RC24670 (W0810-092) October 15, 2008
Computer Science

IBM Research Report

A New Model Selection Criterion for Monte Carlo Sampling Algorithms

Avishy Carmi
Cambridge University
Cambridge, UK

Dimitri Kanevsky
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598
USA

Research Division
Almaden - Austin - Beijing - Cambridge - Haifa - India - T. J. Watson - Tokyo - Zurich

Abstract

We present a new model selection criterion that can be easily approximated by a Monte Carlo sampling algorithm. It is shown that under certain conditions the new criterion is related to both the deviance and the Akaike information criteria. The new criterion can be easily extended to an arbitrary non-negative objective function using the extended Baum-Welch procedure.

Acknowledgment: The authors wish to thank Dan Povey for the helpful discussions which allowed formulating the new criterion.

Model Inference Criterion

Let θ_n be an $\mathbb{R}^{nl}$-valued random parameter vector where l is some integer (e.g., θ_n is a vector constructed out of n $\mathbb{R}^l$-valued random parameters, $\theta_n = \left[\vartheta_1^T, \dots, \vartheta_n^T\right]^T$, $\vartheta_j \in \mathbb{R}^l$). Let also $\pounds(\theta_n \mid z)$ be some likelihood function of θ_n given the data set $z = \{z_i\}_{i=1}^N$. Assuming that the conditional expectation $E[\theta_n \mid z]$ is akin to the maximum likelihood (ML) estimate of θ_n (e.g., $\pounds(\theta_n \mid z)$ is a Gaussian likelihood) we propose the following criteria for finding n, the model dimension

$$\min_n -2\log\pounds\left(E[\theta_n \mid z] \mid z\right) + 2E_{\theta_n|z}\left[\left(\log\pounds(\theta_n \mid z) - E_{\theta_n|z}\left[\log\pounds(\theta_n \mid z)\right]\right)^2\right] \tag{1}$$

the second term above which penalizes the ML estimate is the variance of the log likelihood function computed for a given n, that is

$$E_{\theta_n|z}\left[\left(\log\pounds(\theta_n \mid z) - E_{\theta_n|z}\left[\log\pounds(\theta_n \mid z)\right]\right)^2\right] = \mathrm{Var}\left[\log\pounds(\theta_n \mid z)\right] \tag{2}$$

The expectations in (1) are evaluated with respect to the conditional probability density function (pdf)

$$p(\theta_n \mid z) \propto \pounds(\theta_n \mid z)p(\theta_n) \tag{3}$$

and can be numerically approximated by means of a Monte Carlo sampling algorithm (e.g., Markov chain Monte Carlo, particle filtering) [1, 2].

Relation to Other Criteria

Providing that the conditions given below are satisfied we show that the new criterion in (1) coincides up to the second order terms with both the deviance information criterion (DIC) [3] and the Akaike information criterion (AIC) [4].

Proposition 1. *If the following is satisfied*

$$\frac{1}{\pounds}\frac{\partial^2\pounds}{\partial\theta_n^2}|_{\theta_n=E[\theta_n|z]} = 0 \tag{4}$$

then the new criterion (NC) (1) *coincides up to the 2nd-order terms with the DIC*

$$DIC := -4E_{\theta_n|z}\left[\log\pounds(\theta_n \mid z)\right] + 2\log\pounds(E[\theta_n \mid z] \mid z) \tag{5}$$

Proof. Let us write the 2nd-order Taylor expansion of the log likelihood around $E[\theta_n \mid z]$

$$\begin{aligned}\log\pounds(\theta_n \mid z) \approx \log\pounds(E[\theta_n \mid z] \mid z) + \frac{\partial\log\pounds}{\partial\theta_n}|_{\theta_n=E[\theta_n|z]}\left(\theta_n - E[\theta_n \mid z]\right) \\ + \frac{1}{2}\left(\theta_n - E[\theta_n \mid z]\right)^T A\left(\theta_n - E[\theta_n \mid z]\right)\end{aligned} \tag{6}$$

where $A := \frac{\partial^2\log\pounds}{\partial\theta_n^2}|_{\theta_n=E[\theta_n|z]}$. Taking the expectation of (6) yields

$$E_{\theta_n|z}[\log\pounds(\theta_n \mid z)] \approx \log\pounds(E[\theta_n \mid z] \mid z) + \frac{1}{2}\mathrm{tr}\left[\mathrm{Cov}[\theta_n \mid z]\frac{\partial^2\log\pounds}{\partial\theta_n^2}|_{\theta_n=E[\theta_n|z]}\right] \tag{7}$$

Substituting (7) into (5) gives

$$DIC \approx -2\log\pounds(E[\theta_n \mid z] \mid z) - 2\mathrm{tr}\left[\mathrm{Cov}[\theta_n \mid z]\frac{\partial^2\log\pounds}{\partial\theta_n^2}|_{\theta_n=E[\theta_n|z]}\right] \tag{8}$$

Now, substituting (6) and (7) into (1) while neglecting the 3rd and higher-order terms yields

$$NC \approx -2\log\pounds(E[\theta_n \mid z] \mid z) + 2\mathrm{tr}\left[\mathrm{Cov}[\theta_n \mid z]\left(\frac{\partial\log\pounds}{\partial\theta_n}|_{\theta_n=E[\theta_n|z]}\right)\left(\frac{\partial\log\pounds}{\partial\theta_n}|_{\theta_n=E[\theta_n|z]}\right)^T\right] \tag{9}$$

The proposition follows straightforwardly from (4), (8) and (9) upon recognizing that

$$\frac{\partial^2\log\pounds}{\partial\theta_n^2} = \frac{1}{\pounds}\frac{\partial^2\pounds}{\partial\theta_n^2} - \left(\frac{\partial\log\pounds}{\partial\theta_n}\right)\left(\frac{\partial\log\pounds}{\partial\theta_n}\right)^T \tag{10}$$

□

Corollary 1. *In the case where $\pounds(\theta_n \mid z)$ is Gaussian and the condition* (4) *is satisfied the criterion* (1) *coincides with the AIC*

$$-2 \log \pounds(E[\theta_n \mid z] \mid z) + 2nl \tag{11}$$

Proof. For Gaussian likelihood

$$\mathrm{Cov}[\theta_n \mid z] = R, \quad R \in \mathbb{R}^{nl \times nl} \tag{12}$$

and

$$\frac{\partial^2 \log \pounds}{\partial \theta_n^2} = -R^{-1} \tag{13}$$

Substituting the above into (8) yields

$$NC \approx DIC \approx -2 \log \pounds(E[\theta_n \mid z] \mid z) + 2\mathrm{tr}\left[RR^{-1}\right] = -2 \log \pounds(E[\theta_n \mid z] \mid z) + 2nl \tag{14}$$

□

Gaussian Likelihoods

It was already noted that the condition (4) is equivalent to (see (10))

$$\frac{\partial^2 \log \pounds}{\partial \theta_n^2} = -\left(\frac{\partial \log \pounds}{\partial \theta_n}\right)\left(\frac{\partial \log \pounds}{\partial \theta_n}\right)^T \tag{15}$$

The above condition may hold in the Gaussian scalar case, that is for

$$\pounds(\mu \mid z) = c \exp\left\{-\frac{1}{2}\frac{(z-\mu)^2}{\sigma^2}\right\} \tag{16}$$

Thus,

$$\frac{\partial \log \pounds}{\partial \mu} = \frac{(z-\mu)}{\sigma^2} \tag{17a}$$

$$-\frac{\partial^2 \log \pounds}{\partial \mu^2} = \frac{1}{\sigma^2} \tag{17b}$$

Substituting the above into (15) yields

$$\mu - \sigma = z \tag{18}$$

In other words, if the linear relationship (18) is satisfied then both Proposition 1 and Corollary 1 hold. In what follows we elaborate on the implication of (4) in the multivariate normal case.

Multivariate Normal Distributions

Consider the following likelihood pdf

$$\pounds(\mu_n \mid z) = \prod_{i=1}^{N} \frac{1}{(2\pi)^{n/2} \det(R)^{1/2}} \exp\left\{-\frac{1}{2}(z_i - \mu_n)^T R^{-1} (z_i - \mu_n)\right\} \tag{19}$$

where $z = \{z_i\}_{i=1}^N$ and $R \in \mathbb{R}^{n \times n}$ is a positive definite covariance matrix. Let us assume also that a single observation z_i is related to the model parameters μ_n by

$$z_i = \mu_n + \eta_i \tag{20}$$

Here $\{\eta_i\}_{i=1}^N$ is a zero-mean white sequence, i.e., $\mathrm{Cov}[\eta_i, \eta_j] = R\delta_{ij}$ where δ_{ij} denotes the Kronecker delta.

Proposition 2. *If the likelihood is specified by* (19), *and* (20) *holds then the condition*

$$\frac{1}{\pounds}\frac{\partial^2 \pounds}{\partial \mu_n^2}|_{\mu_n = E[\mu_n \mid z]} = 0 \tag{21}$$

is equivalent to

$$\frac{\partial \log \pounds}{\partial R}|_{\mu_n = E[\mu_n \mid z], R = R^*} = 0 \tag{22}$$

for large enough number of observations N. In other words, the proposition holds for $R = R^*$, the ML estimate of R given μ_n and z.

Proof. Explicitly writing (21) yields

$$\frac{1}{\pounds}\frac{\partial^2 \pounds}{\partial \mu_n^2} = R^{-1}\left[\sum_{i=1}^{N}(z_i - \mu_n)\right]\left[\sum_{i=1}^{N}(z_i - \mu_n)\right]^T R^{-1} - NR^{-1} = 0 \tag{23}$$

Multiplying both sides of (23) by R gives

$$R = \frac{1}{N}\left[\sum_{i=1}^{N}(z_i - \mu_n)\right]\left[\sum_{i=1}^{N}(z_i - \mu_n)\right]^T \tag{24}$$

which approaches the sample covariance for large enough N (see (20)), that is

$$R \approx \frac{1}{N}\sum_{i=1}^{N}(z_i - \mu_n)(z_i - \mu_n)^T \tag{25}$$

Finally, it can be easily shown (see appendix) that the sample covariance in (25) is the ML estimate of R satisfying (22) which thereby completes the proof. □

Modification of the New Criterion

Letting

$$\theta_n = \{\theta_n^{(1)}, \theta_n^{(2)}\} \tag{26}$$

while assuming that the following is satisfied

$$\left\{\frac{\partial \log \pounds}{\partial \theta_n^{(2)}}\Big|_{\theta^{(1)}=E[\theta_n^{(1)}|z],\theta^{(2)}={\theta_n^{(2)}}^*} = 0\right\} \Longrightarrow \left\{\frac{1}{\pounds}\frac{\partial^2 \pounds}{\partial {\theta_n^{(1)}}^2}\Big|_{\theta^{(1)}=E[\theta_n^{(1)}|z],\theta^{(2)}={\theta_n^{(2)}}^*} = 0\right\} \tag{27}$$

allows writing the new criterion as follows

$$\min_n - \log \pounds(E_{\theta_n^{(1)}|z}[\theta_n \mid z] \mid z) + E_{\theta_n^{(1)}|z}\left[\left(\log \pounds(\theta_n \mid z) - E_{\theta_n^{(1)}|z}[\log \pounds(\theta_n \mid z)]\right)^2\right] \tag{28}$$

$$\text{s.t.} \quad \frac{\partial^2 \pounds}{\partial {\theta_n^{(1)}}^2}\Big|_{\theta^{(1)}=E[\theta_n^{(1)}|z],\theta^{(2)}={\theta_n^{(2)}}^*} = 0$$

or

$$\min_n - \log \pounds(E_{\theta_n^{(1)}|z}[\theta_n \mid z] \mid z) + E_{\theta_n^{(1)}|z}\left[\left(\log \pounds(\theta_n \mid z) - E_{\theta_n^{(1)}|z}[\log \pounds(\theta_n \mid z)]\right)^2\right] \tag{29}$$

$$\text{s.t.} \quad \frac{\partial \log \pounds}{\partial \theta_n^{(2)}}\Big|_{\theta^{(1)}=E[\theta_n^{(1)}|z],\theta^{(2)}={\theta_n^{(2)}}^*} = 0$$

or

$$\min_n - \log \pounds(E_{\theta_n^{(1)}|z}[\theta_n \mid z] \mid z) + E_{\theta_n^{(1)}|z}\left[\left(\log \pounds(\theta_n \mid z) - E_{\theta_n^{(1)}|z}[\log \pounds(\theta_n \mid z)]\right)^2\right] \tag{30}$$

$$\text{s.t.} \quad {\theta_n^{(2)}}^* = \arg\max_{\theta^{(2)}} \log \pounds$$

Generalization to an Arbitrary Objective Function

Let $F(\theta_n, z)$ be an arbitrary non-negatve objective function. Then

$$\min_n - \log F(E_{\theta_n^{(1)}|z}[\theta_n \mid z], z) + E_{\theta_n^{(1)}|z}\left[\left(\log F(\theta_n, z) - E_{\theta_n^{(1)}|z}[\log F(\theta_n, z)]\right)^2\right] \tag{31a}$$

$$\text{s.t.} \quad {\theta_n^{(2)}}^* = \arg\max_{\theta^{(2)}} \log F \tag{31b}$$

where (31b) can be solved via extended Baum-Welch [5–7].

Appendix A

ML Estimate of the Covariance Matrix R

Let us write down the log likelihood function

$$\log \pounds(\mu_n \mid z) = c - \frac{1}{2}\sum_{i=1}^{N} \log\det(R) - \frac{1}{2}\sum_{i=1}^{N}(z_i-\mu_n)^T R^{-1}(z_i-\mu_n)$$
$$= c - \frac{1}{2}\left[N \log\det(R) + \operatorname{tr}\left(SR^{-1}\right)\right] \quad \text{(A.1)}$$

where $S := \sum_{i=1}^{N}(z_i-\mu_n)(z_i-\mu_n)^T$. Defining

$$B = S^{1/2}R^{-1}S^{1/2} \quad \text{(A.2)}$$

while recognizing that

$$\operatorname{tr}(SR^{-1}) = \operatorname{tr}(B), \qquad \det(R) = \det(S)\det(B)^{-1} \quad \text{(A.3)}$$

allows rewriting (A.1) as

$$\log \pounds(\mu_n \mid z) = c - \frac{N}{2}\log\det(S) - \frac{1}{2}\left[-N\log\det(B) + \operatorname{tr}(B)\right] \quad \text{(A.4)}$$

Now, the symmetric matrix B can be decomposed as $B = V\Lambda V^T$ where $\Lambda = \operatorname{diag}(\lambda_j)$ and $VV^T = I$ (i.e., V is an orthogonal matrix). Therefore (A.4) can be expressed as

$$\log \pounds(\mu_n \mid z) = c - \frac{N}{2}\log\det(S) - \frac{1}{2}\left[-N\sum_j \log\lambda_j + \sum_j \lambda_j\right] \quad \text{(A.5)}$$

Observing (A.5), the condition (22) is equivalent to

$$\frac{\partial \log \pounds}{\partial \lambda_j} = -N/\lambda_j + 1 = 0 \quad \text{(A.6)}$$

which yields the solution $\lambda_j = N$ and

$$R^* = S^{1/2}B^{-1}S^{1/2} = \frac{S}{N} = \frac{1}{N}\sum_{i=1}^{N}(z_i-\mu_n)(z_i-\mu_n)^T \quad \text{(A.7)}$$

Bibliography

[1] Doucet, A., de Freitas, J. F. G., and Gordon, N. J., *Sequential Monte Carlo Methods in Practice*, New York: Springer-Verlag, 2001.

[2] Ristic, B., Arulampalam, S., and Gordon, N., *Beyond the Kalman Filter*, Artech House, 2004.

[3] Speigelhalter, D. J., Nicola, G. B., Bradley, P. C., and Van der Linde, A., "Bayesian Measures of Models Complexity and Fit (with Discussion)," *Journal of the Royal Statistical Society*, Vol. 64, No. 4, October 2002, pp. 583–639.

[4] Akaike, H., "A New Look at the Statistical Model Identification," *IEEE Transactions on Automatic Control*, Vol. 19, No. 6, 1974, pp. 716–723.

[5] Gopalakrishnan, P., Kanevsky, D., Nahamoo, D., and Nadas, A., "An Inequality for Rational Functions with Applications to Some Statistical Estimation Problems," *IEEE Trans. Information Theory*, Vol. 37, No. 1, January 1991.

[6] Kanevsky, D., "Extended Baum Transformations For General Functions, II," Tech. Rep. RC23645(W0506-120), Human Language Technologies, IBM, 2005.

[7] Carmi, A. and Kanevsky, D., "Matrix form of Extended Baum Transformations," Tech. Rep. RC, Human Language Technologies, IBM, 2008.