of the variational stochastic complexity and evaluated the generalization error,

$$\overline{G}(\boldsymbol{x}_1^n) = K(p(x,y|\tilde{\boldsymbol{w}}^*)||\tilde{p}(x,y|\boldsymbol{x}_1^n)), \qquad (15)$$

where $\tilde{p}(x,y|\boldsymbol{x}_1^n) = \langle p(x,y|\boldsymbol{w})\rangle_{\hat{q}(\boldsymbol{w}|\boldsymbol{x}_1^n)}$ is the (approximate) predictive distribution.

To investigate the difference between $\overline{G}(\boldsymbol{x}_1^n)$ and $\overline{G}^*(\boldsymbol{x}_1^n)$ introduced in Section 3, we also evaluated $\overline{G}^*(\boldsymbol{x}_1^n)$, on the expectation of which we can show that

$$\overline{G}^*(n) \simeq \left\{\frac{M}{2} + (K-1)\alpha_0\right\}\frac{1}{n} + o\left(\frac{1}{n}\right). \qquad (16)$$

Note that the coefficient $\frac{M}{2} + (K-1)\alpha_0$ is exactly equal to $\overline{\lambda}$ in the inequality (14) for the case where $K_0 = 1$ and $\alpha_0 < \frac{M+1}{2}$. This means that the inequality (11) is tight in this case.

Additionally, we calculated the generalization error of the marginal distribution,

$$G(\boldsymbol{x}_1^n) = K(p(x|\boldsymbol{w}^*)||\tilde{p}(x|\boldsymbol{x}_1^n)) \qquad (17)$$

where $\tilde{p}(x|\boldsymbol{x}_1^n) = \langle p(x|\boldsymbol{w})\rangle_{\hat{q}(\boldsymbol{w}|\boldsymbol{x}_1^n)}$ is the marginal predictive distribution.

Fig.1 and Fig.2 show the generalization errors for $n = 100$ and $K_0 = 1$ averaged over 100 trials with different data sets. Fig.1 is for the case of $K = 2$ with different values of the hyperparameter $\alpha_0$. We can see that for small $\alpha_0$, the behavior of the generalization error of the joint predictive distribution is well described by that of $\overline{G}^*(n)$ and hence by the coefficient $\overline{\lambda}$ in the upper bound (14). As $\alpha_0$ tends larger, the average of $\overline{G}(\boldsymbol{x}_1^n)$ also increases, as does that of the generalization error $G(\boldsymbol{x}_1^n)$ of the marginal distribution, although only slightly. This may be caused by overfitting. Fig.2 shows the average of the generalization errors for the case of $\alpha_0 = 0.2$ with different number $K$ of components. Again, we can see that for small $\alpha_0$ the generalization error of the joint predictive distribution is described by $\overline{\lambda}$ in eq.(14) while the generalization error of the marginal distribution stays constant even when the model becomes more redundant.

## 6. CONCLUSION

In this paper, we have investigated the average generalization error of the variational Bayesian approach for latent

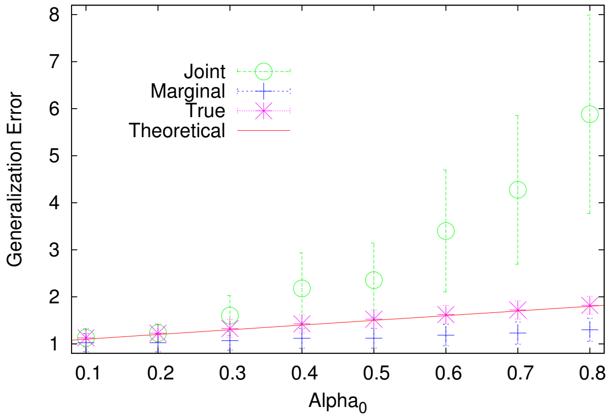

Figure 1. Average generalization errors for $K = 2$ and different $\alpha_0$ with 95%-confidence intervals. $\circ$: Average errors of the joint distribution (15). $+$: Average errors of the marginal distribution (17). $*$: Average errors of the joint distribution with the variational parameter substituted by the true one (6). Solid line: Theoretical values of the average error (16). The generalization errors are multiplied by $n = 100$ for scaling purposes.

variable models by deriving inequalities on the difference of the minimum variational stochastic complexity. We have demonstrated that the coefficient of the asymptotic minimum variational stochastic complexity partly describes the behavior of the generalization error. Thorough investigation of the generalization ability of the variational Bayes algorithm including the case for large $\alpha_0$ and for the marginal predictive distribution will be left for future work.

In the original (not approximate) Bayesian estimation, the universal relation among the quartet, Bayes and Gibbs generalization errors and Bayes and Gibbs training errors, was proved [1]. It is an important undertaking to explore such relationships among the quantities introduced in this paper for the approximate Bayesian estimation.

## 7. ACKNOWLEDGMENTS

This work was partly supported by KAKENHI (23700175) from the Japan Society for the Promotion of Science.

## 8. REFERENCES

[1] S. Watanabe, *Algebraic Geometry and Statistical Learning Theory*, Cambridge University Press, 2009.

[2] K. Yamazaki and S. Watanabe, “Singularities in mixture models and upper bounds of stochastic complexity,” *Neural Networks*, vol. 16, pp. 1029–1038, 2003.

[3] H. Attias, “Inferring parameters and structure of latent variable models by variational Bayes,” in *Uncertainty in Artificial Intelligence*, 1999, pp. 21–30.

[4] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

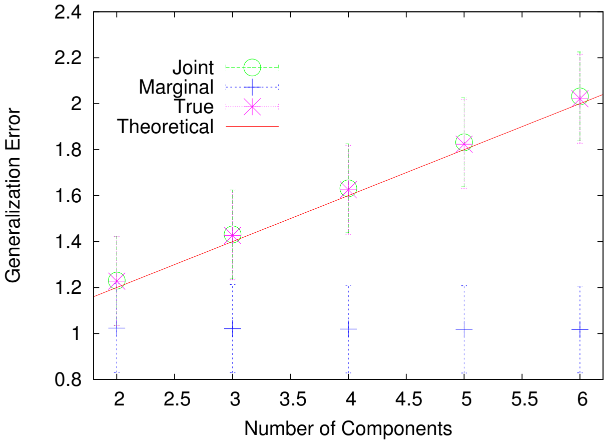

Figure 2. Average generalization errors for different $K$ with 95%-confidence intervals. Symbols are the same as in Fig.1.

[5] K. Watanabe and S. Watanabe, “Stochastic complexities of Gaussian mixtures in variational Bayesian approximation,” *Journal of Machine Learning Research*, vol. 7, pp. 625–644, 2006.

[6] K. Watanabe, “An alternative view of variational Bayes and minimum variational stochastic complexity,” in *Proc. of 3rd Workshop on Information Theoretic Methods in Science and Engineering (WITMSE-10)*. 2010, Tampere International Center for Signal Processing.

[7] S. Nakajima and S. Watanabe, “Variational Bayes solution of linear neural networks and its generalization performance,” *Neural Computation*, vol. 19, pp. 1112–1153, 2007.

[8] K. Yamanishi, *Information Theoretic Learning Theory*, Kyoritsu Shuppan, 2010, (in Japanese).

[9] K. Watanabe, M. Okada, and K. Ikeda, “Divergence measures and a general framework for local variational approximation,” *Neural Networks*, 2011, doi:10.1016/j.neunet.2011.06.004.