EXTENSIONS AND PROBABILISTIC ANALYSIS OF DYNAMIC MODEL SELECTION

Kenji Yamanishi and Ei-ichi Sakurai

The University of Tokyo, Hongo 7-3-1, Bunkyo-ku, Tokyo, 113-8656, JAPAN
yamanishi@mist.i.u-tokyo.ac.jp, Eiichi_Sakurai@mist.i.u-tokyo.ac.jp

ABSTRACT

This paper addresses the issue of tracking structural changes of probabilistic models from time series. Suppose that the structure of probability distributions for data (e.g., the number of parameters) changes over time, we are interested in tracking when and how it changes. We call the issue *dynamic model selection (DMS)*. In previous works batch algorithms for DMS have been proposed on the basis of the minimum description length (MDL) principle and their coding theoretic analysis has been provided. In this paper we give some extensions of existing DMS algorithms and give new probabilistic analysis for DMS from the standpoint of hypothesis testing theory. We thereby obtain a broader family of DMS algorithms and give a new rationale for them. The extensions of DMS algorithms include a sequential DMS algorithm, which outputs a model sequence in linear time every time a data sequence of a fixed size is input. We give a coding theoretic bound for it and show that it runs much more efficiently than existing ones. As for the probabilistic analysis of DMS, we derive bounds on Type 1 and 2 error probabilities to show that they are related to the model transition probabilities and the Kullback-Leibler divergence between the distributions to be switched each other. Throughout the paper we give a new insight of probabilistic modeling for non-stationary data sources from both of the information-theoretic and computational aspects of the MDL principle.

1. INTRODUCTION

In this paper we consider the non-stationary situation where the structure of probabilistic model for information sources may change over time. Here the structure of probabilistic model refers to as, for example, the number of parameters, the number of clusters, the number of latent variables, etc., all of which are to be selected in the context of statistical model selection. We are interested in the issue of tracking when and how the structure of probabilistic model changes. We may call such an issue *dynamic model selection* according to Yamanishi and Maruyama's pioneering work [15](see also [14]).

The model selection criteria such as AIC [1], BIC [11], cannot straightforwardly be applied to the issue of dynamic model selection, since the data source is assumed to be intrinsically non-stationary. In [15] Yamanishi and Maruyama considered the situation where a probabilistic model switches to another as time goes by and proposed algorithms for finding the optimal path of model switching on the basis of the *minimum description length (MDL) principle* [8]. More precisely, the optimal model sequence was defined as the one that attains the shortest code-length required for the encoding of data sequence as well as model sequence itself. Such an optimal sequence was computed using the dynamic-programming method in time $O(n^2)$ (n:sample size). Erven et.al.[4] also independently introduced the notion of *switching distribution* in a different context. They have shown that the switching distribution achieves consistency as MDL does as well as the fastest rate of convergence as AIC does.

DMS is closely related to the issues of *tracking best experts* [5] and *derandomizing stochastic prediction strategies* [13],[2]. In them, under the assumption that the best predictor attaining the smallest loss may change over time, it was addressed how one could realize the prediction strategy attaining almost the same losses which the series of the best predictors suffer from.

In previous works [15], batch algorithms for DMS have been proposed, i.e., they take a data sequence as input and output a model sequence that explains it best. Their coding theoretic analysis has been made. In this paper we give some extensions of existing DMS algorithms and give a new analysis for DMS from the standpoint of hypothesis testing theory. We thereby obtain a broader family of DMS algorithms and give a new rationale for them.

As for the extension of DMS algorithms, we propose two variants of the original DMS algorithms. One is a linear-time sequential algorithm that outputs a model sequence every time a data sequence of a fixed size is input. We give a coding theoretic bound and show that it runs more efficiently than existing ones. The other is a new type of DMS algorithms with model resetting (i.e., probability models are initialized at each change-point so that the model can better fit the data than the original DMS).

As for the probabilistic analysis of DMS, we derive upper bounds Type 1 and 2 error and relate them to model transition probabilities and the Kullback-Leibler divergence between the distributions to be switched each other.

Throughout the paper we give a new insight of probabilistic modeling for non-stationary data sources from both of the information-theoretic and computational aspects of the MDL principle.

2. BATCH DYNAMIC MODEL SELECTION

We first give a mathematical framework of dynamic model selection. Let $\mathcal{X}$ be a finite set or a subset in $\mathbf{R}^s$ for some positive integer s, which we call the domain. Let us denote $X^t = X_1, \cdots, X_t \in \mathcal{X}^t$ and $X_a^b = X_a, \cdots, X_b \in \mathcal{X}^{b-a}$. We write a random variable as X and an observed value as x. We let $\mathcal{C}_k = \{P(X^t|\theta : k) : t = 1, 2, \cdots, \theta \in \Theta_k\}$ be the parametric class of probability mass functions over $\mathcal{X}$ specified by the index k where Θ_k is a compact real-valued parameter space. For example, k is the dimension of Θ_k. We refer to k as the *model*. Let $\mathcal{K} = \{1, \ldots K\}$ for a positive integer K. Suppose that k ranges over $\mathcal{K}$. and we are given $\mathcal{C} = \cup_{k\in\mathcal{K}}\mathcal{C}_k$.

We write conditional probability distributions as: $P(X_t|X^{t-1} : \theta, k) = P(X^t|\theta, k)/P(X^{t-1}|\theta, k)$, and $P(X_t|X_a^{t-1} : \theta, k) = P(X_a^t|\theta, k)/P(X_a^{t-1}|\theta, k)$.

We may employ as the conditional distribution associated with k the *Bayesian distribution* defined as:

$$P(X_t|X^{t-1} : k) = \int P(X_t|\theta : k)\pi(\theta|X^{t-1})d\theta,$$

where $\pi(\theta|X^{t-1}) = \pi(\theta)P(X^{t-1}|\theta : k)/\int P(X^{t-1}|\theta : k)d\theta$ is the Bayesian posterior density of θ.

We may also employ as the conditional distribution associated with k the *plug-in distribution* defined as:

$$P(X_t|X^{t-1} : k) = P(X_t|\hat{\theta}(X^{t-1}) : k),$$

where $\hat{\theta}(X^{t-1}) = \arg\max_{\theta\in\Theta_k} P(X^{t-1}|\theta : k)$ is the maximum likelihood estimator (m.l.e.) of θ from X^{t-1}. The m.l.e. can be replaced with any other estimators such as the MAP estimator or the discounting m.l.e. defined by $\hat{\theta}(X^{t-1}) \stackrel{\text{def}}{=} \arg\max_{\theta\in\Theta_k} \sum_{j=1}^{t-1} r(1-r)^{t-j} \log P(X_j|\theta : k)$, where $0 < r < 1$ is a discounting factor.

Further we may employ as the conditional distribution associated with the model k the *sequentially normalized maximum likelihood* (SNML) distribution [9] defined as:

$$P(X_t|X^{t-1} : k) = \frac{P(X_t \cdot X^{t-1}|\hat{\theta}(X_t \cdot X^{t-1}) : k)}{\sum_X P(X \cdot X^{t-1}|\hat{\theta}(X \cdot X^{t-1}) : k)},$$

where $\hat{\theta}(X_t \cdot X^{t-1})$ is the m.l.e. of θ from $X_t \cdot X^{t-1}$.

For a given positive integer n, $m \le n$, let

$$\mathbf{s} = (t_0, k_0), \ldots, (t_m, k_m)$$
$$(t_0 = 1 < \cdots < t_m,\ t_{m+1} = n + 1,\ k_1, .., k_m \in \mathcal{K})$$

be a sequence of pairs of a model change point t_i and its corresponding model value k_i $(i = 0, \ldots, m)$ where m is the number of change points. We define $\mathcal{S}_n(\mathcal{K})$ by a set of all possible sequences of such pairs for data size n.

We define a *switching distribution* associated with $\mathbf{s} \in \mathcal{S}_n(\mathcal{K})$ as follows:

Definition 1 (Switching distribution) *For given n, we define a* switching distribution *associated with $\mathbf{s} \in \mathcal{S}_n(\mathcal{K})$, in a conditional form: For each $t(= 1, \cdots, n)$,*

$$P_{SW}(X_t|X^{t-1} : \mathbf{s}) \stackrel{\text{def}}{=} P(X_t|X^{t-1} : k_i)$$
$$(t_i \le t \le t_{i+1} - 1,\ \ i = 1, \cdots, m). \qquad (1)$$

We then write the switching joint distribution for X^n as:

$$P_{SW}(X^n|\mathbf{s}) = \prod_{i=0}^{m} \prod_{t=t_i}^{t_{i+1}-1} P(X_t|X^{t-1} : k_i).$$

The batch DMS problem is formulated as follows: *When given a data sequence $x^n = x_1, \cdots, x_n$, find a model sequence $\mathbf{s} \in \mathcal{S}_n(\mathcal{K})$.* Here we may define a criterion for evaluating the goodness of a model sequence from the standpoint of the MDL principle as follows:

$$-\log P_{SW}(x^n|\mathbf{s}) + \ell(\mathbf{s}) \Longrightarrow \text{minimum}, \qquad (2)$$

where $\ell(\mathbf{s})$ is the prefix code-length for $\mathbf{s}$. This is really a two-stage coding version of the MDL principle for model sequence selection.

Let d $(1 \le d < K/2)$ be a positive integer. We make an assumption that the model transition probability distribution takes the following specific form: At each time t,

$$P_t(k_t|k^{t-1} : \alpha) = \begin{cases} P_t(|k_t - k_{t-1}| = r) = \alpha_r \\ \quad (r = 0, 1, \cdots, d), \\ 0 \text{ otherwise} \end{cases}$$

where we set the parameter vector $\alpha = (\alpha_0, \cdots, \alpha_d)$ and $\sum_{r=0}^{d} \alpha_r = 1$. We set $P(k_t = k_{t-1} + r) = P(k_t = k_{t-1} - r) = \alpha_r/2$ if $k_t = k_{t-1} + r$ and $k_t = k_{t-1} - r$ exist.

We may employ as an estimator of α the Krichevsky & Trofimov's estimator [6] given as follows: The i-th component of α at time $t-1$ is given by

$$\hat{\alpha}_{r,t-1} = \frac{N_{r,t-1} + 1/2}{t - 1 + (d+1)/2}, \ \ (r = 0, \cdots, d). \qquad (3)$$

where $N_{r,t-1}$ is the number of occurrences of events: $|k_t - k_{t-1}| = r$. Then the criterion (2) is written as follows:

$$\sum_{t=1}^{n} -\log P(x_t|x^{t-1} : k_t) + \sum_{t=1}^{n} -\log P_t(k_t|k^{t-1} : \hat{\alpha}_{t-1}) \qquad (4)$$

We may call (4) the *batch DMS(Dynamic Model Selection) criterion*.

According to [15], we have the following theorem on the existence of batch DMS algorithm in the case of $d = 1$.

Theorem 2 (Coding-theoretic bound for DMS algorithm for switching distributions [15]) *For the class of switching distributions, there exists a batch DMS algorithm that outputs a model sequence attaining the minimum of (4) and runs in time $O(Kn^2)$. Then the total code-length is upper bounded by:*

$$\min_m \min_{(t_0,k_0),\cdots,(t_m,k_m)} \left\{ \sum_{j=0}^{m} \sum_{t=t_j}^{t_{j+1}-1} -\log P(x_t|x^{t-1} : k_i) \right.$$
$$\left. + nH\left(\frac{m}{n}\right) + \frac{1}{2}\log n + m + \log K + o(\log n) \right\}. \qquad (5)$$

We denote the algorithm as in Theorem 2 as DMS1. Although the details of DMS1 is omitted here, the key idea of the design of DMS1 is summarized as follows:

1) Dynamic programming-based model sequence search. We learn a number of probability distributions with different models in parallel. Then we employ the Viterbi-like dynamic-programming method [12] to search a model path that attains the minumum of the DMS criterion (4).

2) Prequential analysis of model sequences. The total code-length for a model sequence is calculated in terms of the plug-in type predictive code-length, which we call the *predictive stochastic complexity* [8]: $\sum_{t=1}^{n} -\log P_t(k_t|k^{t-1} : \alpha_{t-1})$. The model fitting based on the predictive stochastic complexity is appropriate for the sequential analysis for non-stationary data. Such analysis is also called *prequential analysis* [3].

Theorem 2 implies that the total code-length for DMS1 is upper-bounded by the minimum total sum of code-length for the data sequence relative to the model path associated with $\mathbf{s}$ plus the quantity: $nH(m/n) + (1/2)\log n + m + \log K + o(\log n)$ where the minimum is taken with respect to the number of change points, their locations, and their corresponding model values.

Next we introduce the notion of resetting distribution and extend the DMS framework for them. In switching distributions, at a change point one model switches to another. After the switching, for each model the distributions associated with it is learned from all of the past data. Meanwhile, in resetting distributions, at a change point for each model the distribution associated with it is once initialized and is learned from the data starting from the latest change point. Hence resetting distributions better fit to data sequence than switching ones.

Definition 3 (Resetting distribution) *For given n, we define a* resetting distribution *associated with $\mathbf{s} \in \mathcal{S}_n(\mathcal{K})$, in a conditional form: For each $t(= 1, \cdots, n)$,*

$$P_{RE}(X_t|X^{t-1} : \mathbf{s}) \stackrel{\text{def}}{=} P(X_t|X_{t_i}^{t-1} : k_i) \quad (t_i + 1 \le t \le t_{i+1},\ \ i = 0, \cdots, m). \tag{6}$$

We then write the resetting joint distribution for X^n as:

$$P_{RE}(X^n|\mathbf{s}) = \prod_{i=0}^{m} \prod_{j=t_i+1}^{t_{i+1}} P(X_j|X_{t_i}^{j-1} : k_i).$$

For the class of resetting distributions, similarly with switching distributions, we may define the batch DMS criterion with respect to $\mathbf{s}$ as follows:

$$-\log P_{RE}(x^n|\mathbf{s}) + \ell(\mathbf{s}) \Longrightarrow \text{minimum}, \tag{7}$$

Batch DMS Algorithm: DMSR

k: model index, τ: session length from the latest change point,
$\boldsymbol{N} = (N_0, \cdots, N_d)$ (N_i: number of occurrences of $|k_t - k_{t-1}| = i$),
$S(k, \boldsymbol{N}, \tau, t)$: cumulative code-length until the time t when given $k, \boldsymbol{N}, \tau$ at time t,
$\hat{P}_t(k|k') = \{P_t(|k-k'| = i) : i = 0, \cdots, d\}$: model transistion probability estimated at time t.

Step 1. Initialization.

($t = 1, \forall k, \forall k'$, $P(\cdot|x_0 : k), \hat{P}_1(k|k')$: given)
$\boldsymbol{N} = (0, \cdots, 0)$, $S(k, \boldsymbol{N}, 0, 1) = \log K - \log P(x_1 \mid x_0 : k)$, $\boldsymbol{K}(k, \boldsymbol{N}, 0, 1) = (k)$.

Step 2. Model Sequence Search.

for $t = 2$ **to** $t = n$ **do**
For each $k = 1, \cdots, K$,
[Model Selection]
If $\tau = 1$ **then**

$$S(k, \boldsymbol{N}, \tau, t) = \min_{k' \neq k, \boldsymbol{N}', \tau'} \{S(k', \boldsymbol{N}', \tau', t-1) - \log P(x_t|x_{t-1} : k) - \log \hat{P}_t(k|k')\},$$
$$(\tilde{k}, \tilde{\boldsymbol{N}}, \tilde{\tau}) = \arg \min_{k' \neq k, \boldsymbol{N}', \tau'} \{S(k', \boldsymbol{N}', \tau', t-1) - \log P(x_t|x_{t-1} : k) - \log \hat{P}_t(k|k')\},$$
$$\boldsymbol{K}(k, \boldsymbol{N}, \tau, t) = \boldsymbol{K}(\tilde{k}, \tilde{\boldsymbol{N}}, \tilde{\tau}, t-1) \oplus k,$$

where $\tilde{\boldsymbol{N}} = (\cdots, N_{|k-k'|} - 1, \cdots)$.
else then

$$S(k, \boldsymbol{N}, \tau, t) = S(k, \tilde{\boldsymbol{N}}, \tau - 1, t-1) - \log P(x_t|x_{t-\tau+1}^{t-1} : k) - \log \hat{P}_t(k|k)\},$$
$$\tilde{\boldsymbol{N}} = \arg\min_{\boldsymbol{N}'} \{S(k, \boldsymbol{N}', \tau - 1, t-1) - \log P(x_t|x_{t-\tau+1}^{t-1} : k) - \log \hat{P}_t(k|k)\},$$
$$\boldsymbol{K}(k, \boldsymbol{N}, \tau, t) = \boldsymbol{K}(k, \tilde{\boldsymbol{N}}, \tau - 1, t-1) \oplus k.$$

[Estimation of Model Transition Probabilities]

$$\hat{P}_t(k|k') = \frac{N_i + 1/2}{\sum_{i=0}^{d} N_i + (d+1)/2} \quad (i = 0, \cdots, d).$$

end for

Step 3. Output. ($t = n$)
$(k^*, \boldsymbol{N}^*, \tau^*) = \arg\min_{k, \boldsymbol{N}, \tau} S(k, \boldsymbol{N}, \tau, n)$.
$(k_1^*, \ldots, k_n^*) = \boldsymbol{K}(k^*, \boldsymbol{N}^*, \tau^*, n)$.
Output $\mathbf{s} = \phi(k_1^*, \ldots, k_n^*)$.

Figure 1. Batch DMS Algorithm: DMSR

Similarly with DMS1, we may design a DMS algorithm for resetting distribution, in which the dynamic programming method is employed for model sequence search while the prequnetial method is employed for computing

code-lengths for model sequences. We denote such an algorithm DMSR. It differs from DMS1 in that DMSR conducts dynamic programming-based model search starting from the latest change point in a recursive way while DMSR does that starting from the initial data of the sequence. Hence DMSR requires more computational costs than DMS1 The sketch of DMSR is shown in Fig. 1. The following theorem shows the existence of a batch DMS algorithm for resetting distributions.

Theorem 4 (Coding-theoretic bound for batch DMS algorithms for resetting distributions) *For the class of resetting distributions, there exists a batch DMS algorithm that outputs a model sequence attaining the minimum of (4) and runs in time $O(Kn^3)$. Then the total code-length is upper bounded by:*

$$\min_{m} \min_{(t_0,k_0),\cdots,(t_m,k_m)} \left\{ \sum_{j=0}^{m} \sum_{t=t_j+!}^{t_{j+1}} -\log P(x_t|x_{t_j}^{t-1}:k_i) + nH\left(\frac{n_0}{n},\ldots,\frac{n_d}{n}\right) + \frac{d}{2}\log n + m + \log K + o(\log n) \right\}, \tag{8}$$

where $H(z_0,\ldots,x_m) = -\sum_{i=0}^{d} \log z_i \log z_i$, n_r is the number of occurrences of the event: $|k_t - k_{t-1}| = r$ $(r = 0,\cdots,d)$ in k^n.

By Theorem 2 and 4, we see that the computation time $O(Kn^3)$ for DMSR for resetting distributions is greater than that for DMS1 for switching distributions–$O(Kn^2)$. Meanwhile, in the case of $d = 1$, the bound (5) is greater than (8) in general when the nature of data rapidly changes since the resetting makes better fitting to data.

3. SEQUENTIAL DYNAMIC MODEL SELECTION

Next we extend DMS into the sequential setting. Supposing that data is sequentially given, we wish to sequentially identify change points after seeing a data sequence of some fixed length B so that the total code-length for the data that have ever been observed is minimal.

Below $\mathcal{S}_{(a,b)}(\mathcal{K})$ represents the set of all $\mathbf{s} \in \mathcal{S}_n(\mathcal{K})$ such that $\mathbf{s}$ starts from $t = a$ and ends at $t = b$. A *sequential DMS algorithm* is an algorithm that at each time t takes as input x_t $(t = 1, 2, \cdots)$, $0 < B < \infty$(window size), then given $\mathbf{s}_1^{t-B-1} \in \mathcal{S}_{(1,t-B-1)}(\mathcal{K})$, outputs $\mathbf{s}_{t-B}^{t}$ so that the following quantity is minimum with respect to $\mathbf{s}_{t-B}^{t} \in \mathcal{S}_{(t-B,t)}(\mathcal{K})$.

$$-\log P_{SW}(x^t|\mathbf{s}_1^{t-B-1} \oplus \mathbf{s}_{t-B}^{t}) + \ell(\mathbf{s}_1^{t-B-1} \oplus \mathbf{s}_{t-B}^{t}),$$

where $\oplus$ denotes the addition of a new component; i.e., $(a,b) \oplus c = (a,b,c)$. $\ell(\mathbf{s}_1^{t-B-1} \oplus \mathbf{s}_{t-B}^{t})$ denotes the prefix code-length of $\mathbf{s}_1^{t-B-1} \oplus \mathbf{s}_{t-B}^{t}$), which is computed as the predictive code-length as with the batch DMS criterion.

We introduce a sequential DMS algorithm for switching distributions, which was developed in [10], and we denote as SDMS. The key idea of this algorithm is that it checks the existence of a change-point using the dynamic-programming method after seeing the data sequence of a fixed length B, and this process sequentiality goes on by sliding the window of length B. This mechanism makes the algorithm SDMS work sequentially and efficiently. The flow of SDMS is shown in Fig. 2. The following theorem gives a coding-theoretic analysis of SDMS.

Sequential DMS Algorithm: SDMS
k: model index, u: state, $B(>0)$: constant,
$S(k,u,t)$: cumulative code-length until the time t when given k, u at time t,
$\hat{P}_t(k|k') = \{P_t(|k-k'| = i) : i = 0,\cdots,d\}$: model transistion probability estimated at time t.
Step 1. Initialization.
($t = 1, \forall k, \forall k'$, $P(\cdot|x_0 : k)$, $\hat{P}_1(k|k')$: given)
$S(k,1) = \log K - \log P(x_1 \mid x_0 : k)$, $\boldsymbol{K}(k,1) = (k)$.
Step 2. Model Sequence Search.
For $t = 2$ **to** $t = n - B$ **do**
For each $k = 1,\cdots,K$,
[Model Selection]

$$S(k,t) = S(\tilde{k}, t-1) - \log P(x_t|x^{t-1}:k) - \log \hat{P}_t(k|\tilde{k}),$$
$$\boldsymbol{K}(k,t) = \boldsymbol{K}(\tilde{k}, t-1) \oplus k,$$
$$\tilde{k} = \arg\min_{k'}\{S(k', t-1) + \min_{k_{t+1},\cdots,k_{t+B}} \sum_{\tau=t}^{t+B} \{-\log P(x_\tau|x^{\tau-1}:k_\tau) - \log \hat{P}_\tau(k_\tau|k_{\tau-1})\}\} \quad (k_{t-1} = k', k_t = k).$$

[Estimation of Model Transition Probabilities]

$$\hat{k}^\tau = \boldsymbol{K}(k', t-1) \oplus (k, k_{t+1},\cdots,k_\tau)$$
$$\hat{P}_\tau(k_\tau|k_{\tau-1}) = \frac{N_i(\hat{k}^\tau) + 1/2}{\sum_{i=0}^{d} N_i(\hat{k}^\tau) + (d+1)/2}$$
$$(i = 0,\cdots,d).$$

end for
Step 3. Output. ($t = n$)

$$(k^*, k_{n-B+1}^*,\cdots,k_n^*) = \arg\min_{k', k_{n-B+1},\cdots k_n} \{S(k', n-B) + \sum_{\tau=n-B+1}^{n} \left(-\log P(x_\tau|x^{\tau-1}:k_\tau) - \log \hat{P}_\tau(k_\tau|k_{\tau-1})\right)\}.$$
$$(k_1^*,\ldots,k_n^*) = \boldsymbol{K}(k^*, n-B) \oplus (k_{n-B+1}^*,\cdots,k_n^*).$$
Output $\mathbf{s} = \phi(k_1^*,\ldots,k_n^*)$.

Figure 2. Sequential DMS Algorithm: SDMS

Theorem 5 (Coding-theoretic bound for sequential DMS algorithm for switching distributions [10]) *There exists a sequential DMS algorithm that runs in computation time*

$O(KB^2n)$ *and outputs* s *for which the total code-length for encoding* x^n *and* k^n *is upper-bounded by:*

$$\min_{m} \overset{*}{\min_{(t_0,k_0),\cdots,(t_m,k_m)}} \left\{ \sum_{j=0}^{m} \sum_{t=t_j}^{t_{j+1}-1} -\log P(x_t|x^{t-1}:k_i) + nH\left(\frac{n_0}{n},\ldots,\frac{n_d}{n}\right) + \frac{d}{2}\log n + m + \log K + o(\log n) \right\}, \tag{9}$$

where the notations follow Theorem 4 and the minimum of $\{(t_0,k_0),\ldots(t_m,k_m)\}$ *is taken over the range* $S(B,n) \subset S_n(\mathcal{K})$ *defined as follows:*

$$S(B,n) = \left\{ k^{n-B} \oplus \kappa^B | k^{n-B} \in \hat{D}(B,n), \kappa^B \in \mathcal{K}^B \right\}$$

$$\hat{D}(B,n) = \left\{ k_{op}^{n-B-1}(k) \oplus k | k \in \mathcal{K} \right\}$$

where $k_{op}^{n-B-1}(k) = \arg \min_{k^{n-B-1} \in \hat{D}(B,n-1)} \min_{\kappa^B \in \mathcal{K}^B}$

$$\left\{ \sum_{t=1}^{n} -\log_{k_t} P(x^n|\theta_k^{(t)}) + l(k^{n-B-1} \oplus k \oplus \kappa^B) \right\}.$$

We see that in the case of $d = 1$, SDMS requires less computation time
$O(KB^2n)$ than DMS1–$O(Kn^2)$ while the resulting total code-length for SDMS becomes greater than that for DMS1 due to the constraint for the minimum as above.

4. PROBABILISTIC ANALYSIS

In this section, we simplify the problem of DMS so that there are only two models; M_1 and M_2. We are thereby concerned with the issue of testing whether a model has switched or not. For the sake of simplicity, we consider the case where no real-valued parameters are associated with the models and model transition probabilities $P(M_i|M_j)$ $(i,j = 1,2)$ are known in advance, hence are not to be estimated. This setting is simple but valid enough for analyzing the essential property of tracking latent dynamics with the MDL principle. The batch DMS problem is then reduced to a hypothesis testing problem as follows.

Let t^* be a change-point. Let us the following two hypotheses: H_0 and H_1:

$$H_0: \quad M_1 \qquad \text{for } x_1^n = x_1 \cdots x_n,$$

$$H_1: \quad \begin{cases} M_1 & \text{for } x_1^{t^*} = x_1 \cdots x_{t^*}, \\ M_2 & \text{for } x_{t^*+1}^n = x_{t^*+1} \cdots x_n. \end{cases}$$

Suppose that $P(M_1|M_1) = P(M_2|M_2) = \omega > 1/2$ and $P(M_2|M_1) = 1 - \omega$. Define α by

$$\alpha \overset{\text{def}}{=} \log(\omega/(1-\omega)). \tag{10}$$

Suppose that $\alpha > 0$. We further rewrite $\prod_{t^*+1}^{n} P(x_j|x^{j-1}: M_1) = P(x_{t^*+1}^n|x^{t^*}: M_1)$ and $\prod_{t^*+1}^{n} P(x_j|x^{j-1}: M_2) = P(x_{t^*+1}^n|x^{t^*}: M_2)$. Then in the setting of hypothesis testing as above, DMS1 can be considered as a hypothesis testing algorithm such that H_0 is accepted if

$$-\log P(x_{t^*+1}^n|x^{t^*}: M_1) + \log P(x_{t^*+1}^n|x^{t^*}: M_2) - \alpha < 0 \tag{11}$$

else then H_1 is accepted.

Definition 6 (Type 1 and 2 error probabilities) *For given the length of data sequence* n*, the change-point time* t^**, we define Type 1 error probability for DMS1 by:*

$$Prob[x_{t^*+1}^n \sim P(X_{t^*+1}^n|x^{t^*}: M_1) \text{ and } (11) \text{ doesn't hold}],$$

and Type 2 error probability for DMS1 at time delay $h = n - t^*$ *by:*

$$Prob[x_{t^*+1}^n \sim P(X_{t^*+1}^n|x^{t^*}: M_2) \text{ and } (11) \text{ holds}].$$

Type 1 error probability is the probability that the model change has not yet occurred until time n but the change is incorrectly reported at time t^*. Type 2 error probability is the probability that the model change has already occurred at time t^*, but it is overlooked until time n, hence the time delay is $h = n - t^*$.

We prepare the following notations:

$$D_h(M_2||M_1)|_{x^{t^*}} \tag{12}$$

$$\overset{\text{def}}{=} \sum_{X_{t^*+1}^n} P(X_{t^*+1}^n|x^{t^*}: M_2) \log \frac{P(X_{t^*+1}^n|x^{t^*}: M_2)}{P(X_{t^*+1}^n|x^{t^*}: M_1)},$$

$$\beta \overset{\text{def}}{=} \frac{1}{h}\left(D_h(M_2||M_1)|_{x^{t^*}} - \alpha\right). \tag{13}$$

$D_h(M_2||M_1)|_{x^{t^*}}$ is considered as what we call the *Kullback-Leibler divergence* (the KL-divergence) between $P(X^h|x^{t^*}: M_2)$ and $P(X^h|x^{t^*}: M_1)$. We have the following theorem on Type 1 and 2 error probabilities for DMS1.

Theorem 7 (Probabilistic analysis for DMS1) *For the algorithm for tracking latent dynamics for switching distributions: DMS1, Type 1 error probability is always upper-bounded by* $2^{-\alpha}$*. Let* $x^{t^*} = x_1, \cdots, x_{t^*}$ *be an observed sequence until the change point* t^**. Suppose that for some* $0 < V < \infty$ *the variance of the random variable* $V_j = \log P(X_j|X^{j-1}: M_2)/P(X_j|X^{j-1}: M_1)$ *is upper-bounded by* V *for any* j*. If* $(1/h)D_h(M_2||M_1)|_{x^{t^*}} > \alpha$ *is satisfied for* α *as in (10), Type 2 error probability for DMS1 at time delay* h *is upper-bounded by* $2\exp(-h\beta^2/4V)$ *where* β *is as in (13).*

Proof. First we consider Type 1 error probability for DMS1. In the case where (11) does not hold, we have

$$-\log P(x_{t^*+1}^n|x^{t^*}: M_1) + \log P(x_{t^*+1}^n|x^{t^*}: M_2) - \alpha \geq 0. \tag{14}$$

Then Type 1 error probability for DMS1 is upper-bounded as follows:

$$\begin{aligned} & Prob[x_{t^*+1}^n \sim P(X_{t^*+1}^n|x^{t^*}: M_1) \text{ and } (11) \text{ doesn't holds}] \\ &= \sum_{X_{t^*+1}^n \cdots (14)} P(X_{t^*+1}^n|x^{t^*}: M_1) \\ &\leq \sum_{X_{t^*+1}^n \cdots (14)} P(X_{t^*+1}^n|x^{t^*}: M_2) 2^{-\alpha} \\ &\leq \sum_{X_{t^*+1}^n} P(X_{t^*+1}^n|x^{t^*}: M_2) 2^{-\alpha} \\ &= 2^{-\alpha}. \end{aligned} \tag{15}$$

Next we upper-bound Type 2 error probability for DMS1 as follows:

$$
\begin{aligned}
&Prob[x_{t^*+1}^n \sim P(X_{t^*+1}^n|x^{t^*}:M_2) \text{ and (11) hold}] \\
&= Prob[-\log P(X_{t^*+1}^n|x^{t^*}:M_1) \\
&\quad + \log P(X_{t^*+1}^n|x^{t^*}:M_2) - \alpha < 0] \\
&= Prob\left[\left(\log \frac{P(X_{t^*+1}^n|x^{t^*}:M_2)}{P(X_{t^*+1}^n|x^{t^*}:M_1)} - D_h(M_2||M_1)|_{x^{t^*}}\right)\right. \\
&\quad \left. < -\left(D_h(M_2||M_1)|_{x^{t^*}} - \alpha\right)\right] \\
&\leq Prob\left[\left|\log \frac{P(X_{t^*+1}^n|x^{t^*}:M_2)}{P(X_{t^*+1}^n|x^{t^*}:M_1)} - D_h(M_2||M_1)|_{x^{t^*}}\right| > h\beta\right],
\end{aligned}
$$

where the notations of α, β, and $D_h(M_2||M_1)|_{x^{t^*}}$ follow (10),(13), and (12), respectively. Note that if $(1/h)$ $D_h(M_2||M_1) > \alpha$ is satisfied, $\beta > 0$ holds. Under the assumption that for some V $(0 < V < \infty)$, $\mathrm{Var}[\log P(X_j|X^{j-1}:M_2)/P(X_j|X^{j-1}:M_1)]$ is upper-bounded by V, we use Bernstein's inequality to have the following bound:

$$
\begin{aligned}
&Prob\left[\left|\log \frac{P(X_{t^*+1}^n|x^{t^*}:M_2)}{P(X_{t^*+1}^n|x^{t^*}:M_1)} - D_h(M_2||M_1)|_{x^{t^*}}\right| > h\beta\right] \\
&\leq 2\exp\left(-\frac{h\beta^2}{4V}\right).
\end{aligned}
$$

This completes the proof. □

Theorem 7 shows that Type 1 error probability for DMS1 is always upper-bounded by a constant, which is determined by only the ratio of the two model transition probabilities. We also see that the Type 2 error probability for DMS1 decays in order $O(\exp(-h\beta^2))$, where the exponent factor depends on the ratio of model transition probabilities as well as the KL-divergence between probability distributions associated with M_2 and M_1. The larger the KL-divergence is, the faster Type 2 error probability converges to zero.

The following theorem shows the expected length of delay for DMS1. The proof is omitted due to the space limitation.

Theorem 8 (Expected delay for DMS1) *Suppose that β as in (13) is lower bounded by $\gamma(> 0)$ uniformly with respect to h. Then under the same assumption as in Theorem 7, the expected delay for DMS1 is upper-bounded by $O(1/\gamma^2)$.*

Theorem 8 shows that the smaller the KL-divergence between the switching distributions is, the larger the expected delay becomes in the order of $O(1/\gamma^2)$.

5. CONCLUSION

We have discussed on some extensions and probabilistic analysis of dynamic model selection. We have dealt with both of the batch and sequential DMS problems. For the batch DMS problem, we have introduced the existing algorithm DMS1 and its extension DMSR to the resetting scenario. For the sequential DMS problem, we have proposed the SDMS algorithm, which locally produces an optimal model sequence over a window of a fixed size and sequentially outputs them by moving the window. All of the algorithms are designed on the basis of the MDL principle using the dynamic programming-based model search with prequential analysis of model sequences. We have shown that they run in computation time $O(n^2)$, $O(n^3)$,and $O(n)$, respectively, and their upper bounds on the total code-lengths are in order: $\mathrm{DMSR} < \mathrm{DMS1} < \mathrm{SDMS}$. Further we have given probabilistic analysis to DMS1 from the view of hypothesis testing. We have shown that Type 1 error probability decays exponentially in the ratio of transition probabilities while Type 2 error probability decays where the rate depends on modle transition probabilities and the KL-divergence between probability distributions associated with the switching models. These results may form a new theoretical basis of dynamic model selection.

6. ACKNOWLEDGMENTS

This research has been done in MSR(Microsoft Research) CORE Project. The authors appreciate MICROSOFT CO., LTD. for the support.

7. REFERENCES

[1] H. Akaike: A new look at statistical model identification. *IEEE Trans. on Automatic Control*, 19(6): 716–723, 1974.

[2] N. Cesa-Bianchi and G.Lugosi: *Prediction, Learning, Games*. Cambridge Press, 2006.

[3] A.P. Dawid: Statistical theory: The prequential approach. *The Journal of the Royal Statistical Society A*, 147, Part2:278–292, 1984.

[4] T. van Erven and P.D. Grünwald and S. de Rooij: Catching up faster in Bayesian model selection and model averaging. *Advances in NIPS* 20, 2007.

[5] M. Herbster and M. K. Warmuth. Tracking the best expert. *Journal of Machine Learning*, 30(2):151–178, 1998.

[6] R. E. Krichevsky and V. K. Trofimov. The performance of universal encoding. *IEEE Trans. Inf. Theory*, 27:199–207, 1981.

[7] J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978.

[8] J. Rissanen: *Information and Complexity in Statistical Modeling*, Springer, 2007.

[9] J. Rissanen, T. Roos, and P. Myllymaki: Model selection by sequentially normalized least squares. *Journal of Multivariate Analysis*, vol.101, no.4, pp:829–849, 2009.

[10] E.Sakurai and K.Yamanishi. On liner time algorithm for sequential dynamic model selection. *IEICE Workshop on Information Based Induction Sciences and Machine Learning*, Japan, May, 2010.

[11] G. Schwarz: Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[12] A. J. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans. on Inform. Theory*, IT-13:260–267, 1967.

[13] V. Vovk: Derandomizing stochastic prediction strategies. *Machine Learning*, 35, pp:247–282, 1999.

[14] K. Yamanishi and Y. Maruyama: Dynamic syslog mining for network failure monitoring. *Proc. of KDD2005*, pp: 499-508, ACM Press, 2005.

[15] K. Yamanishi and Y. Maruyama: Dynamic model selection with its applications to novelty detection. *IEEE Trans. on Information Theory*,IT 53(6)：2180-2189, June, 2007.