A new method of Bayesian causal inference in non-stationary environments

Shuji Shinohara[1*], Nobuhito Manome[1,2], Kouta Suzuki[1,2], Ung-il Chung[1], Tatsuji Takahashi[1,3], Hiroshi Okamoto[1], Pegio-Yukio Gunji[4], Yoshihiro Nakajima[5], Shunji Mitsuyoshi[1]

[1] Department of Bioengineering, Graduate School of Engineering, The University of Tokyo, Tokyo, Japan

[2] Department of Research and Development, SoftBank Robotics Group Corp., Tokyo, Japan

[3] School of Science and Engineering, Tokyo Denki University, Saitama, Japan

[4] Department of Intermedia Art and Science, School of Fundamental Science and Technology, Waseda University, Tokyo, Japan

[5] Graduate School of Economics, Osaka City University, Osaka, Japan

* Corresponding author

E-mail: shinohara@bioeng.t.u-tokyo.ac.jp

AUTHOR CONTRIBUTIONS STATEMENT

S. S. conceived the extended confidence model and the extended Bayesian inference, conducted the meta-analysis and the simulation, wrote the manuscript. U. C., T. T., and H. O revised it. N. M., K. S., U. C., T. T., H. O., P.Y.G., Y. N., and S. M. contributed to the interpretation of the study findings. All authors participated in the editing and revision of the final version of the manuscript.

Abstract

Bayesian inference is the process of narrowing down the hypotheses (causes) to the one that best explains the observational data (effects). To accurately estimate a cause, a considerable amount of data is required to be observed for as long as possible. However, the object of inference is not always constant. In this case, a method such as exponential moving average (EMA) with a discounting rate is used to improve the ability to respond to a sudden change; it is also necessary to increase the discounting rate. That is, a trade-off is established in which the followability is improved by increasing the discounting rate, but the accuracy is reduced. Here, we propose an extended Bayesian inference (EBI), wherein human-like causal inference is incorporated. We show that both the learning and forgetting effects are introduced into Bayesian inference by incorporating the causal inference. We evaluate the estimation performance of the EBI through the learning task of a dynamically changing Gaussian mixture model. In the evaluation, the EBI performance is compared with those of the EMA and a sequential discounting expectation-maximization algorithm. The EBI was shown to modify the trade-off observed in the EMA.

Introduction

The objective of Bayesian inference is to deduce the hidden cause behind observed data by retrospectively applying statistical inferences. The relationship between Bayesian inference and brain function has attracted significant attention in recent years in the field of neuroscience [1,2]. In Bayesian inference, the degree of confidence for each hypothesis is updated based on a predefined model for each hypothesis by incorporating the current observational data. In other words, Bayesian inference is a process of narrowing down hypotheses (causal candidates) to one that best explains the observational data (the effects).

As an example, consider a situation in which one attempts to read another's emotions. Because one cannot directly view another's emotions, they can only be inferred from external cues (observation data), such as facial expressions and voice tones. The unknown emotions correspond to the hypotheses in Bayesian inference; i.e., it is the inferred target. In addition, a probability distribution representing what type of facial expression appears in what proportion when the other has a specified emotion corresponds to a model for each hypothesis. For example, if one has a model that "If a person is pleased, the person will smile, with an 80% chance," and if one observes that person to smile frequently, they will be more confident in the hypothesis that "the person is pleased." That is, by observing the data, the "effect" of a "smile," the emotion of "joy" is presumed as the cause of the "smile."

Attention should be paid to the following two points. First, to infer someone's emotions more accurately, it is better to have as much observation data as possible, but only if it is ensured that the emotion will not change during the estimation. Emotions change from moment to moment. In such an unsteady situation, it is necessary to consider whether the observation data are derived from the same emotion throughout the period being observed for the estimation. This may be a problem with online clustering of non-stationary data. The second point is that because a model for a stranger cannot be given in advance, it is necessary to learn and construct it from observed data. If this model is wrong, one cannot obtain the correct results from

observations to determine the emotion of the person.

Regarding the second point, there are methods such as the expectation-maximization (EM) algorithm and the K-means algorithm that perform inference and learning simultaneously. The EM algorithm is a method for obtaining the maximum likelihood estimate in a hidden-variable model [3] and it is often used for mixture models or latent-topic models, such as latent Dirichlet allocation [4]. K-means is the non-stochastic version of the EM algorithm [3].

In the previous example, hidden variables correspond to emotions such as joy or anger because they are not observable directly. By using the EM algorithm, a person's emotions can be estimated from the observed data while creating a model of emotion, based on the observed data. However, in the EM algorithm, it is necessary to provide all the observational data at one time. In practice, there are cases where data processing must be performed sequentially after each time the data are observed.

Various online algorithms have been proposed to deal with this situation [5–13]. For example, Yamanishi et al. [7] proposed a sequential discounting expectation-maximization (SDEM) algorithm that introduced the effect of forgetting to deal with unsteady situations where the inferred target changed. The algorithm is used in the fields of anomaly detection and change point detection. We also proposed EBI, incorporating causal reasoning into Bayesian inference, like an algorithm that performs inference and learning simultaneously [14].

In the field of cognitive psychology, experiments on causal induction have been performed to identify how humans evaluate the strength of causal relations between two events [15–19]. In a regular conditional statement of the form "if p then q," the degree of confidence is considered to be proportional to the conditional probability $P(q \mid p)$, which is the probability of occurrence of q given the existence of p [20]. In contrast, in the case of a causal relation, it has been experimentally demonstrated that humans have a strong sense of causal relation between a cause c and an effect e when $P(c \mid e)$ is high, as well as when $P(e \mid c)$ is high. Specifically, the causal intensity that people feel between c and e can be approximated by

the geometric mean of $P(e|c)$ and $P(c|e)$. This is called the "dual-factor heuristics" (DFH) model [19]. If the causal intensity between c and e is denoted as $DFH(e|c)$, then $DFH(e|c)=\sqrt{P(e|c)P(c|e)}\neq P(e|c)$. Here, note that $DFH(c|e)=DFH(e|c)$ is valid. Such inference is called "symmetry inference."

In this paper, we first describe the EBI, which replaces conditional inference in Bayesian inference with causal inference. Second, we show that the learning effect and forgetting effect are introduced into Bayesian inference by this replacement. Third, we evaluate the estimation performance of the EBI through the learning task of a dynamically changing Gaussian mixture model. In the evaluation, the performance is compared with SDEM, an online EM algorithm.

Methods

Bayesian inference

In Bayesian inference, several hypotheses h_k are first defined and models (probability distributions of data d for the hypotheses) are prepared in the form of conditional probabilities $P(d|h_k)$. This conditional probability is called "likelihood" in the case that the data are fixed, and this probability is considered to be a function of the hypothesis. In addition, the confidence $P(h_k)$ for each hypothesis is prepared as a prior probability. That is, one must have some prior estimate of the probability that h_k is true.

Assuming that the confidence for hypothesis h_k at time t is represented as $P^t(h_k)$ and data d^t were observed, the posterior probability is calculated as follows using Bayes' theorem.

$$P^t(h_k|d^t)=\frac{P^t(h_k)P(d^t|h_k)}{P^t(d^t)} \qquad (1)$$

where $P^t\left(d^t\right)$ is the marginal probability of d^t at time *t* and is defined as follows.

$$P^t\left(d^t\right)=\sum_k P^t\left(h_k\right)P\left(d^t \mid h_k\right) \qquad (2)$$

Next, the posterior probability that resulted from the analysis becomes the new prior estimate in the update.

$$P^{t+1}\left(h_k\right)\leftarrow P^t\left(h_k \mid d^t\right) \qquad (3)$$

By combining formulas (1) and (3), we can get formula (4).

$$P^{t+1}\left(h_k\right)\leftarrow \frac{P^t\left(h_k\right)P\left(d^t \mid h_k\right)}{P^t\left(d^t\right)} \qquad (4)$$

Each time the data are observed, the inference progresses by updating the confidence for each hypothesis using formula (4). Note that in this process, the confidence $P^t\left(h_k\right)$ for each hypothesis changes over time, but the model $P\left(d \mid h_k\right)$ for each hypothesis does not change.

If we focus on the recursiveness of $P^t\left(h_k\right)$, formula (4) can be rewritten as

$$P^{t+1}\left(h_k\right)\leftarrow P^1\left(h_k\right)\prod_{i=1}^{t}\frac{P\left(d^i \mid h_k\right)}{P^i\left(d^i\right)} \qquad (5)$$

Here, the denominator $P^i\left(d^i\right)$ is common to all hypotheses and can be considered as a constant. Therefore, if normalization processing is omitted, formula (5) can be written as follows.

$$P^{t+1}\left(h_k\right)\leftarrow P^1\left(h_k\right)\prod_{i=1}^{t}P\left(d^i \mid h_k\right) \qquad (6)$$

That is, the current confidence for a hypothesis is proportional to the prior probability multiplied by the likelihood of the data observed so far.

Extended Bayesian inference

We proposed the incorporation of such causal induction factor into Bayesian inference in the EBI [14]. In the EBI, first, we defined $C(e|c)$ as a new index representing the strength of the connection between two events *c* and *e* as follows.

$$C(e|c)=\left[(1-\alpha)P(e|c)^m+\alpha P(c|e)^m\right]^{1/m} \quad (7)$$

Similarly, we defined $C(c|e)$ as follows.

$$C(c|e)=\left[(1-\alpha)P(c|e)^m+\alpha P(e|c)^m\right]^{1/m} \quad (8)$$

These formulas represent the generalized weighted averages of $P(e|c)$ and $P(c|e)$. The generalized weighted average of variables *x* and *y* is expressed by the following formula using parameters α and *m*.

$$\mu(x,y|\alpha,m)=\left[(1-\alpha)x^m+\alpha y^m\right]^{1/m} \quad (9)$$

Here, α takes values in the range $0\le\alpha\le1$ and denotes weighting the values of *x* and *y*, while *m* takes values in the range $-\infty\le m\le\infty$ and denotes the manner of taking the mean. For example, suppose α = 0.5 and *m* = 1, then $\mu(x,y|0.5,1)=0.5x+0.5y$, which represents the arithmetic mean. Suppose α = 0.5 and *m* = -1, then $\mu(x,y|0.5,-1)=2xy/(x+y)$ represents the harmonic mean.

Suppose that *m* = 0, then formula (9) is undefinable. If $X\approx0$, the approximation of $\exp(X)\approx1+X$ is established by the Maclaurin expansion. When *x* is finite and $m\approx0$, the approximation $m\log(x)\approx0$ holds true; thus, the approximation $x^m=\exp(\log(x^m))=\exp(m\log(x))\approx1+m\log(x)$ is derived. Similarly, when *y* is finite and $m\approx0$, $y^m\approx1+m\log(y)$ is derived. Therefore, formula (9) can be rewritten as follows.

$$
\begin{aligned}
\mu(x,y\,|\,\alpha,m) &= \left[(1-\alpha)x^m + \alpha y^m\right]^{1/m} \\
&\approx \left[(1-\alpha)(1+m\log x) + \alpha(1+m\log y)\right]^{1/m} \\
&= \left[1+m\log x^{1-\alpha} + m\log y^{\alpha}\right]^{1/m} \\
&= \left[1+m\log x^{1-\alpha}y^{\alpha}\right]^{1/m} \\
&= \left[1+\log\left(x^{1-\alpha}y^{\alpha}\right)^m\right]^{1/m} \\
&\approx \left[\exp\left(\log\left(x^{1-\alpha}y^{\alpha}\right)^m\right)\right]^{1/m} \\
&= \exp\left(\log\left(x^{1-\alpha}y^{\alpha}\right)\right) \\
&= x^{1-\alpha}y^{\alpha}
\end{aligned}
\tag{10}
$$

Therefore, if we define the mean value $\mu(x,y\,|\,\alpha,0)$ as the limit of $m \to 0$, we get $\mu(x,y\,|\,\alpha,0) = x^{1-\alpha}y^{\alpha}$, where if α = 0.5, it denotes the geometric mean $\mu(x,y\,|\,0.5,0) = \sqrt{xy}$. That is, if *m* = 0 and α = 0.5, $C(e\,|\,c) = DFH(e\,|\,c) = \sqrt{P(e\,|\,c)P(c\,|\,e)}$. In contrast, if α = 0, $C(e\,|\,c) = P(e\,|\,c)$, irrespective of the value of *m*. In other words, by introducing parameters *m* and α, the conditional reasoning $P(e\,|\,c)$ and causal reasoning $DFH(e\,|\,c)$ can be seamlessly connected.

Here, we discuss the meaning of *m*. For simplicity, let $\alpha = 0.5,\ x+y=1\ (0 \le x \le 1)$.

$$
\mu(x\,|\,0.5,m) = \left[0.5x^m + 0.5(1-x)^m\right]^{1/m} \tag{11}
$$

When $m=1$, $\mu(x\,|\,0.5,1) = 0.5$, irrespective of the value of *x*. When $m>1$, $\mu(x\,|\,0.5,m)$ is a convex function and the values at both ends, $\mu(0\,|\,0.5,m)$ and $\mu(1\,|\,0.5,m)$, approach 1 as *m* increases.

However, when $.0<m<1.$, $\mu(x\,|\,0.5,m)$ becomes a concave function and as *m* approaches 0, the values $\mu(0\,|\,0.5,m)$ and $\mu(1\,|\,0.5,m)$ at both ends approach 0.

When $m<0$, the values at both ends cannot be defined, but as *x* approaches 0 or 1, $\mu(x\,|\,0.5,m)$ approaches 0. In other words, in the range of $m \le 0$, if either *x* or *y* approaches 0, their mean also

approaches 0.

Next, using Bayes' theorem, formulas (7) and (8) can be transformed as follows [14]:

$$
\begin{aligned}
C(e|c) &= \left[(1-\alpha)P(e|c)^m + \alpha P(c|e)^m\right]^{1/m} \\
&= \left[(1-\alpha)\left(\frac{P(c)P(e|c)}{P(c)}\right)^m + \alpha\left(\frac{P(e)P(c|e)}{P(e)}\right)^m\right]^{1/m} \\
&= \left[(1-\alpha)\left(\frac{1}{P(c)}\right)^m + \alpha\left(\frac{1}{P(e)}\right)^m\right]^{1/m} P(c)P(e|c) \qquad (12) \\
&= \left[(1-\alpha)P(c)^{-m} + \alpha P(e)^{-m}\right]^{1/m} P(c)P(e|c) \\
&= \frac{P(c)P(e|c)}{\left[(1-\alpha)\left(P(c)\right)^{-m} + \alpha\left(P(e)\right)^{-m}\right]^{-1/m}}
\end{aligned}
$$

$$
\begin{aligned}
C(c|e) &= \left[(1-\alpha)P(c|e)^m + \alpha P(e|c)^m\right]^{1/m} \\
&= \frac{P(c)P(e|c)}{\left[(1-\alpha)P(e)^{-m} + \alpha P(c)^{-m}\right]^{-1/m}} \qquad (13)
\end{aligned}
$$

Note that $P(c)P(e|c) = P(e)P(c|e) = P(c,e)$.

Here, if we describe formulas (12) and (13) recursively, and replace c and e with h_k and d^t, respectively, we can get the next formulas.

$$
C^{t+1}\left(d^t \mid h_k\right) \leftarrow \frac{C^t\left(h_k\right)C^t\left(d^t \mid h_k\right)}{\left[(1-\alpha)C^t\left(h_k\right)^{-m} + \alpha C^t\left(d^t\right)^{-m}\right]^{-1/m}} \qquad (14)
$$

$$
C^{t+1}\left(h_k \mid d^t\right) \leftarrow \frac{C^t\left(h_k\right)C^t\left(d^t \mid h_k\right)}{\left[(1-\alpha)C^t\left(d^t\right)^{-m} + \alpha C^t\left(h_k\right)^{-m}\right]^{-1/m}} \qquad (15)
$$

$$
C^t\left(d^t\right) = \sum_k C^t\left(h_k\right)C^t\left(d^t \mid h_k\right) \qquad (16)
$$

Formula (15) can be rewritten as follows by using a Bayesian update.

$$C^{t+1}\left(h_k\right) \leftarrow \frac{C^t\left(h_k\right)C^t\left(d^t \mid h_k\right)}{\left[\left(1-\alpha\right)C^t\left(d^t\right)^{-m} + \alpha C^t\left(h_k\right)^{-m}\right]^{-1/m}} \quad (17)$$

In formula (17), a description of the normalization process for setting the confidence as a probability is omitted.

Assuming α = 0 in formula (17), the same form as formula (4) for Bayesian inference is obtained.

$$C^{t+1}\left(h_k\right) \leftarrow \frac{C^t\left(h_k\right)C^t\left(d^t \mid h_k\right)}{C^t\left(d^t\right)} \qquad (18)$$

If α = 0, $C^t\left(d^t \mid h_k\right)$ does not change by formula (14), as shown below.

$$C^{t+1}\left(d^t \mid h_k\right) \leftarrow C^t\left(d^t \mid h_k\right) \qquad (19)$$

In other words, if α = 0, then formula (14) substantially disappears and the EBI becomes the same as Bayesian inference. In contrast, in the case of α > 0, the likelihood is modified by formula (14). In this study, we only update the likelihood of the hypothesis with the highest confidence at that time instead of updating the likelihood of all hypotheses. That is, the following formula is used instead of formula (14).

$$C^{t+1}\left(d^t \mid h_k\right) \leftarrow \begin{cases} \dfrac{C^t\left(h_k\right)C^t\left(d^t \mid h_k\right)}{\left[\left(1-\alpha\right)C^t\left(h_k\right)^{-m} + \alpha C^t\left(d^t\right)^{-m}\right]^{-1/m}} & \textit{if } h_k = \arg\max_{h_i} C^t\left(h_i\right) \\ C^t\left(d^t \mid h_k\right) & \textit{otherwise} \end{cases} \qquad (20)$$

Hereafter, the hypothesis with the highest confidence at time t is denoted by $h^t_{\max}$. If there are multiple hypotheses with the highest confidence, one of them is selected at random.

In the following, let us analyze the case of m = 0, that is, the geometric mean case. In the case of m = 0, formula (17) can be transformed as follows.

$$C^{t+1}\left(h_k\right) \leftarrow \frac{C^t\left(h_k\right)C^t\left(d^t \mid h_k\right)}{C^t\left(d^t\right)^{1-\alpha}C^t\left(h_k\right)^{\alpha}} = \left[\frac{C^t\left(h_k\right)}{C^t\left(d^t\right)}\right]^{1-\alpha} C^t\left(d^t \mid h_k\right) \quad (21)$$

If we focus on the recursiveness of $C^t\left(h_k\right)$, formula (21) can be rewritten as follows.

$$C^{t+1}\left(h_k\right) \leftarrow \left[C^1\left(h_k\right)\right]^{(1-\alpha)^t} \prod_{i=1}^{t} \frac{\left[C^i\left(d^i \mid h_k\right)\right]^{(1-\alpha)^{t-i}}}{\left[C^i\left(d^i\right)\right]^{(1-\alpha)^{t+1-i}}} \qquad (22)$$

Here, the denominator $C^i\left(d^i\right)$ is common to all hypotheses and can be considered as a constant. Therefore, if the normalization processing is omitted, formula (20) can be written as follows.

$$C^{t+1}\left(h_k\right) \leftarrow \left[C^1\left(h_k\right)\right]^{(1-\alpha)^t} \prod_{i=1}^{t} \left[C^i\left(d^i \mid h_k\right)\right]^{(1-\alpha)^{t-i}} \quad (23)$$

This can be understood as indication that the current confidence for a hypothesis is proportional to its prior probability multiplied by the likelihood designed to weaken the weight of the distant past. In the case of α = 0, that is, Bayesian inference, $C^{t+1}\left(h_k\right) \leftarrow C^1\left(h_k\right) C^1\left(d^1 \mid h_k\right) C^2\left(d^2 \mid h_k\right) \cdots C^t\left(d^t \mid h_k\right)$. This means that the current likelihood and the past likelihood are weighted equally.

However, in the case of α = 1, $C^{t+1}\left(h_k\right) \leftarrow C^t\left(d^t \mid h_k\right)$. It means that the confidence is calculated using only the current likelihood. Thus, it can be said that the EBI introduces the effect of forgetting into Bayesian inference when considering past history.

In the case of $m = 0$, with respect to $h_{\max}^t$, formula (20) can be written as follows.

$$C^{t+1}\left(d^t \mid h_{\max}^t\right) \leftarrow \frac{C^t\left(h_{\max}^t\right) C^t\left(d^t \mid h_{\max}^t\right)}{C^t\left(h_{\max}^t\right)^{1-\alpha} C^t\left(d^t\right)^{\alpha}} = \left[\frac{C^t\left(h_{\max}^t\right)}{C^t\left(d^t\right)}\right]^{\alpha} C^t\left(d^t \mid h_{\max}^t\right) \qquad (24)$$

In the case of $C^t\left(h_{\max}^t\right) > C^t\left(d^t\right)$, the likelihood becomes larger, and in the case of $C^t\left(h_{\max}^t\right) < C^t\left(d^t\right)$, the likelihood becomes smaller. This means that the model is corrected based on the observed data. Thus, it can be said that the EBI introduces the effect of learning into the Bayesian inference.

Testing a normal distribution model

Mean value estimation using normal distribution

In this study, we deal with one-dimensional continuous probability distribution, such as one-dimensional normal distribution as a concrete model for the hypothesis. The model of hypothesis *k* at time *t* is denoted by $F\left(d \mid \theta_k^t\right)$. Here, θ_k^t represents the parameter of the model. In the case of normal distribution, $F\left(d \mid \theta_k^t\right) = N\left(d \mid \theta_k^t\right)$, $\theta_k^t = \left(\mu_k^t, \Sigma_k^t\right)$, where μ and Σ represent the mean and variance, respectively. This section describes the mean value estimation. The variance estimation will be described in the next section. When a normal distribution is used as a model, $C^t\left(d \mid h_k\right)$ and $C^t\left(d\right)$ are probability densities, while $C^t\left(h_k\right)$ is a probability because the number of hypotheses is discrete and finite. Thus, when calculating formulas (23) and (24), a positive number Δ is introduced and approximately calculated as follows.

$$C^{t+1}\left(h_k\right) \leftarrow \left[C^t\left(h_k\right)\right]^{1-\alpha} \Delta F\left(d^t \mid \theta_k^t\right) \qquad (25)$$

$$\begin{aligned} C^{t+1}\left(d^t \mid h_{\max}^t\right) &\leftarrow \frac{1}{\Delta}\left[\frac{C^t\left(h_{\max}^t\right)}{\sum_k C^t\left(h_k\right)\Delta F\left(d^t \mid \theta_k^t\right)}\right]^{\alpha} \Delta F\left(d^t \mid \theta_{\max}^t\right) \\ &= \frac{1}{\Delta^{\alpha}}\left[\frac{C^t\left(h_{\max}^t\right)}{\sum_k C^t\left(h_k\right) F\left(d^t \mid \theta_k^t\right)}\right]^{\alpha} F\left(d^t \mid \theta_{\max}^t\right) \end{aligned} \qquad (26)$$

Here, $\theta_{\max}^t$ represents the parameter of the distribution that is the model for $h_{\max}^t$.

In formula (25), the term Δ is common to all hypotheses and can be canceled by normalization. Thus, if normalization processing is omitted, it can be expressed as follows.

$$C^{t+1}\left(h_k\right) \leftarrow \left[C^t\left(h_k\right)\right]^{1-\alpha} F\left(d^t \mid \theta_k^t\right) \qquad (27)$$

In formula (27), if the confidence for a hypothesis becomes zero once, it remains zero thereafter. To prevent this, normalization processing (smoothing) is performed by adding a small positive constant ε to the confidence of each hypothesis obtained by formula (27).

$$C^{t+1}\left(h_k\right) \leftarrow \frac{C^{t+1}\left(h_k\right)+\varepsilon}{\sum_{j=1}^{K}\left[C^{t+1}\left(h_j\right)+\varepsilon\right]} = \frac{C^{t+1}\left(h_k\right)+\varepsilon}{K\varepsilon+\sum_{j=1}^{K}C^{t+1}\left(h_j\right)} \qquad (28)$$

Here, K represents the total number of hypotheses. In this study, we set $\varepsilon = 10^{-10}$.

Having observed the data d^t, the likelihood is changed to $C^{t+1}\left(d^t \mid h_{\max}^t\right)$ by formula (26). Concomitantly, the parameter of the model for the hypothesis is modified from $\theta_{\max}^t$ to $\theta_{\max}^{t+1}$ so that the following equation is satisfied.

$$C^{t+1}\left(d^t \mid h_{\max}^t\right) = F\left(d^t \mid \theta_{\max}^{t+1}\right) \qquad (29)$$

If F is a normal distribution, equation (29) can be described as follows.

$$C^{t+1}\left(d^t \mid h_{\max}^t\right) = F\left(d^t \mid \theta_{\max}^{t+1}\right) = \frac{1}{\sqrt{2\pi\Sigma_{\max}^{t+1}}}\exp\left[-\frac{\left(d^t-\mu_{\max}^{t+1}\right)^2}{2\Sigma_{\max}^{t+1}}\right] \qquad (30)$$

Updating the variance from $\Sigma_{\max}^t$ to $\Sigma_{\max}^{t+1}$ is described in the next section.

Solving formula (30) for $\mu_{\max}^{t+1}$ leads to the following two solutions.

$$\begin{aligned} \mu_1 &= d^t + \sqrt{-2\Sigma^t \log\left[C^{t+1}\left(d^t \mid h_{\max}^t\right)\sqrt{2\pi\Sigma_{\max}^{t+1}}\right]} \\ \mu_2 &= d^t - \sqrt{-2\Sigma^t \log\left[C^{t+1}\left(d^t \mid h_{\max}^t\right)\sqrt{2\pi\Sigma_{\max}^{t+1}}\right]} \end{aligned} \qquad (31)$$

$\mu_{\max}^t$ reflects the past observed data. We determine $\mu_{\max}^{t+1}$ as the one closer to $\mu_{\max}^t$, among the two solutions μ_1 and μ_2 to account for the past data as much as possible.

$$\mu_{\max}^{t+1} = \begin{cases} \mu_1 & if \quad \left|\mu_1-\mu_{\max}^t\right| \le \left|\mu_2-\mu_{\max}^t\right| \\ \mu_2 & otherwise \end{cases} \qquad (32)$$

However, to solve formula (29), $C^{t+1}\left(d^t \mid h_{\max}^t\right)$ needs to be within the range of $0 < C^{t+1}\left(d^t \mid h_{\max}^t\right) \le \frac{1}{\sqrt{2\pi\Sigma_{\max}^{t+1}}}$. Thus, we set the following restrictions after calculating $C^{t+1}\left(d^t \mid h_{\max}^t\right)$ using formula (26).

$$C^{t+1}\left(d^t \mid h_{\max}^t\right) \leftarrow \min\left(\max\left(C^{t+1}\left(d^t \mid h_{\max}^t\right), \varepsilon\right), \frac{1}{\sqrt{2\pi\Sigma_{\max}^{t+1}}}\right) \quad (33)$$

where $\max(x, y)$ represents the larger of *x* and *y*. Conversely, $\min(x, y)$ represents the smaller of *x* and *y*. We set $\varepsilon = 10^{-10}$.

If *K* = 1, there is no other hypothesis. Therefore, the only hypothesis always becomes $h_{\max}^t$ and the value of confidence is always 1. Consider the situation $C^t\left(h_{\max}^t\right) \approx 1$, including the case of *K* = 1. In this case, the confidence of hypotheses other than $h_{\max}^t$ becomes almost 0 by the constraint $\sum_k C^t\left(h_k\right) = 1$; thus, $C^t\left(d^t\right) = \sum_k C^t\left(h_k\right) C^t\left(d^t \mid h_k\right) \approx C^t\left(d^t \mid h_{\max}^t\right)$ is derived from formula (16). Therefore, formula (26) can be transformed as follows.

$$\begin{aligned} C^{t+1}\left(d^t \mid h_{\max}^t\right) &\leftarrow \frac{1}{\Delta^\alpha}\left[\frac{C^t\left(h_{\max}^t\right)}{\sum_k C^t\left(h_k\right) N\left(d^t \mid \mu_k^t, \Sigma_k^t\right)}\right]^\alpha N\left(d^t \mid \mu_{\max}^t, \Sigma_{\max}^t\right) \\ &\approx \left(\frac{1}{\Delta}\right)^\alpha \left[C^t\left(d^t \mid h_{\max}^t\right)\right]^{1-\alpha} \end{aligned} \quad (34)$$

If formula (34) is denoted by $x^{t+1} = f\left(x^t\right) = \left(\frac{1}{\Delta}\right)^\alpha \left(x^t\right)^{1-\alpha}$, $f\left(x^t\right)$ becomes a concave function. Solving $x^t = f\left(x^t\right)$ results in $x^t = 0, \frac{1}{\Delta}$. The fixed point $\left(x^t, f\left(x^t\right)\right) = \left(\frac{1}{\Delta}, \frac{1}{\Delta}\right)$ is a stable point because $x^t \geq f\left(x^t\right)$ when $x^t > \frac{1}{\Delta}$ and $x^t \leq f\left(x^t\right)$ when $x^t < \frac{1}{\Delta}$. In this study, we set $\Delta = \sqrt{2\pi\Sigma_{\max}^t}$. In this case, $C^{t+1}\left(d^t \mid h_{\max}^t\right)$ approaches the vertex of the normal distribution whenever data d^t are observed.

As shown in formula (29), $\mu_{\max}^{t+1}$ is determined to satisfy the condition $C^{t+1}\left(d^t \mid h_{\max}^t\right) = N\left(d^t \mid \mu_{\max}^{t+1}, \Sigma_{\max}^{t+1}\right)$. This means that $\mu_{\max}^{t+1}$ approaches the observation data d^t. Through the processing described above, the confidences for each hypothesis and the model for the hypothesis with maximum confidence are corrected

whenever the data are observed.

We will hereinafter refer to the latter process of modifying the model for $h^t_{\max}$ as inverse Bayesian inference [21–24]. If the former process of updating the confidences for hypotheses is referred to as inference, inverse Bayesian inference can be called "learning" because it forms a model for a hypothetical instead of an inference. Thus, although the two α s in formulas (26) and (27) are denoted by the same α, they can be called the "learning rate" and "forgetting rate," respectively. We can also set the "learning rate" and "forgetting rate" as two independent parameters. However, when dealing with temporal alteration, like in this study, good performance is achieved when the two parameters have almost identical values. On the contrary, it is preferable to set the parameters separately in spatial clustering.

Variance estimation using gamma distribution

Consider a random variable D that is the sum of n squares of data sampled from a normal distribution $N(0,\Sigma)$ with mean 0 and variance Σ.

$$d^i \sim N(0,\Sigma), \quad D=\sum_{i=1}^{n}\left(d^i\right)^2 \qquad (35)$$

In this case, D follows the gamma distribution with shape parameter $S=n/2$ and scale parameter $\lambda=1/(2\Sigma)$, as shown below.

$$f(D\mid\lambda,S)=\frac{\lambda^S}{\Gamma(S)}D^{S-1}\exp(-\lambda D)=\frac{1}{\Gamma(n/2)(2\Sigma)^{n/2}}D^{n/2-1}\exp(-D/2\Sigma) \qquad (36)$$

The mean of this distribution is $S/\lambda=n\Sigma$. We set n=20, that is S=10. We use the gamma distribution as a model for estimating variance from observed data d^t.

First, the following D_k^t is calculated using the mean estimated value μ_k^t obtained in the previous section. This is used as input data for variance estimation instead of the observation data d^t.

$$\begin{aligned} &T_k\left(n_k \% n+1\right)=t, \quad n_k \leftarrow n_k+1, \qquad if\ h_k^t=h_{\max}^t \\ &D_k^t=\sum_{i=1}^{n}\left(d^{T_k(i)}-\mu_k^{T_k(i)}\right)^2 \end{aligned} \tag{37}$$

Here, n_k represents the number of times that hypothesis k has been the hypothesis with the highest confidence. The initial value of n_k is 0. x % y represents the remainder when integer x is divided by integer y.

The model of hypothesis k at time t is $F\left(D \mid \theta_k^t\right)=f\left(D \mid \theta_k^t\right), \quad \theta_k^t=\left(\lambda_k^t, S\right)$. In this case, formula (29) is rewritten as

$$C^{t+1}\left(D_{\max}^t \mid h_{\max}^t\right)=F\left(D_{\max}^t \mid \theta_{\max}^{t+1}\right)=f\left(D_{\max}^t \mid \lambda_{\max}^{t+1}, S\right)=\frac{\left(\lambda_{\max}^{t+1}\right)^S}{\Gamma(S)}\left(D_{\max}^t\right)^{S-1} \exp\left(-\lambda_{\max}^{t+1} D_{\max}^t\right) \tag{38}$$

With $\Theta_{\max}^{t+1}=-\lambda_{\max}^{t+1} D_{\max}^t / S, \quad Z^{t+1}=-\left(D_{\max}^t C^{t+1}\left(D_{\max}^t \mid h_{\max}^t\right) \Gamma(S)\right)^{1/S} / S$, this equation can be rewritten as

$$Z^{t+1}=\Theta_{\max}^{t+1} \exp\left(\Theta_{\max}^{t+1}\right) \tag{39}$$

For this equation to have a solution with respect to $\Theta_{\max}^{t+1}$ in the range of $\Theta_{\max}^{t+1}<0$, $-1/e \leq Z^{t+1}<0$ must be satisfied. Therefore, the following restrictions are provided.

$$Z^{t+1} \leftarrow \max\left(\min\left(Z^{t+1},-\varepsilon\right),-1/e\right) \tag{40}$$

We set $\varepsilon=10^{-10}$.

When formula (39) is solved for $\Theta_{\max}^{t+1}$, the following two solutions are obtained.

$$\begin{aligned} \Theta_1 &= W_{-1}\left(Z^{t+1}\right) \\ \Theta_2 &= W_0\left(Z^{t+1}\right) \end{aligned} \tag{41}$$

Here, W_{-1} and W_0 are two Lambert W functions that satisfy $Z=xe^x \Leftrightarrow x=W(Z)$.

Similar to the case of estimating the mean value, $\Theta_{\max}^{t+1}$ is determined as

$$\Theta_{\max}^{t+1} = \begin{cases} \Theta_1 & if \quad \left|\Theta_1 - \Theta_{\max}^t\right| \le \left|\Theta_2 - \Theta_{\max}^t\right| \\ \Theta_2 & otherwise \end{cases} \quad (42)$$

Because $\Theta_{\max}^{t+1} = -\lambda_{\max}^{t+1} D_{\max}^t / S$, the scale parameter $\lambda_{\max}^{t+1}$ can be calculated as

$$\lambda_{\max}^{t+1} = -\frac{S\Theta_{\max}^{t+1}}{D_{\max}^t} \quad (43)$$

Further, because $\lambda_{\max}^{t+1} = 1 \big/ \left(2\Sigma_{\max}^{t+1}\right)$, the variance estimate $\Sigma_{\max}^{t+1}$ is calculated as

$$\Sigma_{\max}^{t+1} = \frac{1}{2\lambda_{\max}^{t+1}} \quad (44)$$

We use $\Sigma_{\max}^{t+1}$ as the variance estimate for the next time in generation distribution. That is, $\theta_{\max}^{t+1} = \left(\mu_{\max}^{t+1}, \Sigma_{\max}^{t+1}\right)$.

Regarding the value of Δ, we consider the gamma distribution in which the current input value $D_{\max}^t$ is the mean value S/λ, and define the output value for the input value $D_{\max}^t$ as $1/\Delta$, based on the same arguments as in the case of normal distribution.

$$\Delta = \frac{1}{f\left(D_{\max}^t \mid S/D_{\max}^t, S\right)} = \frac{\Gamma(S) D_{\max}^t \exp(S)}{S^S} \quad (45)$$

The group of processes described in this section and the previous section is summarized as an algorithm below.

1. Set values for parameters α, *m*, ε, *K*.
2. Establish initial values for $\theta_k^1 = \left(\mu_k^1, \Sigma_k^1\right), C^1\left(h_k\right) \quad (k = 1, 2, \cdots K)$.
3. Repeat the following whenever data d^t are observed.
    - Update the confidence $C^{t+1}\left(h_k\right)$ of each hypothesis using formulas (27) and (28).

- Find the hypothesis $h_{\max}^{t}$ with the maximum confidence.
- Create the input data $D_{\max}^{t}$ for variance calculation using formula (37).
- Update the likelihood $C^{t+1}\left(D_{\max}^{t} \mid h_{\max}^{t}\right)$ of the hypothesis $h_{\max}^{t}$ for the input data $D_{\max}^{t}$ using formula (26).
- Correct the variance $\Sigma_{\max}^{t+1}$ of the model for the hypothesis $h_{\max}^{t}$ using formulas (41), (42), (43), and (44) to match the new likelihood $C^{t+1}\left(D_{\max}^{t} \mid h_{\max}^{t}\right)$.
- Update the likelihood $C^{t+1}\left(d^{t} \mid h_{\max}^{t}\right)$ of the hypothesis $h_{\max}^{t}$ for the observed data d^{t} using formula (26).
- Correct the mean $\mu_{\max}^{t+1}$ of the model for the hypothesis $h_{\max}^{t}$ using formulas (31) and (32) to match the new likelihood $C^{t+1}\left(d^{t} \mid h_{\max}^{t}\right)$.
- $\theta_{\max}^{t+1} = \left(\mu_{\max}^{t+1}, \Sigma_{\max}^{t+1}\right)$ is set as the new parameter of the model for the hypothesis $h_{\max}^{t}$.

Sequential Discounting Expectation-Maximization Algorithm (SDEM)

This section describes SDEM, an online EM algorithm proposed by Yamanishi et al. [7]. In SDEM, the E and M steps are executed once for each data observed sequentially. First, in step E, the responsibility is calculated. The responsibility for the normal distribution *k* of the data d^{t} is calculated as follows.

$$q_{k}^{t} = \frac{\pi_{k}^{t} N\left(d^{t} \mid \mu_{k}^{t}, \Sigma_{k}^{t}\right)}{\sum_{j=1}^{K} \pi_{j}^{t} N\left(d^{t} \mid \mu_{j}^{t}, \Sigma_{j}^{t}\right)} \qquad (46)$$

Here, $\sum_{k=1}^{K} \pi_{k}^{t} = 1$ is assumed. π_{k} is called the “mixing weights” and represents the weight of each normal distribution.

Next, in the M step, the mixing weights, means, and variances of each normal distribution are updated.

However, weighting is performed to weaken the influence of older observation data by introducing the discounting rate β(0<β<1).

$$\tilde{\pi}_k^{t+1} \leftarrow (1-\beta)\tilde{\pi}_k^t + \beta q_k^t = \beta \sum_{i=1}^{t} (1-\beta)^{t-i} q_k^i \quad (47)$$

$$\tilde{\mu}_k^{t+1} \leftarrow (1-\beta)\tilde{\mu}_k^t + \beta q_k^t d^t = \beta \sum_{i=1}^{t} (1-\beta)^{t-i} q_k^t d^i \quad (48)$$

$$\tilde{\Sigma}_k^{t+1} \leftarrow (1-\beta)\tilde{\Sigma}_k^t + \beta q_k^t \left(d^t - \mu_k^t\right)\left(d^t - \mu_k^t\right) = \beta \sum_{i=1}^{t} (1-\beta)^{t-i} q_k^t \left(d^i - \mu_k^t\right)\left(d^i - \mu_k^t\right) \quad (49)$$

$$\mu_k^{t+1} \leftarrow \frac{\tilde{\mu}_k^{t+1}}{\tilde{\pi}_k^{t+1}} \quad (50)$$

$$\Sigma_k^{t+1} \leftarrow \frac{\tilde{\Sigma}_k^{t+1}}{\tilde{\pi}_k^{t+1}} \quad (51)$$

Regarding $\tilde{\pi}_k^{t+1}$, smoothing is performed to prevent it from becoming 0 and normalize, similar to the EBI.

$$\pi_k^{t+1} \leftarrow \frac{\tilde{\pi}_k^{t+1} + \gamma}{K\gamma + \sum_{j=1}^{K} \tilde{\pi}_j^{t+1}} \quad (52)$$

We set γ = 0.001 for optimal performance.

In the case of K=1, q_1^t and $\tilde{\pi}_1^t$ are always 1. At this time, formula (50) shows that the new estimated value is obtained as a convex combination of the current estimated value and the current observed data.

$$\mu_k^{t+1} \leftarrow (1-\beta)\mu_k^t + \beta d^t \quad (53)$$

This represents the EMA.

By setting $\pi_k^{t+1} = P^{t+1}(h_k)$, $q_k^i = P^i(h_k \mid d^i)$ in formulas (47) and (52), $P^{t+1}(h_k)$ can be described approximately as follows.

$$P^{t+1}(h_k) \leftarrow \beta \sum_{i=1}^{t} (1-\beta)^{t-i} P^i(h_k \mid d^i) \propto \sum_{i=1}^{t} (1-\beta)^{t-i} P^i(h_k \mid d^i) \quad (54)$$

Taking the logarithm of both sides of formula (23) in the EBI, the following transformation can be made.

$$\log\left(C^{t+1}\left(h_k\right)\right) \leftarrow \sum_{i=1}^{t}(1-\alpha)^{t-i}\log\left(C^{i}\left(d^{i} \mid h_k\right)\right)+(1-\alpha)^{t}\log\left(C^{1}\left(h_k\right)\right) \qquad (55)$$

Comparing formula (55) with formula (54), the EBI differs in that it takes a logarithm and uses likelihood instead of posterior probability.

The group of processes described above is summarized as an algorithm below.

1. Set values for parameters β, γ, K.
2. Establish initial values for the variables $\theta_k^1 = \left(\mu_k^1, \Sigma_k^1\right), \pi_k^1 \quad (k = 1, 2, \cdots K)$.
3. Repeat the following whenever data d^t are observed.
   - E step: Calculate the responsibility q_k^t for each normal distribution k of the observed data d^t using formula (46).
   - M step: Update the mixing weights π_k^{t+1} using formulas (47) and (52).
   - M step: Correct the mean μ_k^{t+1} and variance Σ_k^{t+1} of the normal distribution using formulas (48), (49), (50), and (51).
   - $\theta_k^{t+1} = \left(\mu_k^{t+1}, \Sigma_k^{t+1}\right)$ is set as the new parameter of the model for each hypothesis h_k^t.

Simulation

To investigate the behavior of EBI, a simulation was performed. In the simulation, one random number d^t is generated at each time from a certain normal distribution (the “generation distribution”). Then, the EBI estimates the generation distribution by observing d^t.

In this study, we deal with a task in which the mean and variance of the generation distribution fluctuate randomly at each regular interval. Specifically, every 1000 steps, a random number from a uniform

distribution of the range [0, 5] is generated, and the number is set as a new mean of distribution. Similarly, a random number from a uniform distribution of the range [0, 0.1] is generated, and the number is set as a new variance of the distribution.

$$\mu_{correct}^{t+1} = \begin{cases} 5 \times rnd^t & if\ t\,\%\,1000 = 0 \\ \mu_{correct}^t & otherwise \end{cases} \qquad (56)$$

$$\Sigma_{correct}^{t+1} = \begin{cases} 0.1 \times rnd^t & if\ t\,\%\,1000 = 0 \\ \Sigma_{correct}^t & otherwise \end{cases} \qquad (57)$$

Here, $\mu_{correct}^t$ and $\Sigma_{correct}^t$ represent the mean and variance of the normal distribution used as the generation distribution at time *t*, that is, the correct values in this task.

rnd^t represents a random number generated from a continuous uniform distribution of the range [0, 1] at time *t*. Fig. 1 shows an example of time evolution of observation data d^t in this task.

Fig. 1. Time progress of observed values.

In the simulation, estimations were performed via the EBI, SDEM, and EMA for comparison. The parameter estimated by the EBI at time *t* is that of the model for $h_{\max}^t$, that is $\theta_{\max}^t = \left(\mu_{\max}^t, \Sigma_{\max}^t\right)$. The parameter estimated by SDEM is that of the distribution for which the observed data d^t have the highest responsibility among the normal distributions included in the Gaussian mixture distribution, that is, $\theta_{m^t}^t = \left(\mu_{m^t}^t, \Sigma_{m^t}^t\right)$, $m^t = \arg\max_k q_k^t$.

The mean μ_{EMA}^t and variance Σ_{EMA}^t for the EMA are updated as follows.

$$\mu_{EMA}^{t+1} \leftarrow (1-\beta)\mu_{EMA}^t + \beta d^t \qquad (58)$$

$$\Sigma_{EMA}^{t+1} \leftarrow (1-\beta)\Sigma_{EMA}^t + \beta\left(d^t - \mu_{EMA}^t\right)\left(d^t - \mu_{EMA}^t\right) \qquad (59)$$

Here, β (0<β<1) represents a discount rate.

Results

Fig. 2(a) shows an example of the result by EBI. In this simulation, the initial values of the mean and variance of the model for each hypothesis were set to $\mu_k^1 = 2.5$ and $\Sigma_k^1 = 0.05$, respectively.

Fig. 2 Time progress of the estimated values for the mean of Gaussian. The figures include the correct mean. (a) Estimated values by EBI. $(\alpha, m, K) = (0.018, 0.0, 10)$. (b) Estimated values by inverse Bayesian inference. $(\alpha, m, K) = (0.018, 0.0, 1)$. (c) Estimated values by EMA. $\beta = 0.009, 0.012, 0.15$.

Fig. 2 (a) shows the time evolution of the correct value $\mu_{correct}^t$ and that of the estimated result by the EBI, set to K = 10. Fig. 2(b) shows the result obtained by the EBI set to K = 1 (i.e., the result for inverse Bayesian inference). Fig. 2 (c) shows the estimation results obtained by three types of EMA with different discounting rates β. It is evident that for larger discounting rates, the responses to sudden changes are quicker, as expected, but the fluctuations are increased during the stable period.

In the case of EBI, initially, it takes time to follow up when the correct value suddenly changes. However, there are cases where changes can be handled instantly over time. In contrast, in the cases of inverse Bayesian inference and EMA, the follow-up performances are not improved over time at all.

Fig. 3 (a) shows the time evolution of the means μ_k^t of the models for ten hypotheses used in the simulation of Fig. 2 (a). Fig. 3 (b) shows the time evolution of the hypothesis $h_{\max}^t$ with the maximum confidence. Initially, all hypothesis models are the same, but various hypothesis models are formed by learning over time. Additionally, it is evident that it is possible to follow quickly by appropriately switching

the hypotheses. When the EBI is set to K = 1 and EMA, because the hypotheses cannot be switched, such quick tracking cannot be achieved.

Fig. 3 Internal state of extended Bayesian inference. (a) Time progress of the mean of normal distribution for each hypothesis. (b) Time progress of the hypothesis with the greatest degree of confidence.

To evaluate the estimation performance of each method, the root-mean-square error (RMSE) between the estimated value and correct value is calculated as follows.

$$RMSE = \sqrt{\frac{\sum_{t=j}^{T+j-1}\left(\hat{\mu}^t - \mu_{correct}^t\right)^2}{T}} \quad (60)$$

Here, $\hat{\mu}^t$ and $\mu_{correct}^t$ represent the estimated value and correct value at time t, respectively. T represents a period for evaluation.

Each interval of 1000 steps, from a change in the generation distribution to the next change, is divided into two halves. We use the RMSE of the first half as a measure of the inability to follow rapid changes and that of the second half as a measure of the inaccuracy of the estimation in the stable period.

Fig. 4 shows the relationship between the followability errors and estimation accuracy errors for each method. Note that the x-axis and y-axis in this figure indicate the RMSE; therefore, the closer to the origin, the higher the estimation performance. In Fig. 4 (a) and (b), the simulations for each method were performed in the cases of K = 1, K = 2, K = 5, and K = 10. Each figure also shows the EMA results as a baseline. The simulations were performed by changing the discounting rate of each method from 0.003 to 0.15 in increments of 0.003. These values are obtained by dividing the interval from time 0 to 10000 into intervals of every 1000 steps, calculating the followability errors and accuracy errors in each interval, and averaging them.

Fig. 4 Relationship between the followability errors and estimation accuracy errors. (a) EBI. (b) SDEM.

The values shown in the figure are the average values of 100 trials with different random seeds. This also applies to Fig. 7. The initial values μ_k^1 of the center of each component (normal distribution) were set to random numbers generated from a uniform distribution of the range [0, 5] in both methods. Similarly, the initial values Σ_k^1 of the variance of each normal distribution were set to random numbers generated from a uniform distribution of the range [0, 0.1] in both methods.

For the EMA, it is evident that there is a trade-off, i.e., the accuracy decreases as followability increases. The EBI can modify the trade-off observed in the EMA. In the case of $K = 1$, that is, even if only the inverse Bayesian inference is used, the trade-off can be improved, but the performance is improved as the number of components is increased. SDEM can also modify the trade-off but there is no noticeable difference depending on the number of components.

Fig. 5 shows the time evolution of the mean estimated value of each method where the number of components and discounting rate are selected to achieve the best performance. The correct value is also shown in the figure.

Fig. 5 Time progress of the estimated values for the mean of Gaussian. The figure includes the correct mean. (a) Estimated values by EBI. $(\alpha, m, K) = (0.036, 0.0, 10)$. (b) Estimated values by EMA. $\beta = 0.036$. (c) Estimated values by SDEM. $(\beta, K) = (0.036, 10)$.

Fig. 6 shows the time evolution of the estimated value of variance for each method. The figure also shows the correct value. In the EMA and SDEM, the bursts of estimates at the points where the variances change can be observed.

Fig. 6 Time progress for the estimated values of the variance of the Gaussian. The figure includes the correct variance. (a) Estimated values by EBI. $(\alpha, m, K) = (0.03, 0.0, 10)$. (b) Estimated values by EMA. $\beta = 0.012$. (c) Estimated values by SDEM. $(\beta, K) = (0.012, 10)$.

In the results shown above, the EBI simulations were performed for only $m = 0$. The results of EBI when m is changed are shown below. The values of α and m were shifted, with increments of 0.05 and 0.1 in the interval [0.05, 0.5] and [-2, 2], respectively, and simulations were performed to obtain the total RMSE calculated from the entire interval for each pair of parameters. Fig. 7 shows the total RMSE for each pair of parameter values. In most α regions, the RMSE is low when $m \leq 0$. When m exceeds 0, the RMSE increases.

Fig. 7 RMSE for each pair of parameter values. (a) In all cases. (b) In the case of $\alpha = 0.25$.

Discussion and Conclusions

In general, if a method such as the EMA with the discounting rate is used to improve the followability to a sudden change, it is necessary to increase the discounting rate. This means that in the estimation, the recent data are weighted more extensively. That is, as long as a constant discount rate is used, a trade-off exists; the followability is improved by when the discounting rate is high, but the accuracy is reduced.

In this study, we simulated the task of estimating the distributions for data generation in a non-stationary situation wherein the distributions change suddenly. Consequently, the EBI proposed in this study successfully modified the trade-off observed in the EMA.

In addition, we compared the estimation performance of EBI with that of SDEM. The EBI showed higher estimation performance.

However, as shown in Fig. 7, *m* must be 0 or less to achieve high performance. In the literature [14], we derived α and *m* that best fit the causal strength felt by humans from formula (7) and the eight types of experimental data shown in the literature [15–19]. Accordingly, the values of α were in the range of 0.25 to 0.6. In other words, they were far from α = 0, which implies conditional probability. In contrast, the values of *m* were interestingly in the range of -2.0 to -0.25, that is, were negative in all eight experiments [14]. We did not determine the cause of these negative values in this study. This is a question for further study.

As shown in Fig. 6, some bursts were observed in the variance estimates by the EMA and SDEM. The changes in the mean and variance occurred simultaneously in this simulation. Therefore, the delay in following the changes in the mean may cause confusions between the changes in the mean and those in the variance.

In the EBI, various models for the hypotheses are formed by inverse Bayesian inference, even if appropriate models are not given beforehand. After some models are accumulated, rough inference is performed by switching them and fine adjustment is performed by inverse Bayesian inference, thereby achieving both followability and accuracy. The situations where both learning and inference are performed also exist in daily life. For example, in estimating the emotions of others, one cannot have a complete model for someone else's emotions because "you" are not "them." Assume that someone's facial expression suddenly changed when you estimated that the person feels happy based on your currently incomplete model. Further, assume that it was the first facial expression you saw. At this time, it is possible to think that the person's emotion has changed from joy to another emotion. However, it is also possible to consider that it is a new facial expression representing joy.

It is more difficult to detect a sign of change immediately at a time when a change is occurring than to detect a change point by looking back at the past after a change has occurred and persisted. This is because while detecting a change, the decision must be made in a situation where there is no model regarding the

new stage after the change. That is, in the example above, when the next facial expression model is not completely created. Under such circumstances, while efficiently learning a model from the observed data, there is a need for a technique for making an appropriate decision using the model. In the future, such techniques can be expected to be applied for, for example, the detection of the signs of a disease.

The EBI can be regarded as introducing the effects of forgetting and learning into the Bayesian inference due to the action of exponential smoothing using α. With the introduction of discounting rate α, the influence of much older data is weakened. Simultaneously, α represents the learning rate and the learning process modifies the model for the hypothesis based on the observed data. In this framework, even with the same hypothesis, the content (model) changes over time; therefore, it is not possible to simply accumulate experiences from the past. For this reason, it is reasonable to include the effect of forgetting.

The EBI framework is very similar to the SDEM framework proposed by Yamanishi et al [7]. However, there are some differences. For example, when the history is considered for updating the weight of the Gaussian mixture distribution, there is a difference between the consideration of posterior probability or likelihood, and whether they are accumulated as addition or multiplication. In the future, we would like to clarify the difference in effectiveness due to these differences through the simulation of various tasks.

As limitations, in this simulation, only one-dimensional distribution was handled. In future work, we will extend our model to multidimensional distribution.

Acknowledgments

This research is partially supported by the Center of Innovation Program from the Japan Science and Technology Agency, JST. This work was supported by the JSPS KAKENHI Grant Numbers JP16K01408. We thank Editage [http://www.editage.com] for English language editing.

Competing interests

The authors [N. M. and K. S.] are employed by SoftBank Robotics Group Corp. This does not alter our adherence to PLOS ONE policies on sharing data and materials.

References

1. Dehaene S. Consciousness and the brain: Deciphering how the brain codes our thoughts. New York: Viking; 2014.

2. Chater N, Oaksford M. The probabilistic mind: Prospects for Bayesian cognitive science. England: Oxford University Press; 2008. doi: https://doi.org/10.1093/acprof:oso/9780199216093.003.0001

3. Bishop CM. Pattern recognition and machine learning. New York: Springer; 2011.

4. Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. J Mach Learn Res. 2003;3: 993–1022. doi: http://dx.doi.org/10.1162/jmlr.2003.3.4-5.993.

5. Neal RM, Hinton GE. A view of the EM algorithm that justifies incremental, sparse, and other variants. In: Jordan MI, editor. Learning in Graphical Models. Dordrecht: Kluwer Academic Publishers; 1998. p. 355–368. doi: https://doi.org/10.1007/978-94-011-5014-9_12.

6. Sato M, Ishii S. On-line EM algorithm for the normalized Gaussian network. Neural Comput. 2000;12: 407–432. doi: https://doi.org/10.1162/089976600300015853.

7. Yamanishi K, Takeuchi J, Williams G, Milne P. On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms. Data Min Knowl Disc. 2004;8: 275–300. doi: https://doi.org/10.1023/b:dami.0000023676.72185.7c.

8. Liang P, Klein D. Online EM for unsupervised models. Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics; 2009: 611–619. doi: https://doi.org/10.3115/1620754.1620843.

9. Cappe O, Moulines E. On-line expectation–maximization algorithm for latent data models. J Roy Stat Soc Ser B Stat Methodol. 2009;71: 593–613. doi: https://doi.org/10.1111/j.1467-9868.2009.00698.x.

10. Dias JG, Cortinhal MJ. The SKM algorithm: A k-means algorithm for clustering sequential data. In: Geffner H, Prada R, Machado AI, David N, editors. Advances in Artificial Intelligence – IBERAMIA 2008. IBERAMIA 2008. Lecture Notes in Computer Science. New York: Springer; 2008. p. 2008173–182. doi: https://doi.org/10.1007/978-3-540-88309-8_18.

11. Zhong S. Efficient online spherical k-means clustering. Proceedings of IEEE International Joint Conference on Neural Networks. 2005;5: 3180–3185. doi: https://doi.org/10.1109/ijcnn.2005.1556436.

12. Andrews NO, Fox EA. Recent developments in document clustering. Technical Report TR-07-35 https://vtechworks.lib.vt.edu/handle/10919/19473 (2007).

13. Tehrani AF, Ahrens D. Modified sequential k-means clustering by utilizing response: A case study for fashion products. Expert Syst. 2017;34: e12226. doi: https://doi.org/10.1111/exsy.12226.

14. Shinohara S, Manome N, Suzuki K, Chung U-I, Takahashi T, Gunji P-Y, Nakajima Y, Mitsuyoshi S. Extended Bayesian inference incorporating symmetry bias. 2020;190: 104104. doi: https://doi.org/10.1016/j.biosystems.2020.104104.

15. Buehner MJ, Cheng PW, Clifford D. From covariation to causation: A test of the assumption of causal power. J Exper Psychol Learn Mem Cogn. 2003;29: 1119–1140, doi: http://dx.doi.org/10.1037/0278-7393.29.6.1119.

16. Lober K, Shanks DR. Is causal induction based on causal power? Critique of Cheng (1997). Psychol Rev. 2000;107: 95–212. doi: http://dx.doi.org/10.1037/0033-295X.107.1.195.

17. White PA. Making causal judgments from the proportion of confirming instances: the pCI rule. J Exper Psychol Learn Mem Cogn. 2003;29: 710–727. doi: http://dx.doi.org/10.1037/0278-7393.29.4.710.

18. Waxman S, Kosowski T. Nouns mark category relations: Toddlers' and preschoolers' word-learning biases. Child Dev. 1990;61: 1461–1473. doi: https://doi.org/10.1111/j.1467-8624.1990.tb02875.x.

19. Hattori M, Oaksford M. Adaptive non-interventional heuristics for covariation detection in causal induction: model comparison and rational analysis. Cogn Sci. 2007;31: 765–814. doi: https://doi.org/10.1080/03640210701530755.

20. Evans J, Handley S, Over D. Conditionals and conditional probability. J Exp Psychol Learn Mem Cogn. 2003;29: 321–335. doi: https://doi.org/10.1037/0278-7393.29.2.321.

21. Arecchi F. Phenomenology of consciousness: from apprehension to judgment. Nonlinear Dynamics Psychol Life Sci. 2011;15: 359–375.

22. Gunji Y-P, Shinohara S, Haruna T, Basios V. Inverse Bayesian inference as a key of consciousness featuring a macroscopic quantum logical structure. Biosystems. 2016;152: 44–65. doi: https://doi.org/10.1016/j.biosystems.2016.12.003.

23. Gunji Y-P, Murakami H, Tomaru T, Basios V. Inverse Bayesian inference in swarming behaviour of soldier crabs. Phil Trans Roy Soc A. 2018;376. doi: http://dx.doi.org/10.1098/rsta.2017.0370.

24. Horry Y, Yoshinari A, Nakamoto Y, Gunji Y-P. Modeling of decision-making process for moving straight using inverse Bayesian inference. Biosystems. 2018;163: 70–81. doi: https://doi.org/10.1016/j.biosystems.2017.12.006.

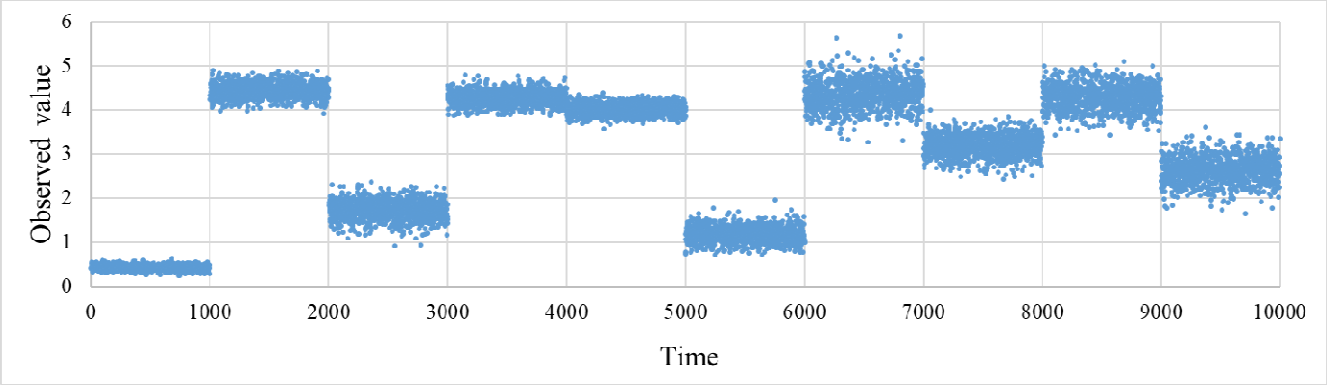
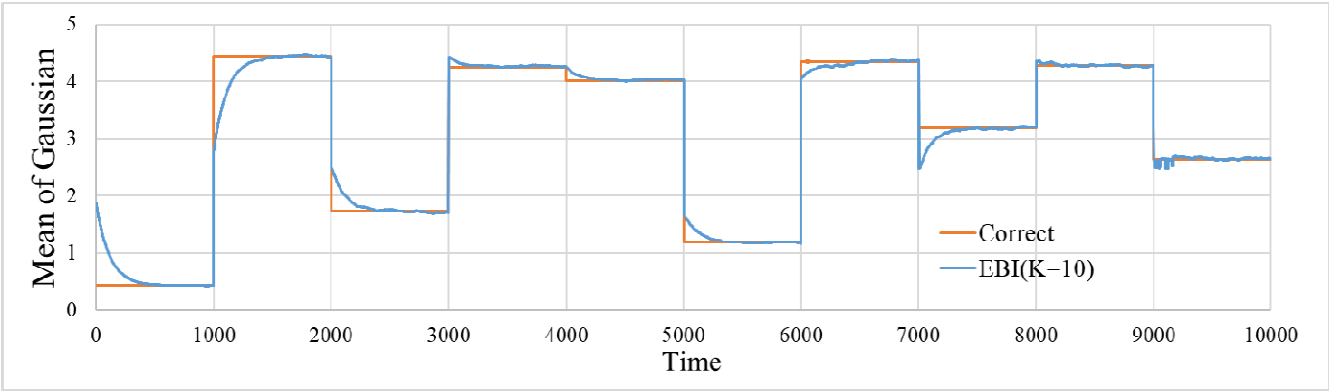
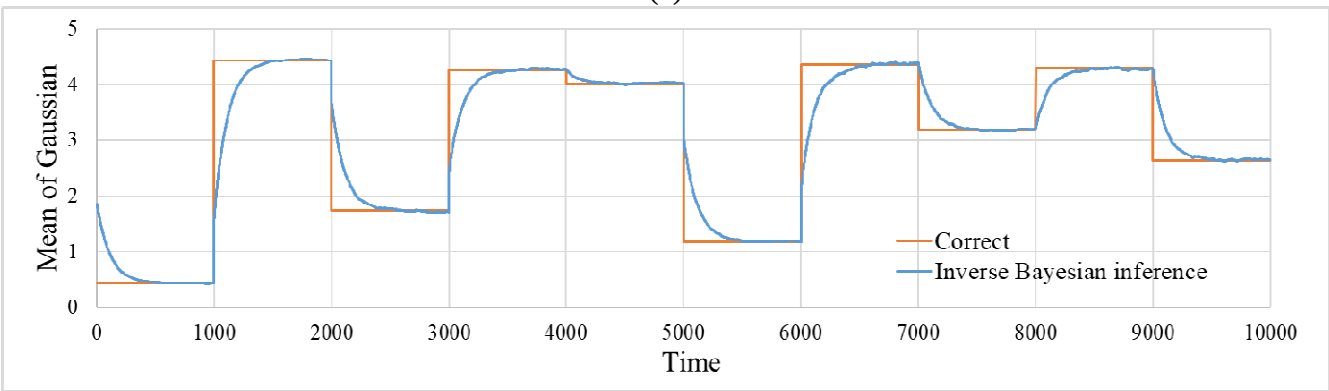


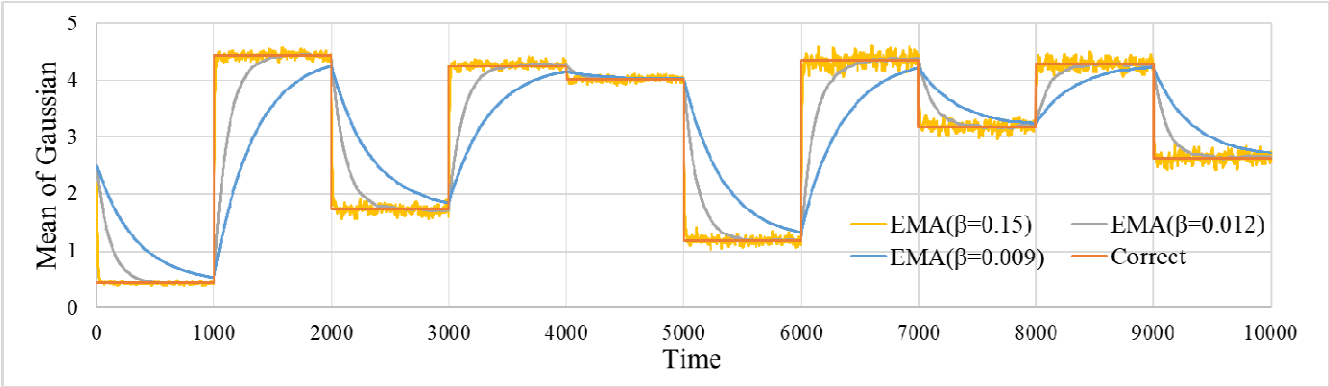
Fig. 1. Time progress of observed values.



(a)

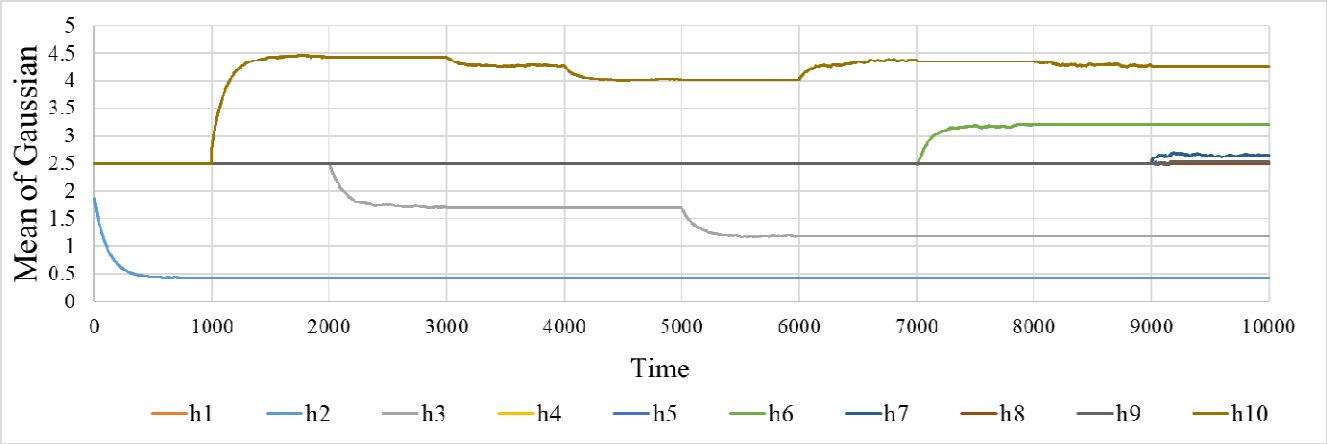


(b)

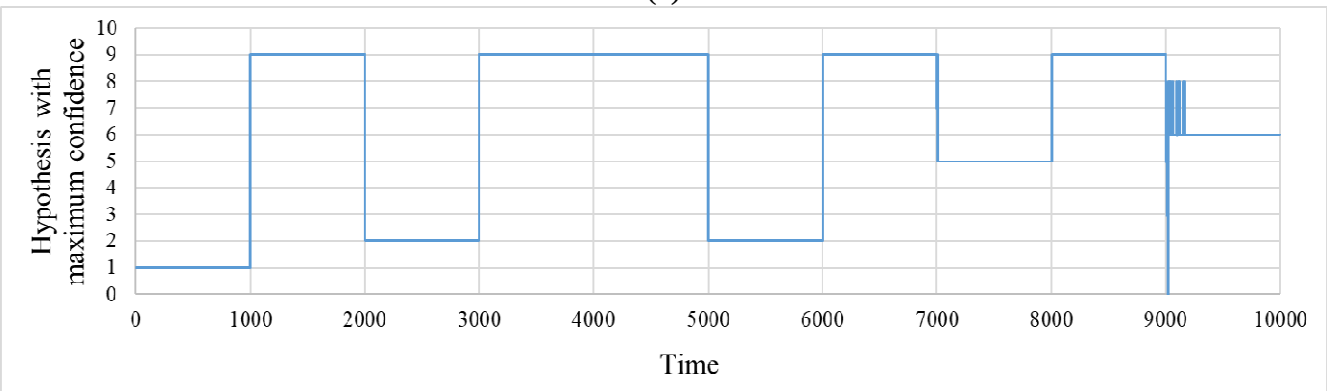


(c)

Fig. 2 Time progress of the estimated values for the mean of Gaussian. The figures include the correct mean. (a) Estimated values by EBI. $(\alpha, m, K) = (0.018, 0.0, 10)$. (b) Estimated values by inverse Bayesian inference. $(\alpha, m, K) = (0.018, 0.0, 1)$. (c) Estimated values by EMA. $\beta = 0.009, 0.012, 0.15$.

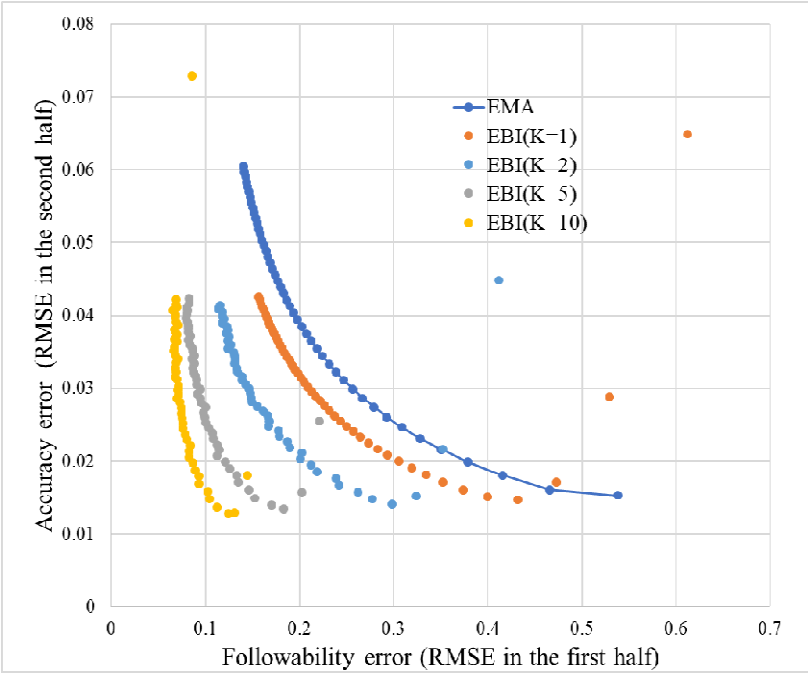


(a)

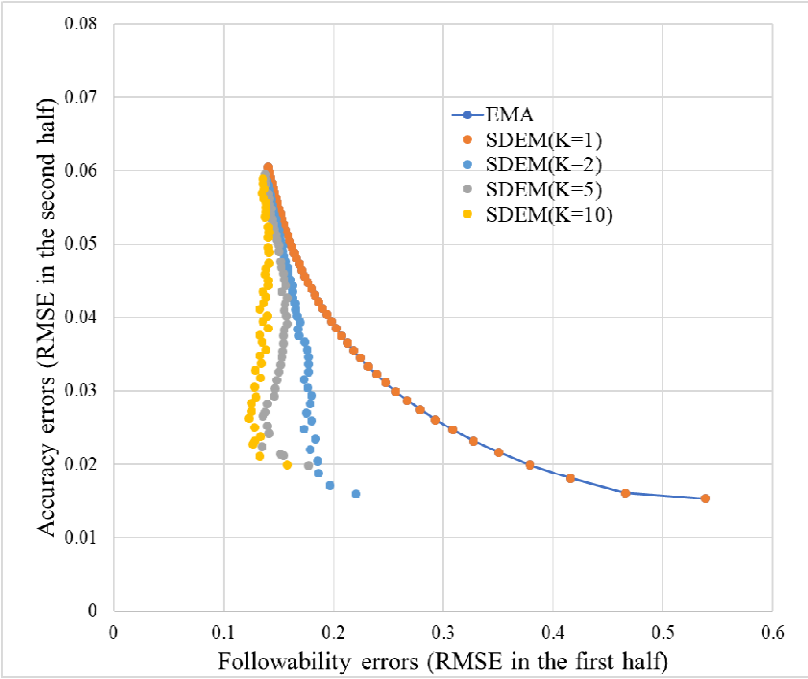


(b)

Fig. 3 Internal state of extended Bayesian inference. (a) Time progress of the mean of normal distribution for each hypothesis. (b) Time progress of the hypothesis with the greatest degree of confidence.

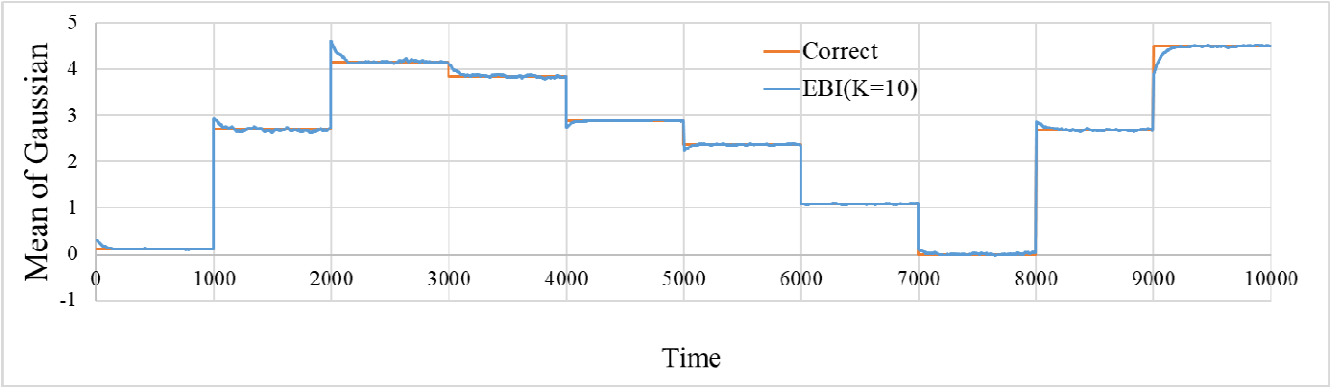


(a)

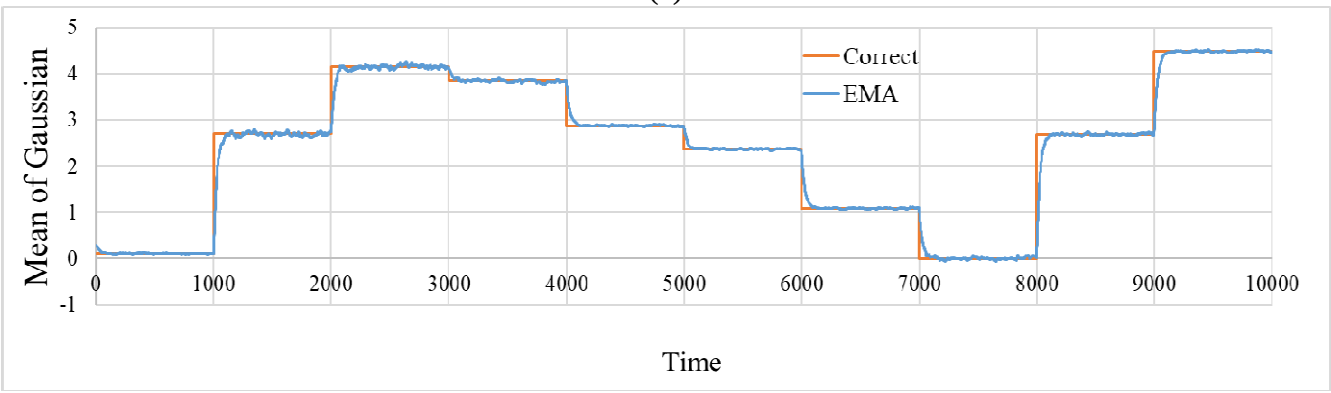


(b)

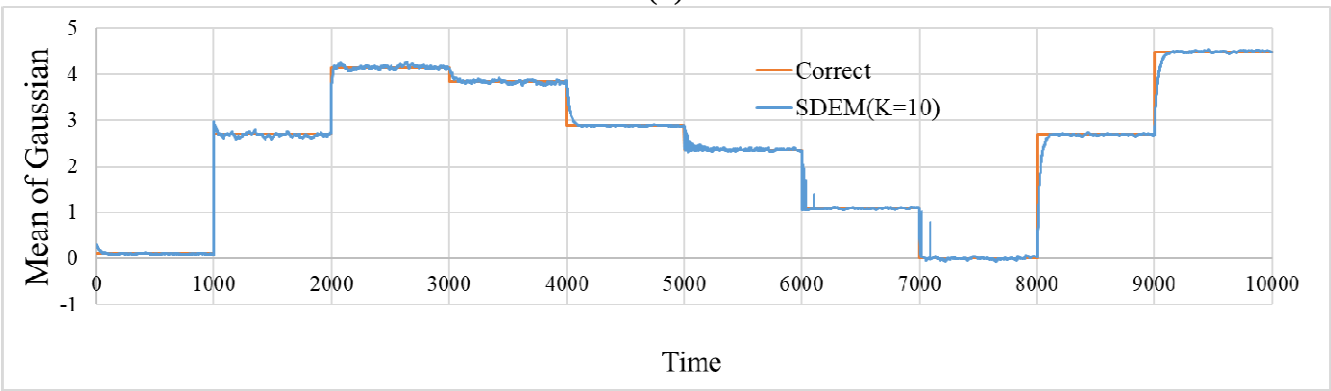
Fig. 4 Relationship between the followability errors and estimation accuracy errors. (a) EBI. (b) SDEM.



(a)

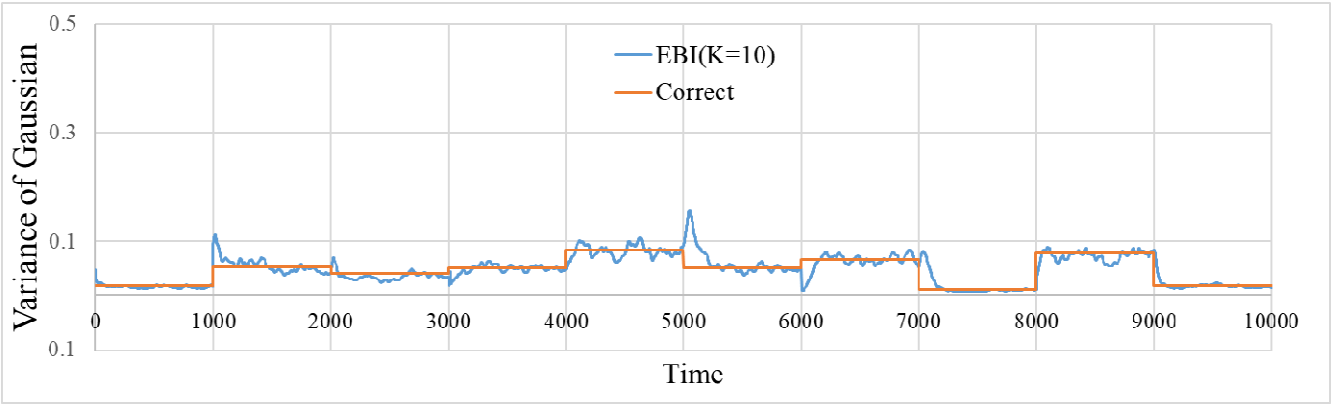


(b)

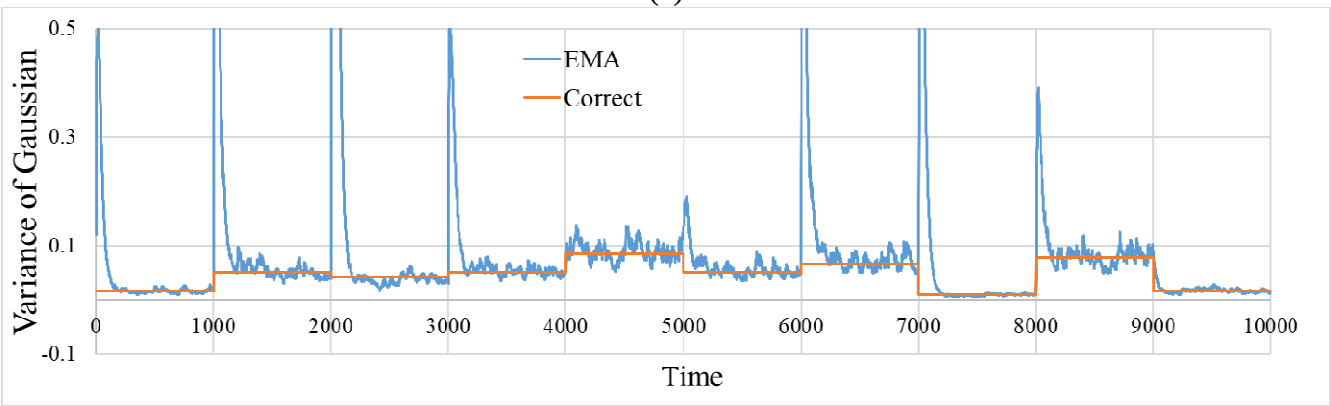


(c)

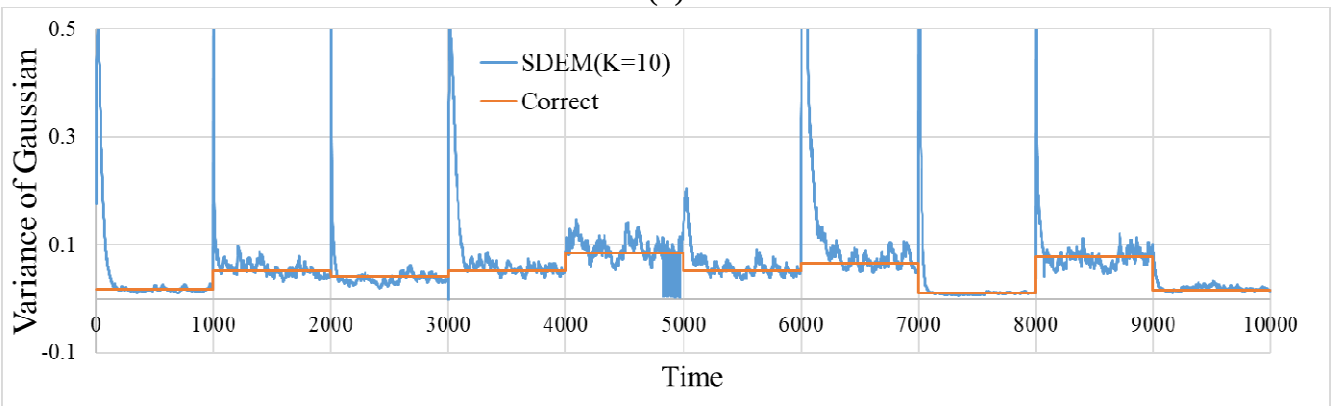
Fig. 5 Time progress of the estimated values for the mean of Gaussian. The figure includes the correct mean. (a) Estimated values by EBI. $(\alpha, m, K) = (0.036, 0.0, 10)$. (b) Estimated values by EMA. $\beta = 0.036$. (c) Estimated values by SDEM. $(\beta, K) = (0.036, 10)$.



(a)



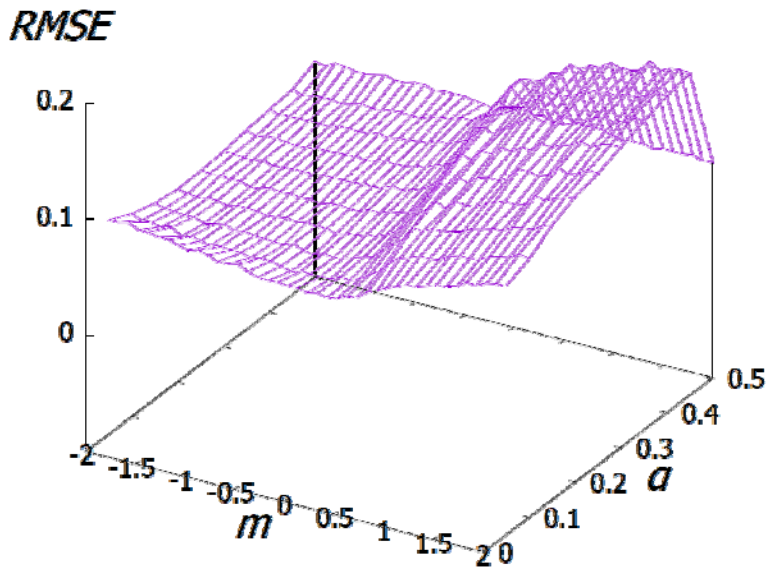
(b)



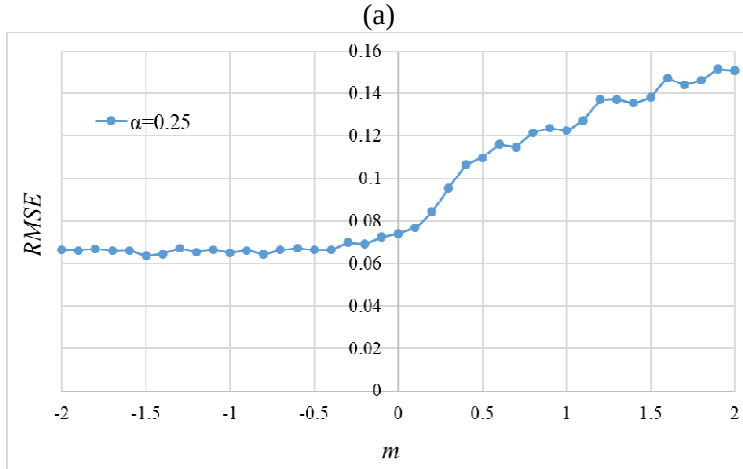
(c)

Fig. 6 Time progress for the estimated values of the variance of the Gaussian. The figure includes the correct variance. (a) Estimated values by EBI. $(\alpha, m, K) = (0.03, 0.0, 10)$. (b) Estimated values by EMA. $\beta = 0.012$. (c) Estimated values by SDEM. $(\beta, K) = (0.012, 10)$.

628



(a)



(b)

Fig. 7 RMSE for each pair of parameter values. (a) In all cases. (b) In the case of $\alpha = 0.25$.

629