Free Energy of Stochastic Context Free Grammar on Variational Bayes

Tikara Hosino[1,2], Kazuho Watanabe[1], and Sumio Watanabe[3]

[1] Computational Intelligence and System Science, Tokyo Institute of Technology, Mailbox R2-5, 4259 Nagatsuda, Midori-ku, Yokohama, 226-8503 Japan
[2] Nihon Unisys, Ltd. 1-1-1 Toyosu, Koutou-ku, Tokyo 135-8560 Japan
[3] Precision and Intelligence Laboratory, Tokyo Institute of Technology, 4529 Nagatsuda, Midori-ku, Yokohama, 226-8503 Japan,
thosino@cs.pi.titech.ac.jp

Abstract. Variational Bayesian learning is proposed for approximation method of Bayesian learning. In spite of efficiency and experimental good performance, their mathematical property has not yet been clarified. In this paper we analyze variational Bayesian Stochastic Context Free Grammar which includes the true distribution thus the model is non-identifiable. We derive their asymptotic free energy. It is shown that in some prior conditions, the free energy is much smaller than identifiable models and satisfies eliminating redundant non-terminals.

1 Introduction

Stochastic Context Free Grammar (SCFG) is a statistical model used for inferring a grammar from given symbols. Whereas Hidden Markov Model (HMM) is a stochastic version of a regular grammar, SCFG is that of a context free grammar which is one step higher order in the Chomsky's hierarchy. Until now, SCFG has been mainly used in a natural language processing, now come into be used in a RNA analysis and the knowledge discovery whose problems has a secondary or nested structure that cannot be modeled in HMM. However SCFG enjoys an increasing demand, their fundamental mathematical property, such as the generalization error and the theoretically founded model selection, has not yet been clarified.

The main obstacle for analyzing SCFG is non-identifiability. In a learning machine $p(x|\theta)$ which has parameters θ, we define the machine is identifiable if the mapping of the parameter to the machine is one to one, otherwise we define the machine is non-identifiable. Many learning machines with hidden variables such as SCFG are non-identifiable. In a non-identifiable machine, Fisher information matrix is degenerated, therefore we cannot apply conventional methods which assume positive definiteness of the matrix.

In Bayesian learning, including a non-identifiable machine, asymptotic behavior of the free energy was clarified based on the algebraic analysis[11]. The result shows that in a non-identifiable machine, the free energy and the generalization error are much smaller than an identifiable machine in contrast to the

maximum likelihood method has larger generalization error than an identifiable machine. This result shows the effectiveness of a non-identifiable machine with Bayesian learning. However, in Bayesian learning, we need multiple integrals in averaging by the posterior distribution. It was well known that this operation is difficult in general. Therefore, we need to seek some approximation methods.

Recently, as an efficient approximation method, variational Bayes was proposed. Variational Bayes approximates a true posterior distribution by a trial distribution which minimizes the distance from the true posterior measured by Kullback information. In real world problems, many reports show variational learning achieves the small computational cost and the good generalization performance[1, 2, 5]. However theoretical properties of the method such as the approximation accuracy and the behavior of the free energy is not yet been clarified.

In variational Bayes, it was already analyzed that the asymptotic form of the free energy in the mixture of exponential family, linear neural networks and HMM[11, 8, 4]. In this paper, we give the analysis of variational Bayesian SCFG.

2 SCFG

In this section, we define SCFG. First, without loss of generality, we can assume the grammar was written by Chomsky's normal form. Let the model has K nonterminal symbols and M terminal symbols. The observation sequence of length L is written by $X = \{x_1, \cdots, x_L\} \in \{1, \cdots, M\}^L$. Then, the learning machine is defined by

$$p(x|\theta) = \sum_{t \in T} p_t(x|\theta), \tag{1}$$

$$\theta = \{a, b\},\ a = \{a^i_{jk}\}(1 \leq i, j, k \leq K),\ b = \{b_{im}\}(1 \leq i \leq K, 1 \leq m \leq M)$$

where T is the set of trees that generates a sequence of length L, and t is an element of the set. Moreover the parameter a^i_{jk} represents a probability that the nonterminal symbol i emits the pair of nonterminal symbols (j, k) and b_{im} represents a probability that the nonterminal symbol i emits the terminal symbol m. Those parameters $\{a, b\}$ have constraints

$$a^i_{ii} = 1 - \sum_{(j,k) \neq (i,i)} a^i_{jk},\ b_{iM} = 1 - \sum_{m=1}^{M-1} b_{im}$$

respectively.

3 Bayesian Learning and Variational Approximation

3.1 Bayesian Learning

In this section we describe Bayesian learning. Let we observe n samples $X^n = \{X_1, \ldots, X_n\}$ that was generated by the true distribution $p_0(x)$. We define the

learning machine $p(x|\theta)$ that have parameters θ and the prior distribution $\varphi(\theta)$ of the parameters.

Then, the posterior distribution of the parameters was written by

$$p(\theta|X^n) = \frac{1}{Z(X^n)} \prod_{i=1}^{n} p(X_i|\theta)\varphi(\theta),$$

where $Z(X^n)$ is a normalizing constant.

We define free energy as a negative logarithm of the normalizing constant

$$F(X^n) = -\log Z(X^n).$$

The free energy is equivalent to a logarithm of the marginal likelihood and also called a stochastic complexity. This quantities is used in model selection or estimation of hyperparameters. In Bayes learning, it was well known that a generalization error $G(n)$ is written by increase of free energy.

$$G(n) = F(n+1) - F(n).$$

Including a non-identifiable case, the asymptotic form of the free energy was obtained [10]. We define the expected free energy as

$$F(n) = \langle -\log \int \prod_{i=1}^{n} p(X_i|\theta)\varphi(\theta)d\theta \rangle_{p_0(X^n)} - nS$$

where the notation $\langle\rangle_{p(x)}$ represents an expectation over the distribution $p(x)$ and S is an entropy of the true distribution defined by

$$S = -\int p_0(x) \log p_0(x) dx,$$

whose value is independent of a learning machine.

Then, $F(n)$ has a following asymptotic form

$$F(n) = \lambda \log n - (m-1) \log \log n + O(1),$$

where the constants $-\lambda$ is a positive rational number and m is a natural number. We call λ as a learning coefficient of Bayes learning.

In an identifiable model, a learning machine that has d model parameters has $\lambda = \frac{d}{2}$. However, in non-identifiable models, a learning coefficient λ is much smaller than $\frac{d}{2}$.

In Bayesian SCFG, an upper bound of the learning coefficient was obtained[12]. We assume following conditions,

(A1) The true distribution has K_0 nonterminal symbols.

(A2) The learning machine is given by (1).

(A3) The prior distribution is strictly positive at singular points.

Then, using an algebraic geometrical method, it was shown that an upper bound of the learning coefficient satisfies inequality

$$\lambda \leq \frac{K_0(K_0^2-1)+K_0(M-1)+K_0(K^2-K_0^2)}{2}. \tag{2}$$

Note, this quantity is much smaller than a number of parameters

$$\frac{K(K^2-1)+K(M-1)}{2}.$$

3.2 Variational Bayes [1]

Next, we describe variational Bayesian learning. We define a complete sample as a pair of the observation sample $X^n = \{X_1, \ldots, X_n\}$ and the corresponding hidden variables $Y^n = \{Y_1, \ldots, Y_n\}$. Then Bayes free energy is written by

$$\begin{aligned} F(X^n) &= -\log \int \sum_{Y^n} \prod_{i=1}^{n} p(X_i, Y_i|\theta)\varphi(\theta)d\theta \\ &= -\log \int \sum_{Y^n} p(X^n, Y^n|\theta)\varphi(\theta)d\theta \end{aligned}$$

where the sum $\sum_{Y^n}$ takes all pair of the hidden variables.

Variational Bayes learning approximates Bayes free energy by the arbitrary trial distribution $q(Y^n, \theta)$. By using Jensen's inequality, the free energy is upper bounded as

$$\begin{aligned} F(X^n) &= -\log \int \sum_{Y^n} q(Y^n, \theta) \frac{p(X^n, Y^n|\theta)\varphi(\theta)}{q(Y^n, \theta)} d\theta \\ &\leq -\int \sum_{Y^n} q(Y^n, \theta) \log \frac{p(X^n, Y^n|\theta)\varphi(\theta)}{q(Y^n, \theta)} d\theta \equiv F_v(X^n). \end{aligned}$$

Rewrite this equation by using $p(X^n, Y^n, \theta) = p(Y^n, \theta|X^n)p(X^n)$,

$$F_v(X^n) = F(X^n) + \int \sum_{Y^n} q(Y^n, \theta) \log \frac{q(Y^n, \theta)}{p(Y^n, \theta|X^n)} d\theta.$$

From this equation, we can see that a minimization of $F_v(X^n)$ is equivalent to a minimization of Kullback information from the trial posterior $q(Y^n, \theta)$ to the true posterior $p(Y^n, \theta|X^n)$ and equality holds if and only if the true posterior coincides with the trial posterior. Additionally, we can evaluate an approximation accuracy by the difference of Bayes free energy and variational free energy.

In variational Bayes, for computational efficiency, we constrain the trial distribution $q(Y^n, \theta)$ to a independent form $q(Y^n)r(\theta)$ between the hidden variables

and the model parameters. We call the constrained free energy as variational free energy $\overline{F}_v(X^n)$ which is given by

$$F(X^n) \leq -\int \sum_{Y^n} q(Y^n) r(\theta) \log \frac{p(X^n, Y^n|\theta)\varphi(\theta)}{q(Y^n) r(\theta)} d\theta$$
$$= -\int \sum_{Y^n} q(Y^n) r(\theta) \log \frac{p(X^n, Y^n|\theta)}{q(Y^n)} d\theta + \int r(\theta) \log \frac{r(\theta)}{\varphi(\theta)} d\theta \equiv \overline{F}_v(X^n).$$

$\overline{F}_v(X^n)$ is a functional over the arbitrary distribution $q(Y^n)$ and $r(\theta)$. When we select the learning machine $p(X^n, Y^n)$ from exponential family with hidden variables and choose the corresponding conjugate prior, we can derive a very efficient EM like algorithm. In SCFG, it was the efficient algorithm known as Inside-Outside algorithm [7] and the extension to variational Bayes was also proposed [5].

4 Main Result

Let $\overline{F}_v(n)$ as expectation over the sample set. In this paper, we clarify an asymptotic form of $\overline{F}_v(n)$ as number of samples goes to infinity. We define the expected variational free energy as

$$\overline{F}_v(n) = \langle \overline{F}(X^n) \rangle_{p_0(X^n)} - nS$$
$$= \langle \min_{q,r} \{ -\langle \log \frac{p(X^n, Y^n|\theta)}{q(Y^n)} \rangle_{q(Y^n) r(\theta)} + \log p_0(X^n)$$
$$+ \int r(\theta) \log \frac{r(\theta)}{\varphi(\theta)} d\theta \} \rangle_{p_0(X^n)} \tag{3}$$

and assume the following (A1), (A2), (A3), (A4) conditions.

(A1) True distribution $p_0(x)$ has K_0 nonterminal symbols and M terminal symbols and written with constants θ^* as

$$p_0(x|\theta^*) = \sum_{t \in T} p_{0t}(x|\theta^*)$$
$$\theta^* = \{a^*, b^*\}$$
$$a^* = \{a^{*i}_{jk}\} (1 \leq i, j, k \leq K_0, (j,k) \neq (i,i))$$
$$b^* = \{b^*_{im}\} (1 \leq i \leq K, 1 \leq m \leq M-1)$$
$$a^{*i}_{ii} = 1 - \sum_{(j,k) \neq (i,i)} a^i_{jk}, b^*_{iM} = 1 - \sum_{m=1}^{M-1} b_{im}.$$

However SCFG has a non-trivial non-identifiability as HMM [6, 3], we assume K_0 is the smallest number of non-terminal symbols on this parametrization.

(A2) The learning machine is given by (1) and includes the true distribution, namely, the number of nonterminals K satisfies the inequality $K_0 \leq K$.

(A3) The prior distribution of parameters $a = \{a^i_{jk}\}$ and $b = \{b_{im}\}$ are Dirichlet distribution with hyperparameter ϕ_0 and ξ_0.

$$\varphi(a) = \prod_{i=1}^{K} \frac{\Gamma(K^2\phi_0)}{\Gamma(\phi_0)^{K^2}} \prod_{j=1,k=1}^{K} (a^i_{jk})^{\phi_0 - 1},$$
$$\varphi(b) = \prod_{i=1}^{K} \frac{\Gamma(M\xi_0)}{\Gamma(\xi_0)^{M}} \prod_{m=1}^{M} b_{im}^{\xi_0 - 1}.$$

(A4) Variational Bayes estimator is consistent.

Theorem 1. *Under the condition (A1) to (A4), expectation of variational free energy satisfies as number of samples goes to infinity*

$$\overline{F}_v(n) = \overline{\lambda} \log n + O(1) \quad (n \to \infty).$$

Where the learning coefficient $\overline{\lambda}$ *is given by*

$$\overline{\lambda} = \begin{cases} \frac{K_0(K_0^2-1)+K_0(M-1)}{2} + K_0(K^2 - K_0^2)\phi_0 \\ \quad (\phi_0 \leq \frac{K_0^2+KK_0+K^2+M-2}{2(K_0^2+KK_0)}), \\ \frac{K(K^2-1)+K(M-1)}{2} \\ \quad (\phi_0 > \frac{K_0^2+KK_0+K^2+M-2}{2(K_0^2+KK_0)}). \end{cases} \tag{4}$$

Proof. In SCFG, the log likelihood of a sequence of the complete sample $\{X, Y\}$ is given by

$$\log p(X, Y|\theta) = \sum_{t=2}^{L} \sum_{i,j,k=1}^{K} y^i_{t,jk} \log a^i_{jk} + \sum_{t=1}^{L} \sum_{i=1}^{K} \sum_{m=1}^{M} \hat{y}_{t,i} x_{t,m} \log b_{im}$$

where the state i of $\hat{y}_{t,i}$ indicates the tth leaf from the left of the tree that has L leafs. Moreover, we define expected sufficient statistics as

$$n^i_{jk} = \sum_{t=2}^{L} \langle y^i_{t,jk} \rangle_{q(Y)}, \ n_i = \sum_{j,k=1}^{K} n^i_{jk}$$
$$n_{im} = \sum_{t=1}^{L} \langle \hat{y}_{t,i} \rangle_{q(Y)} x_{t,m}, \ \hat{n}_i = \sum_{m=1}^{M} n_{im}.$$

Then, the posterior distributions are given by

$$r(a) = \prod_{i=1}^{K} \frac{\Gamma(n_i + K^2\phi_0)}{\prod_{j,k=1}^{K} \Gamma(n^i_{jk} + \phi_0)} \prod_{j,k=1}^{K} (a^i_{jk})^{n^i_{jk} + \phi_0 - 1},$$
$$r(b) = \prod_{i=1}^{K} \frac{\Gamma(\hat{n_i} + M\xi_0)}{\prod_{m=1}^{M} \Gamma(n_{im} + \xi_0)} \prod_{m=1}^{M} b_{im}^{n_{im} + \xi_0 - 1}.$$

First, we consider a lower bound. About the first term, using Jensen's inequality, we can evaluate as follows

$$
\begin{aligned}
&-\int \sum_{Y^n} q(Y^n) r(\theta) \log \frac{P(X^n, Y^n|\theta)}{q(Y^n)} d\theta \\
&\geq \ -\log \langle p(X^n|\theta) \rangle_{r(\theta)} \ \geq \ -\log p(X^n|\theta^*).
\end{aligned}
$$

Where θ^* is the maximum likelihood parameter and using the fact that the predictive distribution $\langle p(X^n|\theta)\rangle_{r(\theta)}$ has a same form as the learning machine $p(X^n|\theta^*)$. In finite dimensional models, as samples goes to infinity, it was known that the log likelihood ratio converged to $O(1)$ under weak conditions. SCFG that has a finite number of terminals and finite length of a sequence satisfies these assumptions. Therefore, the first term of θ^* satisfies inequality

$$
\begin{aligned}
&-\langle \log \frac{p(X^n, Y^n|\theta)}{q(Y^n)} \rangle_{q(Y^n)r(\theta)} + \log p_0(X^n) \qquad (5) \\
&\geq \log \frac{p_0(X^n)}{p(X^n|\theta^*)} \to O_p(1) \quad (n \to \infty).
\end{aligned}
$$

where the right hand side is the log likelihood ratio.

Next, we consider a second term of $\overline{F}_v(n)$.

$$
\int r(\theta) \log \frac{r(\theta)}{\varphi(\theta)} d\theta = K(r(\theta)||\varphi(\theta))
$$

Above equation is Kullback information from the posterior to the prior. Specifically, it is written by expected sufficient statistics as

$$
\begin{aligned}
K(r(\theta)||\varphi(\theta)) = &\sum_{i=1}^{K} \Big\{ \log \Gamma(n_i + K^2\phi_0) - n_i \Psi(n_i + K^2\phi_0) \\
&- \sum_{j,k=1}^{K} \{ \log \Gamma(n^i_{jk} + \phi_0) - n^i_{jk} \Psi(n^i_{jk} + \phi_0) \} \Big\} \\
&+ \sum_{i=1}^{K} \Big\{ \log \Gamma(\hat{n}_i + M\xi_0) - \hat{n}_i \Psi(\hat{n}_i + M\xi_0) \\
&- \sum_{m=1}^{M} \{ \log \Gamma(n_{im} + \xi_0) - n_{im} \Psi(n_{im} + \xi_0) \} \Big\} \\
&+ const, \qquad (6)
\end{aligned}
$$

An asymptotic form of the equation is written with asymptotic forms of Psi function $\Psi(x)$ and log Gamma function $\log \Gamma(x)$

$$
\begin{aligned}
&\Psi(x) = \log x + O(1), \\
&\log \Gamma(x) = (x - \frac{1}{2}) \log x - x + O(1) \quad (x \to \infty),
\end{aligned}
$$

$$
\begin{aligned}
&K(r(\theta)||\varphi(\theta)) \\
&= \sum_{i=1}^{K}\Big\{(K^2\phi_0 - \frac{1}{2})\log(n_i + K^2\phi_0) - \sum_{j,k=1}^{K}\{(\phi_0 - \frac{1}{2})\log(n_{jk}^i + \phi_0)\}\Big\} \\
&+ \sum_{i=1}^{K}\Big\{(M\xi_0 - \frac{1}{2})\log(\hat{n_i} + M\xi_0) - \sum_{m=1}^{M}\{(\xi_0 - \frac{1}{2})\log(n_{im} + \xi_0)\}\Big\} + O(1).
\end{aligned} \tag{7}
$$

To evaluate the coefficient of the leading term, let each n_i has order α_ith power of n and n_{jk}^i has order β_{jk}^ith power of n.

$$
\begin{aligned}
n_i &= p_i n^{\alpha_i} + o(n^{\alpha_i}),\ (0 < p_i, 0 \le \alpha_i \le 1, i = 1, \ldots, K), \\
n_{jk}^i &= p_{jk}^i n^{\beta_{jk}^i} + o(n^{\beta_{jk}^i}),\ (0 < p_{jk}^i, 0 \le \beta_{jk}^i \le 1, j, k = 1, \ldots, K).
\end{aligned}
$$

Then, using the assumption (A2),

$$
\begin{aligned}
&\log(n_i + K^2\phi_0) = \log(\hat{n_i} + M\xi_0) = \alpha_i \log n + o(\log n), \\
&\log(n_{jk}^i + \phi_0) = \beta_{jk}^i \log n + o(\log n), \\
&\log(n_{im} + \xi_0) = \alpha_i \log n + o(\log n)
\end{aligned} \tag{8}
$$

is holds. Substitute (8) to (7), we get

$$
\begin{aligned}
\frac{K(r(\theta)||\varphi(\theta))}{\log n} &= \sum_{i=1}^{K}\{(K^2\phi_0 - \frac{1}{2})\alpha_i - \sum_{j,k=1,\alpha_i\neq 0}^{K}(\phi_0 - \frac{1}{2})\beta_{jk}^i\} \\
&+ \sum_{i=1}^{K}\{(\frac{M}{2} - \frac{1}{2})\alpha_i\} + o(1)\ \ (n \to \infty).
\end{aligned} \tag{9}
$$

where the inner sum is taken only $\alpha_i \neq 0$ by the constraint of $n_i = \sum_{j,k=1}^{K} n_{jk}^i$.

To consider minimization of (9), using (A4), we divide the sum into the true K_0 terms and the redundant $K - K_0$ terms. We can see that in case $\phi_0 < \frac{1}{2}$, minimization is achieved by eliminating redundant non-terminals, and in case $\phi_0 \ge \frac{1}{2}$, minimization is achieved by α_i having same order of β_{jk}^i. Therefore, assume the learning machine uses redundant $l(0 \le l \le K - K_0)$ non-terminals, we get

$$
\begin{aligned}
\frac{K(r(\theta)||\varphi(\theta))}{\log n} &= \sum_{i=1}^{K_0}\{(K^2\phi_0 + \frac{M}{2} - 1)\alpha_i + \sum_{j,k=1}^{K_0}(\frac{1}{2} - \phi_0)\beta_{jk}^i\} \\
&+ \sum_{i=K_0+1}^{K}\{(K^2\phi_0 + \frac{M}{2} - 1)\alpha_i - \sum_{\substack{j,k=K_0+1 \\ \alpha_i\neq 0}}^{K}(\phi_0 - \frac{1}{2})\beta_{jk}^i\} + o(1) \\
&= \sum_{i=1}^{K_0}\{(K^2\phi_0 + \frac{M}{2} - 1)\alpha_i - \sum_{j,k=1}^{K_0}(\phi_0 - \frac{1}{2})\beta_{jk}^i\} + g(l) + o(1),
\end{aligned}
$$

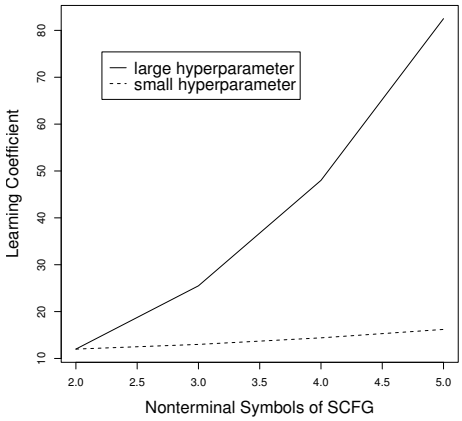

Fig. 1. Schematic diagram of variational free energy.
In case of the true distribution has 2 non-terminals and 10 terminals. The horizontal axis is number of non-terminals of learning machine and the vertical axis is the learning coefficient $\bar{\lambda}$. The solid line is large ϕ_0 and coincides with BIC. The dotted line is $\phi_0 = 0.1$.

where $g(l)$ is given by

$$g(l) = (K^2\phi_0 + \frac{M}{2} - 1)l + (\frac{1}{2} - \phi_0)((k_0 + l)^3 - k_0^3).$$

Then, the l that minimize $g(l)$ is given by

$$\begin{cases} l = 0 & (\phi_0 \leq \frac{K0^2 + KK_0 + K^2 + M - 2}{2(K_0^2 + KK_0)}) \\ l = K - K_0 & (\phi_0 > \frac{K0^2 + KK_0 + K^2 + M - 2}{2(K_0^2 + KK_0)}). \end{cases}$$

In case of a upper bound, these conditions satisfies including the true distribution, then we can set n_i and n_{jk}^i appropriately so that $\overline{F}_v(n) \leq \bar{\lambda} \log n + O(1)$ as n tends to infinity. Finally, substitute $(K_0 + l)$ to (9) completes the proof. □

5 Discussion

First, the asymptotic variational free energy is divided into two case by the hyperparameter of non-terminals ϕ_0. In case of large ϕ_0, the learning coefficient $\bar{\lambda}$ is coincide with half of the model parameters and equivalent to BIC [9]. In case of small ϕ_0, the minimum of the free energy satisfies eliminates redundant non-terminals and much smaller than model parameters (figure1). In the case of eliminating the redundant non-terminals, Variational Bayes learning is achieved with small number of estimated parameters than maximum likelihood method and has the effect of avoiding the overfitting.

Second, to evaluate approximation accuracy, we compare variational free energy to Bayes free energy. Compare the case of the model has uniform prior distribution ($\phi_0 = 1.0$). Then, we can see that variational free energy is always greater than Bayes upper bound. Hence, variational Bayes posterior distribution does not coincide with Bayes one even asymptotically.

6 Conclusion

In variational Bayesian SCFG, we evaluate the asymptotic form of variational free energy which is objective function of the method. In some prior conditions, variational free energy is much smaller than identifiable models and satisfies eliminating redundant non-terminals. Moreover, this result also gives a criterion for learning algorithms of variational Bayes estimators.

References

1. H. Attias, "Inferring parameters and structure of latent variable models by variational Bayes", in *Proc. 15th Conference on Uncertainty in Artificial Intelligence,* 1999, pp. 21-20.
2. M. J. Beal, *Variational Algorithms for Approximate Bayesian Inference,* PhD thesis, University College London, 2003.
3. E. Gassiat and S. Boucheron. "Optimal error exponents in hidden Markov models order estimation.", *IEEE Transactions on Information Theory,* vol. 49, no.2, pp. 964-980, 2003.
4. T. Hosino, K. Watanabe, and S. Watanabe. "Stochastic Complexity of Variational Bayesian Hidden Markov Models", *International Joint Conference on Neural Networks,* 2005.
5. K. Kurihara and T. Sato, "An Application of the Variational Bayesian Approach to Probabilistic Context-Free Grammars.", *International Joint Conference on Natural Language Processing,* 2004.
6. H. Ito, S.-I. Amari, and K. Kobayashi. "Identifiability of hidden Markov information sources and their minimum degrees of freedom", *IEEE Transactions on Information Theory,* vol. 38, no.2, pp. 324–333, 1992.
7. K. Lari and S. Young, "The estimation of stochastic context-free grammars using the inside-outside algorithm", *Computer Speech and Language,* vol. 4, pp. 33-56, 1990.
8. S. Nakajima and S. Watanabe, "Generalization Error and Free Energy of Linear Neural Networks in Variational Bayes Approach", *The 12th International Conference on Neural Information Processing,* 2005.
9. G. Schwarz, "Estimating the dimension of a model", *Annals of Statistics,* vol. 6, no. 2, pp. 461-464, 1978.
10. S. Watanabe, "Algebraic analysis for non-identifiable learning machines," *Neural Computation,* vol. 13, no. 4, pp. 899-933, 2001.
11. K.Watanabe, S.Watanabe, "Variational bayesian stochastic complexity of mixture models," *Advances in Neural Information Processing Systems 18*, MIT Press, 2006, to appear.
12. K. Yamazaki and S. Watanabe, "Generalization Errors in Estimating of Stochastic Context-Free Grammar," *Artificial Intelligence and Soft Computing,* 2005.