The Thirteenth International Symposium on Artificial Life and Robotics 2008(AROB 13th '08),
B-Con Plaza, Beppu, Oita, Japan, January 31-February 2, 2008

On the Distribution of Posterior Probability in Bayesian Inference with a Large Number of Observations

Akira Date
Faculty of Engineering, University of Miyazaki, Miyazaki 889-2192
date@cs.miyazaki-u.ac.jp

Abstract: Probabilistic generative models work in many applications of image analysis and speech recognition. In general, there is an observation vector $\vec{y}$ and a state vector $\vec{x}$, and a joint dependency structure among them. The object of interest is, given $\vec{y}$, the most likely configuration $\vec{x}_{\rm MAP}$ and its posterior distribution. In practice, the exact value of posterior probability of $\vec{x}_{\rm MAP}$ is impossible to obtain, especially when there is a large number of observed variables. Here we analyzed the distribution of posterior probabilities of $\vec{x}_{\rm MAP}$ when there are $N = 200 \sim 1000$ observations. We used a probabilistic model with simple linear dependency structure in which the exact value of posterior probability of $\vec{x}_{\rm MAP}$ is obtainable. Computer experiments show that even an identical model generate a variety of posterior distributions, which suggest difficulties in understanding the meaning of posterior probability. Finally, we propose a method to know the confidence of the estimator $\vec{x}_{\rm MAP}$ by computing $P(\vec{x}'|\vec{y})$'s where $\vec{x}'$'s are neighbors of $\vec{x}_{\rm MAP}$.

Keywords: Bayesian inference, maximum a posteriori estimator, hidden Markov model, dependency graph

1 Introduction

In Bayesian formations, we formalize the relevant prior information, as a probability distribution, say $P(\vec{x})$. Then, given a prior distribution $P(\vec{x})$ and an "observation" $\vec{y}$, we form the posterior distribution $P(\vec{x}|\vec{y})$: the conditional distribution given what is observed. Given an observation we seek the most likely configuration $\vec{x}_{\rm MAP}$ that maximizes the posterior distribution.

Suppose we can compute $P(\vec{x}|\vec{y})$, the posterior probability of $\vec{x}_{\rm MAP}$, although it is in practice difficult to compute the exact value of posterior probability when there is a large number of observed variables. What does $P(\vec{x}|\vec{y})$ signal to us? Even if we know the posterior probability of $\vec{x}_{\rm MAP}$, its meaning is problem-dependent. When, for example, $P(\vec{x}_{\rm MAP}|\vec{y}) = 0.01$, can we say $\vec{x}_{\rm MAP}$ is not so believable? The answer depend on the problem. Even if we fix the problem setting, it is difficult to answer. It seems that larger the configuration space of $\vec{x}$, smaller the $P(\vec{x}_{\rm MAP}|\vec{y})$. However, the value of posterior probability, of course, signals something. When, for example, the posterior probability of $\vec{x}_{\rm MAP}$ is 0.98, it signals that $\vec{x}_{\rm MAP}$ is most likely interpretation with high confidence. In general, how can we use the value of posterior probability effectively for processing in next stage ?

In this paper, we used a probabilistic model with simple linear dependency structure in which the exact value of posterior probability of the most likely configuration $\vec{x}_{\rm MAP}$ is obtainable. We, then, analyzed the distribution of posterior probabilities when there are $n = 200 \sim 1000$ observations. We performed a large number of computer experiments to obtain the distribution of $P(\vec{x}_{\rm MAP}|\vec{y})$. The results suggest that there is a variety of distributions of $P(\vec{x}_{\rm MAP}|\vec{y})$. We computed not only $P(\vec{x}_{\rm MAP}|\vec{y})$ but $P(\vec{x'}_{\rm MAP}|\vec{y})$'s where $\vec{x}'$ is neighbor of $\vec{x}_{\rm MAP}$. The presentation will be nontechnical and by example, highlighting the meaning of posterior probability, such as, what posterior probability signals or whether low posterior probability signals that the estimator is relatively not believable.

2 Bayesian Inference

2.1 Linear dependency graph

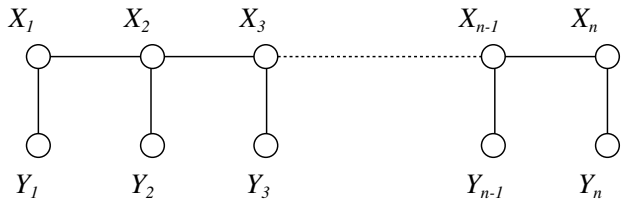
As a simple example, suppose that $X_1, X_2, \ldots$ is a first-order Markov process with state space $X \in \{0,1\}$, initial probability distribution

$$p_0 \equiv \mathrm{Prob}(X_1 = 0) = 0.5 \quad (1)$$
$$p_1 \equiv \mathrm{Prob}(X_1 = 1) = 0.5 \quad (2)$$

and transition probability matrix $P \equiv \begin{matrix} p_{00} & p_{10} \\ p_{01} & p_{11} \end{matrix} \equiv$

$$\begin{matrix} \Pr(X_{i+1}=0|X_i=0) & \Pr(X_{i+1}=0|X_i=1) \\ \Pr(X_{i+1}=1|X_i=0) & \Pr(X_{i+1}=1|X_i=1) \end{matrix} = \begin{matrix} 0.99 & 0.03 \\ 0.01 & 0.97 \end{matrix}.$$

a



b

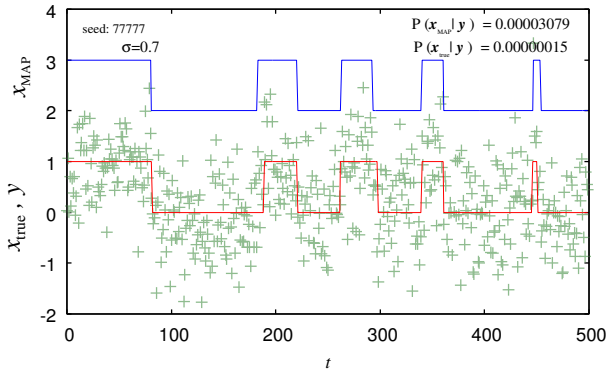


Figure 1: *a*, Linear dependency graph. The nodes of the X-graph represent a source sequence of 0 and 1's generated from a Markov chain, and the Y nodes represent a noisy observation. *b*, A source data $\vec{x}$ (represented as a line), an observation $\vec{y}$ (points), and the maximum a posteriori estimator $\vec{x}_{\mathrm{MAP}}$ (plotted $\vec{x}_{\mathrm{MAP}}+3$) and its posterior probability. $Y_i = X_i + Z_i$, $Z_i \sim \mathcal{N}(0,\sigma^2), \sigma = 0.7$, $1 \leqq i \leqq n = 500$. There was 17/500 discrepancies between $\vec{x}_{\mathrm{MAP}}$ and $\vec{x}_{\mathrm{true}}$.

Suppose we observe corrupted signal $Y_1, Y_2, \ldots Y_n$ where

$$Y_i = X_i + \eta_i \tag{3}$$

with η_i iid $\mathcal{N}(0,\sigma^2), \sigma = 0.7$;

$$\mathrm{Prob}(Y_i < y_i | x_i) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{y_i} \exp\left\{-\frac{(y-x_i)^2}{2\sigma^2}\right\} \mathrm{d}y \tag{4}$$

The goal is to estimate X where the estimation will be based upon the corrupted observations, and based upon the model, i.e., the prior distribution.

Y itself is not Markov. Nevertheless, the conditional distribution of X given Y remain simple, where X given Y is still first-order Markov. In general, the combination of a rich marginal structure for Y and a simple posterior structure for X makes hidden Markov process a common modeling tool [1].

2.2 Most likely configurations

Given an observation $\vec{y} = (y_1, \cdots, y_n)^{\mathrm{T}}$ we seek $\vec{x} = (x_1, \cdots, x_n)^{\mathrm{T}}$ that maximizes the posterior distribution (the so called MAP estimator);

$$\vec{x}_{\mathrm{MAP}} = \underset{\vec{x}}{\mathrm{argmax}}\, \mathrm{Prob}(\vec{x}|\vec{y}) \tag{5}$$

where

$$\mathrm{Prob}(\vec{x} \mid \vec{y}) \quad = \quad \frac{\mathrm{Prob}(\vec{x},\, \vec{y})}{\mathrm{Prob}(\vec{y})}. \tag{6}$$

Since $\mathrm{Prob}(\vec{y})$ is a positive constant, the goal is to compute

$$\vec{x}_{\mathrm{MAP}} = \underset{\vec{x}}{\mathrm{argmax}}\, \mathrm{Prob}(\vec{x},\, \vec{y}). \tag{7}$$

There are legendary practical problems with the actual implementation of Bayesian methods. We want to compute most likely states with respect to these posterior distributions, but usually direct evaluation is already impossible with one hundred dimensions. Fortunately, for random fields with linear graph, this posterior distribution is itself Markov, which has some striking computational implications.

$$\vec{x}_{\mathrm{MAP}} = \underset{\vec{x}}{\mathrm{argmax}} \left\{ \log p_{x_1} + \sum_{t=2}^{n} \log p_{x_{t-1}x_t} + \sum_{t=1}^{n} \log q_{x_t y_t} \right\} \tag{8}$$

In particular, dynamic programming methods can be used: computationally-feasible algorithms exist for estimating the most likely interpretation $\vec{x}$ of a given signal $\vec{y}$. In concrete, first we compute

$$C_1(i) \quad = \quad \log(p_i) + \log(q_{iy_1}) \tag{9}$$

Then, we sequentially compute for $t = 1, \cdots, n-1$

$$S_{t+1}(j) \quad = \quad \underset{i}{\mathrm{argmax}} \left\{ C_t(i) + \log p_{ij} + \log q_{jy_{t+1}} \right\} \tag{10}$$

$$C_{t+1}(j) \quad = \quad C_t(S_{t+1}(j)) + \log p_{S_{t+1}(j)j} + \log q_{jy_{t+1}} \tag{11}$$

where $i, j \in \{0,1\}$, and

$$q_{x_t y_t} \quad = \quad \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(y_t - x_t)^2}{2\sigma^2}\right\} \Delta y \tag{12}$$

Finally, we obtain

$$\hat{x}_n \quad = \quad \underset{i}{\mathrm{argmax}}\, C_n(i) \tag{13}$$

Then the $\hat{x}_t$, $t = n-1, \cdots, 1$ will be obtained.

$$\hat{x}_t \quad = \quad S_{t+1}(\hat{x}_{t+1}) \tag{14}$$

2.3 Posterior probability

We want to compute the conditional probability of the most likely state, given the observation

$$\mathrm{Prob}(\vec{x} \mid \vec{y}) \quad = \quad \frac{\mathrm{Prob}(\vec{x},\ \vec{y})}{\mathrm{Prob}(\vec{y})}. \tag{15}$$

Still it is difficult to know the posterior probability when there is a large number n of observed variables. If we compute the inverse of the conditional probability, rather than conditional probability itself, we can obtain the posterior probability for relatively large n [2].

$$\frac{1}{p(\vec{x}|\vec{y})} \quad = \quad \frac{p(\vec{y})}{p(\vec{x},\vec{y})} = \frac{\sum_{\tilde{x}} p(\tilde{x},\vec{y})}{p(\vec{x},\vec{y})} \tag{16}$$

$$= \frac{\sum_{i_1,i_2,i_3,\cdots,i_n} p_{i_1,i_2} p_{i_2,i_3} q_{i_1,y_1} q_{i_2,y_2} q_{i_3,y_3} \cdots q_{i_n,y_n}}{p_{x_1,x_2} p_{x_2,x_3} q_{x_1,y_1} q_{x_2,y_2} q_{x_3,y_3} \cdots q_{x_n,y_n}}$$

$$= \frac{\sum_{i_n} q_{i_n,y_n}}{q_{x_n,y_n}} \cdots \frac{\sum_{i_3} p_{i_3,i_4} q_{i_3,y_3}}{p_{x_3,x_4} q_{x_3,y_3}} \frac{\sum_{i_2} p_{i_2,i_3} q_{i_2,y_2}}{p_{x_2,x_3} q_{x_2,y_2}} \frac{\sum_{i_1} p_{i_1,i_2} q_{i_1,y_1}}{p_{x_1,x_2} q_{x_1,y_1}}$$

$$= \frac{\sum_{i_n} q_{i_n,y_n}}{q_{x_n,y_n}} \cdots \frac{\sum_{i_3} p_{i_3,i_4} q_{i_3,y_3}}{p_{x_3,x_4} q_{x_3,y_3}} \frac{\sum_{i_2} p_{i_2,i_3} q_{i_2,y_2} r_2(i_2)}{p_{x_2,x_3} q_{x_2,y_2}}$$

$$= \frac{\sum_{i_n} q_{i_n,y_n}}{q_{x_n,y_n}} \cdots \frac{\sum_{i_3} p_{i_3,i_4} q_{i_3,y_3} r_3(i_3)}{p_{x_3,x_4} q_{x_3,y_3}}$$

$$= \frac{\sum_{i_n} q_{i_n,y_n} r_n(i_n)}{q_{x_n,y_n}}$$

where we put

$$r_2(i_2) = \frac{\sum_{i_1} p_{i_1,i_2} q_{i_1,y_1}}{p_{x_1,x_2} q_{x_1,y_1}}$$

and for $t = 3, \cdots, n$

$$r_t(i_t) = \frac{\sum_{i_{t-1}} p_{i_{t-1},i_t} q_{i_{t-1},y_{t-1}} r_{t-1}(i_{t-1})}{p_{x_{t-1},x_t} q_{x_{t-1},y_{t-1}}}.$$

3 Computer Experiments

3.1 Posterior probability of MAP estimator

Computer simulation was carried out for $n = 200$. Typical sequences of source signal $\vec{x}_{\mathrm{TRUE}}$ and observation $\vec{y}$ and its MAP estimator $\vec{x}_{\mathrm{MAP}}$ are illustrated in Fig.2 (left). Posterior probability of the most likely state $P(\vec{x}_{\mathrm{MAP}}|\vec{y})$ and that of true state $P(\vec{x}_{\mathrm{TRUE}}|\vec{y})$ ($\vec{x}_{\mathrm{TRUE}}$ is unknown to observer) is shown on the upper-right corner for typical four cases. $P(\vec{x}_{\mathrm{MAP}}|\vec{y})$ seems to be distributed broadly, and they were 0.113, 0.164, 0.019, and 0.466 in these four cases.

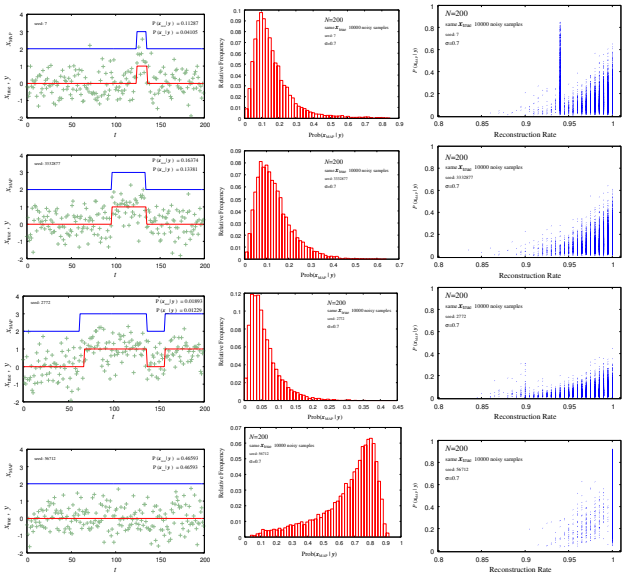


Figure 2: Posterior probability distribution of $\vec{x}_{\mathrm{MAP}}$ for typical four cases of $\vec{x}_{\mathrm{TRUE}}$. Left: See Fig.1 for explanation. Middle: Distribution of $P(\vec{x}_{\mathrm{MAP}}|\vec{y})$ for 10,000 different $\vec{y}$'s generated from an identical $\vec{x}_{\mathrm{TRUE}}$. Right: Relationship between posterior probability $P(\vec{x}_{\mathrm{MAP}}|\vec{y})$ and reconstruction rate $d(\vec{x}_{\mathrm{MAP}},\vec{x}_{\mathrm{TRUE}})/n$.

3.2 Distribution of posterior probability of $\vec{x}_{\mathrm{MAP}}$

We generated a set of 10,000 different observation $\vec{y}$'s from an identical signal $\vec{x}_{\mathrm{TRUE}}$. Figure 2 (middle) show the distribution of $P(\vec{x}_{\mathrm{MAP}}|\vec{y})$ for these 10,000 $\vec{y}$'s. The result shows that there are variety type of the distribution of $P(\vec{x}_{\mathrm{MAP}}|\vec{y})$. We asked whether these posterior probabilities signal to us in this problem setting by examining the relationship between reconstruction rate and posterior probability. The results are shown in Fig.2(right). As a reconstruction rate, we used the normalized Hamming distance

$$\frac{d_{\mathrm{ham}}(\vec{x}_{\mathrm{MAP}},\vec{x}_{\mathrm{TRUE}})}{n}$$

where $n(=200)$ is number of observations. These results show a tendency that higher the $P(\vec{x}_{\mathrm{MAP}}|\vec{y})$, larger the reconstruction rate, although it is not so simply interpreted.

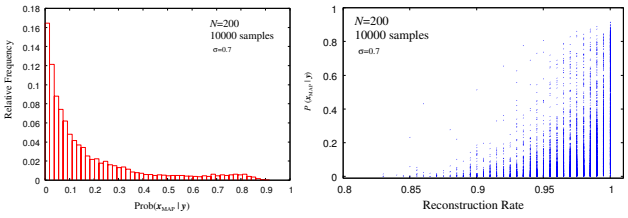


Figure 3: Posterior probability distribution of $\vec{x}_{\mathrm{MAP}}$. Left: Distribution of $P(\vec{x}_{\mathrm{MAP}}|\vec{y})$ for 10,000 different $\vec{y}$'s generated from 10,000 different samples of $\vec{x}_{\mathrm{true}}$. Right: Relationship between posterior probability $P(\vec{x}_{\mathrm{MAP}}|\vec{y})$ and reconstruction rate $d_{\mathrm{ham}}(\vec{x}_{\mathrm{MAP}},\vec{x}_{\mathrm{TRUE}})/n$.

Figure 3 shows the distribution of $P(\vec{x}_{\mathrm{MAP}}|\vec{y})$ for 10,000 totally different sample $\vec{x}_{\mathrm{TRUE}}$'s and different sample observation $\vec{y}$'s. The value of $P(\vec{x}_{\mathrm{MAP}}|\vec{y})$ decrease as n being large (compare Fig.3 (left) to Fig.4 which is the case of $n = 500$).

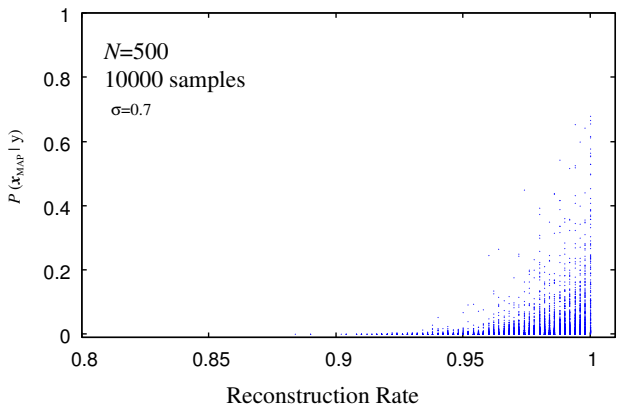


Figure 4: Reconstruction rate $d_{\mathrm{ham}}(\vec{x}_{\mathrm{MAP}},\vec{x}_{\mathrm{TRUE}})/n$ and posterior probability $P(\vec{x}_{\mathrm{MAP}}|\vec{y})$ for $n = 500$.

4 Discussion

The value of posterior probability of $\vec{x}_{\mathrm{MAP}}$ does not tell us much except in the case of $P(\vec{x}_{\mathrm{MAP}}|\vec{y})$ being extremely large since we do not know in advance the structure of posterior probability distribution. To use $P(\vec{x}_{\mathrm{MAP}}|\vec{y})$ effectively in interpretation, we consider the distribution of posterior probability of perturbed $\vec{x}_{\mathrm{MAP}}$ in which we generate a set of $\vec{x}$'s which is close to $\vec{x}_{\mathrm{MAP}}$ as a vector, and make a histogram of $P(\vec{x}|\vec{y})$. For example there are 5 transition points ($0 \to 1$ or $1 \to 0$) in $\vec{x}_{\mathrm{TRUE}}$ exemplified in Fig.5. We generated a set of $3^5 = 243$ $\vec{x}$'s in each of which each transition point of $\vec{x}_{\mathrm{TRUE}}$ was systematically shifted -1,0,or 1. We computed $P(\vec{x}|\vec{y})$ for each thus generated $\vec{x}$ (see Fig.5). From this analysis, we see the value of $P(\vec{x}_{\mathrm{MAP}}|\vec{y}) = 0.003$ has something to tell us; $\vec{x}_{\mathrm{MAP}}$ is most likely interpretation given the observation $\vec{y}$ with high confidence.

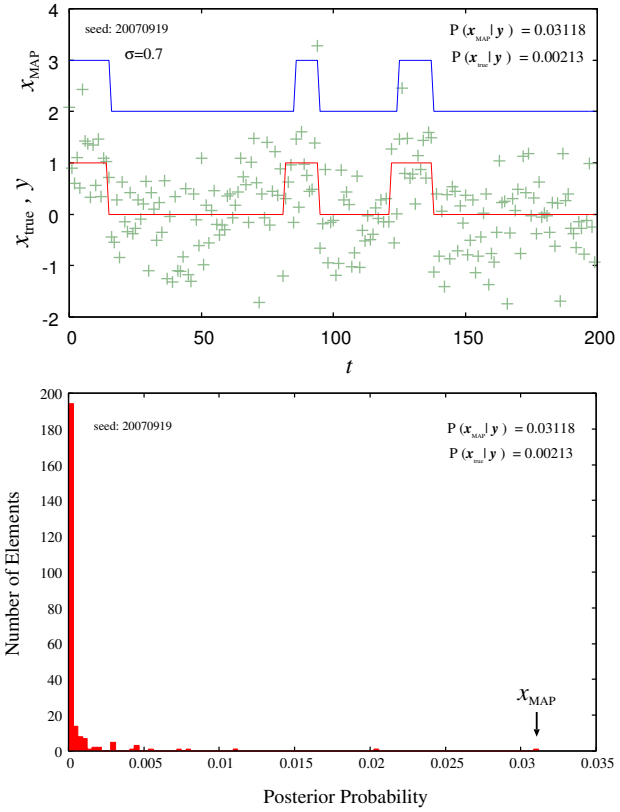


Figure 5: Meaning of posterior probability of $\vec{x}_{\mathrm{MAP}}$. Upper: $P(\vec{x}_{\mathrm{MAP}}|\vec{y}) = 0.03$. See the caption of Fig.1 for explanation in detail. Lower: Distribution of $P(\vec{x'}|\vec{y})$ where 242 $\vec{x'}$ similar to $\vec{x}_{\mathrm{MAP}}$ was generated.

Acknowledgments

This works was partly supported by Grant-in-Aid (19500257) from the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT), and by the Inamori Grant from the Inamori Foundation.

References

[1] H. Künsch, S. Geman, and A. Kehagias, "Hidden Markov random fields," *Annals of Applied Probability*, vol. 5, pp. 557–602, 1995.

[2] S. Geman and K. Kochanek, "Dynamic programming and the graphical representation of error-correcting codes," *IEEE Transactions on Information Theory*, vol. 47, pp. 549–567, 2001.