2009－CVIM－166 (7)
2009／3／13

# ベイズ学習の実装
## －MCMC/SMC/DPEM－

松本隆†

†早稲田大学理工学術院 〒169-8555 新宿区大久保 3-4-1
E-mail: †takashi@matsumoto.elec.waseda.ac.jp

あらまし ベイズ学習の実装法のいくつかを紹介する。
キーワード ベイズ学習，モンテカルロ法，先験分布，ディリクレ過程、EM

# Implementations of Bayesian Learning
## －MCMC/SMC/DPEM－

Takashi MATSUMOTO†

† Faculty of Science and Engineering, Waseda University 3-4-6 Okubo Shinjuku-ku, Tokyo, 169-8555 Japan
E-mail: †takashi@matsumoto.elec.waseda.ac.jp

**Abstract** Several implementation schemes are reviewed for Bayesian learning.
**Keyword** Bayesian Learning, Monte Carlo Schemes, Prior Distributions, Dirichlet Processes, EM

### 1. Introduction

Several implementation schemes are reviewed for Bayesian learning:

a. Sequential Marginal Likelihood Online Change Detector
b. Natural Sequential Prior for Online Bayesian Learning
c. Monte Carlo HMM
d. EM Algorithm for Density Esitimation with Dirichlet Process Prior

Descriptions will be brief. Details will be available in the references cited.

### 2. Sequential Marginal Likelihood Online Change Detector

Consider sequential data

$$y_{1:t} := (y_1,..., y_t) = (y_{1:t-1}, y_t),\ t = 1,2,.., \qquad (1)$$

from a system with changing parameter(s). Challenges arise when parameter changes abruptly where one sometimes wants to detect changes.

Under the present setting, a plausible likelihood function can be of the form

$$P(y_t \mid \theta_t, \beta_t) \qquad (2)$$

where $\theta_t$ is the parameter vector and $\beta_t$ is a possible hyperparameter.

One method of learning parameters as well as hyperparameter for use if the parameter vector changes are unknown, is to consider stochastic search dynamics (sequential prior) for parameters and hyperparameter:

$$P(\theta_t, \beta_t \mid \theta_{t-1}, \beta_{t-1}; y_{t-1}) \qquad (3)$$

where in [1] $y_{t-1}$, the information available at $t$-1, is taken into account instead of random walk, where no information is taken into account. An attempt is made to perform change detection by examining the time dependency of the *sequential marginal likelihood*

$$t \mapsto P(y_t \mid y_{1:t-1})$$

where

$$\begin{cases} P(y_t \mid y_{1:t-1}) = \iint P(y_t \mid \theta_t, \beta_t) P(\theta_t, \alpha_t \mid y_{1:t-1}) d\theta_t d\alpha_t \\ P(y_{1:t}) = \prod_{s=1}^{t} P(y_s \mid y_{1:s-1}) \end{cases} \qquad (4)$$

with $\alpha_t := (\beta_t, \gamma_t)$. One of the main reasons for using $P(y_t|y_{1:t-1})$ as change detector is that this quantity can be regarded as the "degree of surprise" of new data $y_t$ with respect to the past sequential data $y_{1:t-1}$. Equation (4) is evaluated by Sequential Monte Carlo so that linearity as well as Gaussian assumptions are

not made.

Note that evaluating the marginal likelihood by

$$P(y_{1:t}) = \int \cdots \int P(y_{1:t} \mid \theta_{1:t}, \alpha_{1:t}) \\ P(\theta_{1:t}, \alpha_{1:t}) P(\theta_0, \alpha_0) d\theta_{1:t} d\alpha_{1:t} d\theta_0 d\alpha_0 \quad (5)$$

is often difficult with Monte Carlo because the likelihood function landscape can be complicated and the dimension of the multiple integral (5) can be high. This study uses the decomposition

$$P(y_{1:t}) = \prod_{s=1}^{t} P(y_s \mid y_{1:s-1})$$

and evaluate marginal likelihood sequentially.

The likelihood function in this study is defined by

$$P(y_t \mid \theta_t, \beta_t) := \frac{1}{Z(\beta_t)} \exp(-\frac{\beta_t}{2}(y_t - f(\theta_t))^2)$$

with parameter vector $\theta_t$ and $\beta_t$ is a hyperparameter that represents the uncertainty level of the observation with normalizing constant $Z(\beta_t)$. This likelihood function includes a supervised learning problem in which the data set is given as a pair ($y_t$, $x_t$), with $x_t$ as input while $y_t$ as output, so that $f(\theta_t) = f(x_t; \theta_t)$. The likelihood function also includes time series data $x_{0:t}$.For the sake of notational simplicity, the dependency on $x_t$ as well as on $x_{t-\tau::t-1}$ will be suppressed in the following arguments as long as confusion does not arise.

In order to perform Monte Carlo evaluation of marginal likelihood (5), assume that the draws at the previous step

$$(\theta_{0:t-1}^{(i)}, \alpha_{0:t-1}^{(i)}) \sim P(\theta_{0:t-1}, \alpha_{0t-1} \mid y_{1:t-1}), i = 1,...,N$$

are available. Use equation (4) to generate

$$(\tilde{\theta}_t^{(i)}, \tilde{\alpha}_t^{(i)}), i = 1,...,N$$

and evaluate

$$\hat{P}(y_t \mid y_{1:t-1}) = \sum_{i=I}^{N} P(y_t \mid \tilde{\theta}_t^{(i)}, \tilde{\alpha}_t^{(i)}) \tilde{w}_{0:t-1}^{(i)}$$

Change Detection in Nonlinear Dynamical System

Consider a noise-corrupted version of the well-known Roessler dynamical system, as described by

$$\frac{dx}{dt} = -y - z + v_x, \frac{dy}{dt} = x + ay + v_y, \frac{dz}{dt} = bx - cz + xz + v_z$$

where $v_x$, $v_y$ and $v_z$ are noise processes.

Change in this particular demonstration is incurred by

$$\begin{cases} a = a_1 = 0.22, & 0 \le t \le 500 \\ a = a_2 = 0.35, & 500 < t \le 700 \end{cases}$$

where $(b,c)$=(0.4,4.5) is fixed in the Roessler system. The noise amplitude is set at $\sigma$=0.01. Fig. 1 gives the $x$ trajectory between $t$=400 and $t$=600. The change, which occurs at $t$ =600, appears to be subtle. Using the delay-coordinate embedding [1], one can attempt to embed the 3-dimensional dynamics onto a 1-dimensional delay coordinate system [1]. Fig.2 plots $P(x_t|x_{0:t-1})$ in log scale. A reasonable dip is discernible shortly after $t$=500.

Application to online face detection with video sequences is found in [2].

## 3. Natural Sequential Prior for Online Learning

In [3] the authors consider the class of priors of the form

$$P(\theta_t \mid \theta_{t-1}; \Sigma_t) := \frac{1}{Z((\Sigma_t)^{-1})} e^{-\frac{1}{2}(\theta_t - \theta_{t-1})^T \Sigma_t^{-1} (\theta_t - \theta_{t-1})}$$

and design $\Sigma_t$ which minimizes approximate *expected K-L divergence* between $P(\cdot \mid \theta_t)$ and $P(\cdot \mid \theta_{t-1})$ with respect to the prior:

$$Expected\ KL := \int D_{KL}(\theta_t \mid \theta_{t-1}) P(\theta_t \mid \theta_{t-1}; \Sigma_t) d\theta_t$$

$$D_{KL}(\theta_t \mid \theta_{t-1}) := \int P(y \mid \theta_{t-1}) \log \frac{P(y \mid \theta_{t-1})}{P(y \mid \theta_t)} dy$$

while the *Shannon entropy* of the prior

$$S(P; \theta_{t-1}; \Sigma_t) := -\int P(\theta_t \mid \theta_{t-1}; \Sigma_t) \log P(\theta_t \mid \theta_{t-1}; \Sigma_t) d\theta_t$$

held constant. This amounts to minimizing the expected KL divergence with stochastic search volume held constant. Note that.

$$D_{KL}(\theta_t \mid \theta_{t-1}) = \\ \frac{1}{2}(\theta_t - \theta_{t-1})^T \mathcal{F}_{t-1} (\theta_t - \theta_{t-1}) + O(\| \theta_t - \theta_{t-1} \|^3)$$

where $\mathcal{F}_{t-1}$ stands for the Fisher information matrix at $t$-1 so that an approximate solution to this constrained minimization is given by

$$\Sigma_t \propto \mathcal{F}_{t-1}$$

provided that $\|\theta_t - \theta_{t-1}\|$ is small. This gives rise to the proposed sequential prior

$$P(\theta_t \mid \theta_{t-1}; \mathcal{F}_{t-1}) := \frac{1}{Z(\mathcal{F}_{t-1})} e^{-\frac{1}{2}(\theta_t-\theta_{t-1})^T \mathcal{F}_{t-1}(\theta_t-\theta_{t-1})} \quad (6)$$

It should be observed that if the covariance matrix is taken to be the identity matrix, then (6) reduces to the conventional random walk sequential prior where no observation information is taken into account.

The proposed scheme often outperforms the conventional random walk sequential prior [3].

## 4. Monte Carlo HMM Sports Event Detector

Consider a sports event detection problem where one wants to perform an automatic indexing of, e.g., kick off, corner kick, free kick, throw in, goal kick, among others. This is a nontrivial task because of the difficulties associated with manual indexing i.e., cost, time, and human resources/errors to name a few.

Reference [4] formulates the problem via HMM:

*stochastic dynamics*

$$P(x \mid \theta) = \prod_{t=1}^{T} P(x_t \mid x_{t-1}, \theta) P(x_0 \mid \theta)$$

*observation*

$$P(y \mid x, \theta) = \prod_{t=1}^{T} \prod_{l=1}^{L} P(f_{l,t} \mid x_t, \theta) P(e_t \mid x_t, \theta)$$

where $(y_1, \ldots, y_T) := y = (f, e)$ with $f$ feature, $e$ event, $(x_1, \ldots, x_T) := x$, trajectory of stochastic dynamics, $(a, b, c, \pi) := \theta$, *parameters, and* $L :=$ #(symbols)(see Fig.4)).

In [4], features include position/velocity of each player, average velocity, variance, median of the players, among others. The number of the available features are large so that one needs to perform several preprocessing steps to extract important features. The likelihood functions are multinomial while the priors for the parameters are assumed to be Dirichlet. A schematic picture of the algorithm is depicted in Fig5.

Performance is evaluated, by defining *event predictive capability ratio* [4]:

$$S(event) := \frac{\hat{P}(e_t^{new} \mid f_t^{new})(when\ event\ actually\ occurs)}{\hat{P}(e_t^{new} \mid f_t^{new})(when\ event\ does\ not\ occur)}$$

One observes that

$$S(event): \begin{cases} > 1, high\ predictive\ capability \\ < 1, low\ predictive\ capability \\ = 1, little\ use \end{cases}$$

Table 1 shows this quantity for several events. Fig. 6 shows a detection example where the solid rectangle shows that the event (corner kick) actually takes place. The performance appears reasonable.

## 5. DPEM

Consider the Gaussian mixture model

$$y_i \sim p(y \mid \theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(y; \mu_k, \sigma_k^2)$$
$$i = 1,2,\ldots,N$$
$$\theta := (\mu_1, \ldots, \mu_K, \sigma_1^2, \ldots, \sigma_K^2, \pi_1, \ldots, \pi_K)$$
$$\theta \sim G_0(\theta)$$

where $G_0$ stands for the underlying base distribution. If $K$ is known, then one can perform, e.g., EM to estimate the parameters involved. If, however, the number of mixture components is ***unknown***, the problem becomes nontrivial. One possible approach is to choose $K$ in terms of appropriate information criterion, e.g., AIC/BIC. Another possible algorithm is via the Dirichlet process prior:

$$y_i \sim p(y_i \mid \theta_i) = \mathcal{N}(y_i; \mu_i, \sigma_i^2)$$
$$i = 1,2,..$$
$$\theta_i := (\mu_i, \sigma_i^2) \sim G(\theta_i)$$
$$G \sim \mathcal{DP}(G_0, \alpha)$$

An EM algorithm is proposed in [5] with stick-breaking construction (Fig.7). Several applications are found in [5].

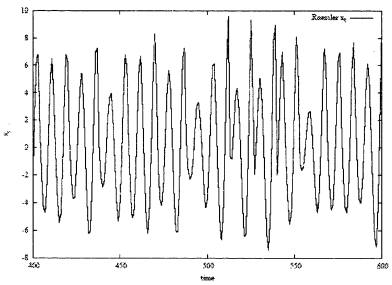


Fig. 1 $x$-coordintate of the Roessler system. The change at $t$=500 appears to be subtle. (from [1] with permission.) © IEEE

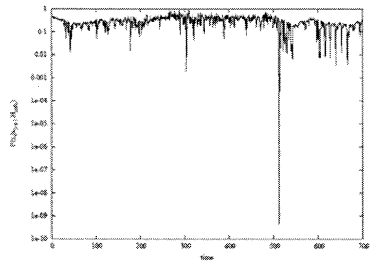

Fig. 2 A discernible dip is seen in the Sequential Marginal Likelihood. (from [1] with permission.) © IEEE

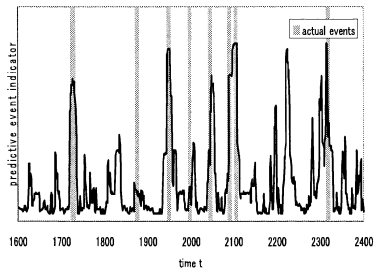


**Fig. 6 Detection example**

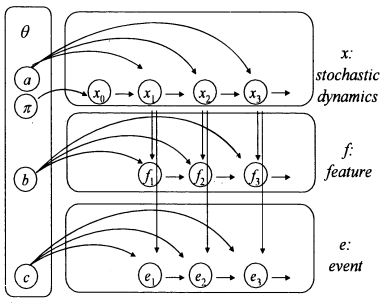


**Fig. 4 HMM sports event detection**

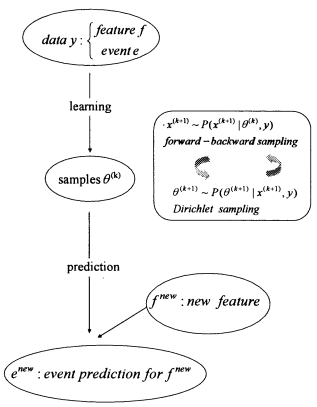


**Fig. 5 Monte Carlo HMM**

| kick off | corner kick | free kick | throw in | goal kick |
|---|---|---|---|---|
| 12.6 | 7.45 | 1.15 | 2.09 | 1.97 |

**Talbe 1 Event predictive capability ratio**

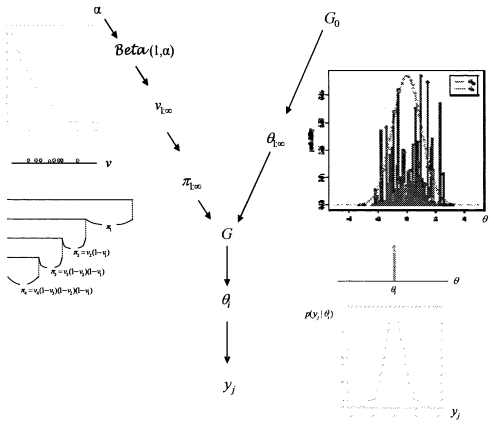


**Fig. 7 Stick-breaking construction of DP**

## References


[1] T. Matsumoto and K. Yosui, "Adaptation and Change Detection with a Sequential Monte Carlo Scheme", IEEE Transactions on Systems, Man and Cybernetics, vol. 37, No. 3, pp. 592-606, June 2007

[2] A. Matsui, S. Clippingdale and T. Matsumoto, "Bayesian Sequential Face Detection with Automatic Re-initialization", Proc. International Conference on Pattern Recognition, Tampa, Florida, Dec. 8-11, 2008

[3] K. Sega, Y. Nakada, and T. Matsumoto, "Online Bayesian Learning for Dynamical Classification Problem Using Natural Sequential Prior", Proc. IEEE International Workshop on Machine Learning for Signal Processing, pp.392-397, Cancun, Mexico Oct.16-19, 2008

[4] S. Motoi, Y. Nakada, T. Misu, T. Yazaki, T. Matsumoto and N. Yagi, "A Hierarchical Bayesian Hidden Markov Model for Multi-Dimensional Discrete Data", in *Frontiers in Robotics, Automation and Control,* In-Tech Publications, pp.357-374, Oct. 2008

[5] T. Kimura, Y. Nakada, T. Matsumoto and A. Doucet. "Recursive EM Algorithm for Parameter Estimation in Mixture Models". to appear