A role of 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

Shuji Shinohara

Mathematical Engineering of Morality Emotions, Department of Bioengineering, Graduate School of Engineering, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, 113-8656, Japan,

+81-3-5841-0439

shinohara@bioeng.t.u-tokyo.ac.jp

Abstract

Bayesian inference is a process of narrowing down hypotheses (causes) to one that best explains 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 lessened. Here, we propose an extended Bayesian inference (EBI), incorporating human-like causal inference into Bayesian inference. We show that both a learning effect and a forgetting effect are introduced into Bayesian inference by incorporating the causal inference. We evaluate the estimation performance of EBI through the learning task of a dynamically changing Gaussian mixture model. In the evaluation, the EBI performance is compared with those of EMA and an online expectation-maximisation algorithm SDEM. The EBI was shown to modify the trade-off seen in EMA.

INTRODUCTION

Bayesian inference deduces the hidden cause behind observed data by retrospectively applying statistical inferences. The relationship between Bayesian inference and brain function has been attracting 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 current observational data. In other words, Bayesian inference is a process of narrowing down hypotheses (causal candidates) to one which best explains observational data (the effects).

As an example, consider a situation in which one attempts to read another's emotions. Since 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 emotion corresponds to the hypothesis in Bayesian inference, in the sense that it is the inferred target. In addition, a probability distribution representing what kind 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 a lot of smiles on the person, 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 the someone's emotion 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 or not the observation data is derived from the same emotion throughout the period that one is estimating the emotion from the data. 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-maximisation (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 is often used for mixture models or latent-topic models such as latent Dirichlet allocation (LDA) [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, it is possible to estimate the other's emotion 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] have proposed a sequential discounting expectation-maximisation (SDEM) algorithm that introduces the effect of forgetting in order to deal with unsteady situations where the inferred target changes. The algorithm is used in the fields of anomaly detection and change point detection. We have 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 carried out, seeking 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]. On the other hand, 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 $P(e \mid c)$ and $P(e \mid c)$. 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 $DFH(c|e)=DFH(e|c)$ holds. Such inference is called 'symmetry inference.'

In this paper, we first describe 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 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 the 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 $P^t(d^t)$ is the marginal probability of d^t at time t and is defined as follows:

$$P^t(d^t)=\sum_k P^t(h_k)P(d^t|h_k) \qquad (2)$$

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) into one, 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 follows:

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

Here, the denominator $P^i\left(d^i\right)$ is common to all hypotheses and can be regarded as a constant. Therefore, if normalisation 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 have proposed in the EBI that such a causal induction factor be incorporated into Bayesian inference [14]. In the EBI: first, we defined $C\left(e \mid c\right)$ as a new index representing the strength of the connection

between two events c and e as follows.

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

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

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

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

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

Here α takes values in the range $0.0\le\alpha\le 1.0$ and denotes weighting the values of x and y, and m takes values in the range $-\infty\le m\le\infty$ and denotes the manner of taking the mean. For example, suppose $\alpha=0.5$ and $m=1.0$, then $\mu(0.5,1.0)=0.5x+0.5y$, and represents the arithmetic mean. Suppose $\alpha=0.5$ and $m=-1.0$ then $\mu(0.5,-1.0)=2xy/(x+y)$ representing the harmonic mean.

Supposing $m=0.0$, then the formula (9) is undefinable. If, however, we define the mean value $\mu(\alpha,0.0)$ as the limit of $m\to 0.0$ we get $\mu(\alpha,0.0)=x^{1-\alpha}y^{\alpha}$, where if $\alpha=0.5$, it is the geometric mean $\mu(0.5,0.0)=\sqrt{xy}$. That is, if $m=0.0$ and $\alpha=0.5$, $C(e|c)=DFH(e|c)=\sqrt{P(e|c)P(c|e)}$. On the other hand, if $\alpha=0.0$, $C(e|c)=P(e|c)$ regardless of the value of m. In other words, by introducing parameters m and α, conditional reasoning $P(e|c)$ and causal reasoning $DFH(e|c)$ can be seamlessly connected.

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

$$C(e|c)=\left[\alpha P(e|c)^m+(1-\alpha)P(c|e)^m\right]^{1/m}=\frac{P(c)P(e|c)}{\left[(1-\alpha)P(c)^{-m}+\alpha P(e)^{-m}\right]^{-1/m}} \quad (10)$$

$$C(c\,|\,e)=\left[\alpha P(c\,|\,e)^{m}+(1-\alpha)P(e\,|\,c)^{m}\right]^{1/m}=\frac{P(c)P(e\,|\,c)}{\left[(1-\alpha)P(e)^{-m}+\alpha P(c)^{-m}\right]^{-1/m}} \qquad (11)$$

Here, if we describe formula (10) and (11) in recursive form, and then replace c and e with h_k and d^t, respectively, we can get the next formulas:

$$C^{t+1}\left(d^t\,|\,h_k\right)\leftarrow\frac{C^t\left(h_k\right)C^t\left(d^t\,|\,h_k\right)}{\left[(1-\alpha)C^t\left(h_k\right)^{-m}+\alpha C^t\left(d^t\right)^{-m}\right]^{-1/m}} \qquad (12)$$

$$C^{t+1}\left(h_k\,|\,d^t\right)\leftarrow\frac{C^t\left(h_k\right)C^t\left(d^t\,|\,h_k\right)}{\left[(1-\alpha)C^t\left(d^t\right)^{-m}+\alpha C^t\left(h_k\right)^{-m}\right]^{-1/m}} \qquad (13)$$

$$C^t\left(d^t\right)=\sum\nolimits_k C^t\left(h_k\right)C^t\left(d^t\,|\,h_k\right) \qquad (14)$$

Formula (13) can be rewritten as follows by using a Bayesian update:

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

In the formula (15), a description of the normalisation process for setting the confidence as a probability is omitted.

Assuming $\alpha=0.0$ in formula (15), the same form as formula (4) for Bayesian inference is obtained:

$$C^{t+1}\left(h_k\right)\leftarrow\frac{C^t\left(h_k\right)C^t\left(d^t\,|\,h_k\right)}{C^t\left(d^t\right)} \qquad (16)$$

Supposing $\alpha=0.0$, formula (12) becomes a tautology as shown below:

$$C^{t+1}\left(d^t\,|\,h_k\right)\leftarrow C^t\left(d^t\,|\,h_k\right) \qquad (17)$$

In other words, if $\alpha=0.0$, then formula (12) substantially disappears, and the EBI becomes the same as Bayesian inference. On the other hand, in the case of $\alpha>0.0$, the likelihood is modified by formula (12).

In this paper, we do not update the likelihood of all hypotheses, but update only the likelihood of the hypothesis with the highest confidence at that time. That is, the following formula is used instead of formula (12).

$$C^{t+1}\left(d^t \mid h_k\right) \leftarrow \begin{cases} \dfrac{C^t\left(h_k\right)C^t\left(d^t \mid h_k\right)}{\left[(1-\alpha)C^t\left(h_k\right)^{-m} + \alpha C^t\left(d^t\right)^{-m}\right]^{-1/m}} & \text{if } h_k = \arg\max_{h_i} C^t\left(h_i\right) \\ C^t\left(d^t \mid h_k\right) & \text{otherwise} \end{cases} \quad (18)$$

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

In the following, let us analyse the case of $m = 0.0$, that is, the geometric mean case. In the case of $m = 0.0$, formula (15) can be transformed as follows:

$$C^{t+1}\left(h_k\right) \leftarrow \frac{C^t\left(h_k\right)C^t\left(d^t \mid h_k\right)}{C^t\left(d^t\right)^{1-\alpha} C^t\left(h_k\right)^{\alpha}} = \left[\frac{C^t\left(h_k\right)}{C^t\left(d^t\right)}\right]^{1-\alpha} C^t\left(d^t \mid h_k\right) \quad (19)$$

If we focus on the recursiveness of $C^t\left(h_k\right)$, formula (19) can be rewritten as follows:

$$C^{t+1}\left(h_k\right) \leftarrow \left[C^1\left(h_k\right)\right]^{(1-\alpha)^t} \prod_{i=1}^{t} \frac{\left[C^i\left(d^i \mid h_k\right)\right]^{(1-\alpha)^{t-i}}}{\left[C^i\left(d^i\right)\right]^{(1-\alpha)^{t+1-i}}} \quad (20)$$

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

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

This can be understood as indicating 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 $\alpha = 0.0$, 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.

On the other hand, in the case of $\alpha = 1.0$, $C^{t+1}\left(h_k\right) \leftarrow C^t\left(d^t \mid h_k\right)$. It means that the confidence is calculated using only the current likelihood. Thus, it can be said that EBI introduces the effect of forgetting into Bayesian inference when considering past history.

In the case of $m = 0.0$, with respect to $h_{\max}^t$, formula (18) can be written as follows.

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

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 EBI introduces the effect of learning into Bayesian inference.

Testing a normal distribution model

In this paper, we deal with the following one-dimensional normal distribution as a concrete model for the hypothesis. For simplicity, we assume that the variance is the same Σ at all times and for all hypotheses, and only the mean μ_k^t is different among the hypotheses.

$$C^t\left(d \mid h_k\right) = N\left(d \mid \mu_k^t, \Sigma\right) = \frac{1}{\sqrt{2\pi\Sigma}} \exp\left[-\frac{\left(d - \mu_k^t\right)^2}{2\Sigma}\right] \qquad (23)$$

In the case that 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. For this reason, when calculating formula (21) and (22), a positive number Δ is introduced and the calculation is made approximately as follows.

$$C^{t+1}\left(h_k\right) \leftarrow \left[C^t\left(h_k\right)\right]^{1-\alpha} \Delta N\left(d^t \mid \mu_k^t, \Sigma\right) \qquad (24)$$

$$C^{t+1}\left(d^t \mid h_{\max}^t\right) \leftarrow \frac{1}{\Delta}\left[\frac{C^t\left(h_{\max}^t\right)}{\sum_k C^t\left(h_k\right)\Delta N\left(d^t \mid \mu_k^t, \Sigma\right)}\right]^{\alpha} \Delta N\left(d^t \mid \mu_{\max}^t, \Sigma\right)$$
$$= \frac{1}{\Delta^{\alpha}}\left[\frac{C^t\left(h_{\max}^t\right)}{\sum_k C^t\left(h_k\right) N\left(d^t \mid \mu_k^t, \Sigma\right)}\right]^{\alpha} N\left(d^t \mid \mu_{\max}^t, \Sigma\right) \quad (25)$$

Here, $\mu_{\max}^t$ represents the mean of the normal distribution that is the model for $h_{\max}^t$.

In formula (24), the term Δ is common to all hypotheses and can be cancelled by normalisation. For this reason, if normalisation processing is omitted, it can be expressed as follows.

$$C^{t+1}\left(h_k\right) \leftarrow \left[C^t\left(h_k\right)\right]^{1-\alpha} N\left(d^t \mid \mu_k^t, \Sigma\right) \quad (26)$$

In formula (26), if the confidence for a hypothesis becomes zero once, it remains zero thereafter. In order to prevent this, normalisation processing (smoothing) is performed by adding a small positive constant ε to the confidence of each hypothesis obtained by formula (26).

$$C^{t+1}\left(h_k\right) \leftarrow \frac{C^{t+1}\left(h_k\right)+\varepsilon}{\sum_{j=1}^{K}\left[C^{t+1}\left(h_j\right)+\varepsilon\right]} = \frac{C^{t+1}\left(h_k\right)+\varepsilon}{K\varepsilon+\sum_{j=1}^{K} C^{t+1}\left(h_j\right)} \quad (27)$$

Here, K represents the total number of hypotheses. In this paper, we set to $\varepsilon = 10^{-300}$.

Having observed the data d^t, the likelihood is changed to $C^{t+1}\left(d^t \mid h_{\max}^t\right)$ by formula (25). Concomitantly, the mean of the model for the hypothesis is modified from $\mu_{\max}^t$ to $\mu_{\max}^{t+1}$ so that the following equation is satisfied:

$$C^{t+1}\left(d^t \mid h_{\max}^t\right) = N\left(d^t \mid \mu_{\max}^{t+1}, \Sigma\right) = \frac{1}{\sqrt{2\pi\Sigma}} \exp\left[-\frac{\left(d^t-\mu_{\max}^{t+1}\right)^2}{2\Sigma}\right] \quad (28)$$

Solving equation (28) for $\mu_{\max}^{t+1}$ leads to the following two solutions:

$$\mu_1 = d^t + \sqrt{-2\Sigma \log\left[C^{t+1}\left(d^t \mid h_{\max}^t\right)\sqrt{2\pi\Sigma}\right]}$$
$$\mu_2 = d^t - \sqrt{-2\Sigma \log\left[C^{t+1}\left(d^t \mid h_{\max}^t\right)\sqrt{2\pi\Sigma}\right]} \quad (29)$$

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

$$\mu_{\max}^{t+1} = \begin{cases} \mu_1 & if \quad \left|\mu_1 - \mu_{\max}^{t}\right| \le \left|\mu_2 - \mu_{\max}^{t}\right| \\ \mu_2 & otherwise \end{cases} \quad (30)$$

However, in order to solve equation (28), $C^{t+1}\left(d^t \mid h_{\max}^{t}\right)$ needs to be within the range of $0 < C^{t+1}\left(d^t \mid h_{\max}^{t}\right) \le \frac{1}{\sqrt{2\pi\Sigma}}$. For this reason, we set the following restrictions after calculating $C^{t+1}\left(d^t \mid h_{\max}^{t}\right)$ using formula (25).

$$C^{t+1}\left(d^t \mid h_{\max}^{t}\right) \leftarrow \min\left(\max\left(C^{t+1}\left(d^t \mid h_{\max}^{t}\right), \varepsilon\right), \frac{1}{\sqrt{2\pi\Sigma}}\right) \quad (31)$$

Where $\max(x, y)$ represents the larger of x and y. Conversely, $\min(x, y)$ represents the smaller of x and y. We set $\varepsilon = 10^{-300}$.

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 $\sum_k C^t\left(h_k\right) = 1$, so $C^t\left(d^t\right) = \sum_k C^t\left(h_k\right) C^t\left(d^t \mid h_k\right) \approx C^t\left(d^t \mid h_{\max}^{t}\right)$ is derived from formula (14). Therefore, formula (25) can be transformed as follows.

$$C^{t+1}\left(d^t \mid h_{\max}^{t}\right) \leftarrow \frac{1}{\Delta^{\alpha}}\left[\frac{C^t\left(h_{\max}^{t}\right)}{\sum_k C^t\left(h_k\right) N\left(d^t \mid \mu_k^t, \Sigma\right)}\right]^{\alpha} N\left(d^t \mid \mu_{\max}^{t}, \Sigma\right) \approx \left(\frac{1}{\Delta}\right)^{\alpha}\left[C^t\left(d^t \mid h_{\max}^{t}\right)\right]^{1-\alpha} \quad (32)$$

If formula (32) 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 convex function. Solving equation $x^t = f\left(x^t\right)$ 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 paper, we set $\Delta = \sqrt{2\pi\Sigma}$. In this case, $C^{t+1}\left(d^t \mid h_{\max}^t\right)$ approaches $\frac{1}{\Delta} = \frac{1}{\sqrt{2\pi\Sigma}}$; that is, the vertex of the normal distribution defined by the formula (23), whenever the data d^t is observed.

As shown in equation (28), $\mu_{\max}^{t+1}$ is determined so as to satisfy the condition $C^{t+1}\left(d^t \mid h_{\max}^t\right) = N\left(d^t \mid \mu_{\max}^{t+1}, \Sigma\right)$. 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 $h_{\max}^t$ as inverse Bayesian inference [21–24]. If the former process of updating confidences for hypotheses is referred as inference, inverse Bayesian inference can be called 'learning' in the sense that it forms a model for a hypothetical, rather than inference. In this sense, although the two α in formulas (25) and (26) are denoted by the same α, they can be called 'the learning rate' and 'the forgetting rate,' respectively. As a matter of course, we can also set 'the learning rate' and 'the forgetting rate' as two independent parameters. However, when dealing with temporal alteration as in this paper, good performance is achieved in the case that the two parameters almost have the same value. On the other hand, in spatial clustering, it is better to set them separately.

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

1. Set values for parameters $\alpha, \beta, m, \varepsilon, K$.
2. Give variables $\Sigma, \mu_k^1, C^1\left(h_k\right)$ $(k = 1, 2, \cdots K)$ initial values.
3. Repeat the following whenever the data d^t are observed.
    - Find the hypothesis $h_{\max}^t$ with the maximum confidence.

- Update the confidence $C^{t+1}\left(h_k\right)$ of each hypothesis using formulas (26) and (27).
- 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(25).
- Correct the mean $\mu_{\max}^{t+1}$ of the model for the hypothesis $h_{\max}^t$ using formula (29) and (30) to match the new likelihood $C^{t+1}\left(d^t \mid h_{\max}^t\right)$.

Fig. 1 shows the comparison between Bayesian inference and EBI.

EM algorithm for Gaussian mixture learning

The Gaussian mixture distribution is expressed as a weighted average of K normal distributions as follows.

$$P\left(d\right) \leftarrow \sum_{k=1}^{K} \pi_k N\left(d \mid \mu_k, \Sigma_k\right) \qquad (33)$$

Here, $\sum_{k=1}^{K} \pi_k = 1$ is assumed. π_k is called 'mixing weights' and represents the weighting of each normal distribution. When observing the data d^i, the responsibility q_k^i for the distribution k of the data is calculated as follows.

$$q_k^i = \frac{\pi_k N\left(d^i \mid \mu_k, \Sigma_k\right)}{\sum_{j=1}^{K} \pi_j N\left(d^i \mid \mu_j, \Sigma_j\right)} \qquad (34)$$

This calculation process is called the 'E step.' When observing a plurality of data, the responsibility of each data is calculated for each distribution.

In the EM algorithm, after calculating the responsibility in E step, the M step is executed, and the mixing weights, mean, and variance of each normal distribution are updated.

When N data are observed, the mixing weight of each normal distribution is expressed as the average of the responsibilities of N data as follows.

$$\pi_k = \frac{1}{N}\sum_{i=1}^{N} q_k^i \qquad (35)$$

Normalise to be $\sum_{k=1}^{K}\pi_k = 1$.

$$\pi_k \leftarrow \frac{\pi_k}{\sum_{j=1}^{K}\pi_j} \qquad (36)$$

Here, the weight of each data for each distribution is defined as follows.

$$w_k^i = \frac{q_k^i}{\sum_{j=1}^{N} q_k^j} \qquad (37)$$

The mean of each normal distribution is calculated as a weighted average of each observation data.

$$\mu_k \leftarrow \sum_{i=1}^{N} w_k^i d^i \qquad (38)$$

Similarly, the variance of each normal distribution is calculated as follows.

$$\Sigma_k \leftarrow \sum_{i=1}^{N} w_k^i \left(d^i - \mu_k\right)\left(d^i - \mu_k\right) \qquad (39)$$

The EM algorithm repeats the E step and M step shown above alternately.

Here we compare the EBI with EM algorithm. By setting $\pi_k = P\left(h_k\right)$, $N\left(d^i \mid \mu_k, \Sigma_k\right) = P\left(d^i \mid h_k\right)$, formula (34) can be written as follows.

$$q_k^i = \frac{\pi_k N\left(d^i \mid \mu_k, \Sigma_k\right)}{\sum_{j=1}^{K}\pi_j N\left(d^i \mid \mu_j, \Sigma_j\right)} = \frac{P\left(h_k\right)P\left(d^i \mid h_k\right)}{\sum_{j=1}^{K} P\left(h_j\right)P\left(d^i \mid h_j\right)} = P\left(h_k \mid d^i\right) \qquad (40)$$

In other words, the responsibility can be regarded as the posterior probability after data observation.

In the case of $N = 1$, it becomes $\pi_k = q_k^1$ from formula (35), so the following formula is derived.

$$P\left(h_k\right) = \pi_k = q_k^1 = P\left(h_k \mid d^1\right) \qquad (41)$$

This is consistent with the Bayesian update expressed by formula (3). In other words, EBI can be said to have introduced the effect of forgetting in consideration of the order of observation data.

Sequential Discounting Expectation-Maximisation Algorithm (SDEM)

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

$$q_k^t = \frac{\pi_k^t N\left(d^t \mid \mu_k^t, \Sigma_k^t\right)}{\sum_{j=1}^{K} \pi_j^t N\left(d^t \mid \mu_j^t, \Sigma_j^t\right)} \qquad (42)$$

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 $\beta\left(0<\beta<1\right)$:

$$\tilde{\pi}_k^{t+1} \leftarrow \left(1-\beta\right)\tilde{\pi}_k^t + \beta q_k^t = \beta \sum_{i=1}^{t} \left(1-\beta\right)^{t-i} q_k^i \quad (43)$$

$$\mu_k^{t+1} \leftarrow \left(1-\beta\right)\mu_k^t + \beta \frac{q_k^t}{\tilde{\pi}_k^t} d^t = \beta \sum_{i=1}^{t} \left(1-\beta\right)^{t-i} \frac{q_k^t}{\tilde{\pi}_k^t} d^i \quad (44)$$

$$\Sigma_k^{t+1} \leftarrow \left(1-\beta\right)\Sigma_k^t + \beta \frac{q_k^t}{\tilde{\pi}_k^t}\left(d^t-\mu_k\right)\left(d^t-\mu_k\right) = \beta \sum_{i=1}^{t} \left(1-\beta\right)^{t-i} \frac{q_k^t}{\tilde{\pi}_k^t}\left(d^i-\mu_k\right)\left(d^i-\mu_k\right) \qquad (45)$$

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

$$\pi_k^{t+1} \leftarrow \frac{\tilde{\pi}_k^{t+1} + \gamma}{K\gamma + \sum_{j=1}^{K} \tilde{\pi}_j^{t+1}} \qquad (46)$$

We set $\gamma = 0.001$ for optimal performance. In the case of $K=1$, q_1^t and $\tilde{\pi}_1^t$ are always 1. At this time, formula (44) shows that the new estimated value is obtained as a convex combination of the current estimated value and of the current observed data.

$$\mu_k^{t+1} \leftarrow \left(1-\beta\right)\mu_k^t + \beta d^t \qquad (47)$$

This represents the EMA.

By setting to $\pi_k^{t+1} = P^{t+1}(h_k)$, $q_k^i = P^i(h_k \mid d^i)$ in formula (43) and (46), $P^{t+1}(h_k)$ can be described approximately as follows:

$$P^{t+1}(h_k) \leftarrow \sum_{i=1}^{t} (1-\beta)^{t-i} P^i(h_k \mid d^i) \quad (48)$$

Taking the logarithm of both sides of formula (21) in the EBI, the following transformation can be made:

$$\log\left(C^{t+1}(h_k)\right) \leftarrow \sum_{i=1}^{t} (1-\alpha)^{t-i} \log\left(C^i(d^i \mid h_k)\right) + (1-\alpha)^t \log\left(C^1(h_k)\right) \quad (49)$$

Comparing formula (49) with formula (48), the EBI differs in that it takes a logarithm, and uses likelihood instead of posterior probability.

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

1. Set values for parameters β, γ, K.
2. Give variables $\Sigma_k^1, \mu_k^1, \pi_k^1 \quad (k = 1, 2, \cdots K)$ initial values.
3. Repeat the following whenever the data d^t is observed.
   - E step: Calculate the responsibility q_k^t for each normal distribution k of the observed data d^t, using formula (42).
   - M step: Update the mixing weights π_k^{t+1} using formulas (43) and (46).
   - M step: Correct the mean μ_k^{t+1} and the variance Σ_k^{t+1} of the normal distribution using formulas (44) and (45).

Simulation

To investigate the behaviour 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. However, for simplicity, for this research only the mean of the generation distribution was estimated, and the variance was not estimated.

In this paper, we deal with a task in which the mean of the generation distribution fluctuates randomly every regular interval. Specifically, every 1000 steps, a random number from a uniform distribution of the range [0.0, 5.0] is generated, and the number is set as a new mean of a distribution.

$$\mu_{correct}^{t+1} = \begin{cases} 5 \times rnd^t & if\ t\%1000 = 0 \\ \mu_{correct}^t & otherwise \end{cases} \qquad (50)$$

Here, $\mu_{correct}^t$ represents the mean of the normal distribution used as the generation distribution at time t, that is, the correct value in this task.

rnd^t represents a random number generated from a continuous uniform distribution of the range [0.0, 1.0] at time t. $x\%y$ represents the remainder when integer x is divided by integer y. The variance of the generation distribution was set to $\Sigma_{correct} = 0.09$. Fig. 2 shows an example of time evolution of observation data d^t in this task.

In the simulation, not only EBI but also estimation by SDEM, and EMA were performed for comparison. The estimated value at time t of the EBI is the mean of the model for $h_{\max}^t$, that is $\mu_{\max}^t$ in formula (25). The estimated value by SDEM is the mean of the distribution for which the observed data d^t has the highest responsibility among the normal distributions included in the Gaussian mixture distribution, that is, $\mu_{m^t}^t$, $m^t = \arg\max_k q_k^t$.

RESULTS

Fig. 3 shows an example of the result by EBI. In this simulation, the initial value of the mean of the model for each hypothesis was set to $\mu_k^1 = 2.5$. The distribution variances were the same for all hypotheses and set to $\Sigma = 0.0025$.

Fig. 3 (a) shows time evolution of correct value $\mu_{correct}^{t}$ and that of the estimated result by EBI set to $K=10$. 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. On the other hand, Fig. 3(b) shows the result by the EBI set to $K=1$, that is, that by only the inverse Bayesian inference. In this case, the follow-up performance is not improved over time at all. Fig. 3 (c) shows the estimation results by three types of EMA with different discounting rates β. It can be seen that the larger the discounting rate, as expected, the quicker the response to sudden changes, but the greater the fluctuation during the stable period.

Fig. 4 (a) shows the time evolution of the means μ_{k}^{t} of the models for 10 hypotheses used in the simulation of Fig. 3 (a). Fig. 4 (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. Along with that, it can be seen that it is possible to follow quickly by appropriately switching hypotheses. In the case of $K=1$, because hypotheses cannot be switched, such quick tracking cannot be achieved.

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

$$RMSE=\sqrt{\frac{\sum_{t=j}^{T+j-1}\left(\hat{\mu}^{t}-\mu_{correct}^{t}\right)^{2}}{T}} \quad (51)$$

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

Each interval of 1000 steps, from a change of 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 the RMSE of the second half as a measure of the inaccuracy of the estimation in the stable period.

Fig. 5 shows the relationship between followability errors and estimation accuracy errors of each method. Note that the x-axis and y-axis in this figure indicate RMSE, so the closer to the origin, the higher the estimation performance. In Fig. 5 (a) and (b), the simulations for each method were performed in the cases of $K=1, K=2, K=5, K=10$. Each figure also shows the EMA results as a baseline. The simulations were performed by changing the discounting rate of each method from 0.01 to 0.3 in increments of 0.005. These values are obtained by dividing the interval from time 0 to 10000 into intervals of every 1000 steps, calculating followability errors and accuracy errors in each interval, and averaging them.

The values shown in the figure are average values of 100 trials with different random seeds. This also applies to Fig. 6 and Fig. 7. The initial values μ_k^1 of the centre of each component (normal distribution), were set to random numbers generated from a uniform distribution of the range [0.0, 5.0] in both methods. In the EBI, variances were time-invariant, the same for all hypotheses, and set to $\Sigma = 0.0025$. In SDEM, variances were set to $\Sigma_k^1 = \Sigma_{correct} = 0.09$, which is the correct value for this task.

As for EMA, it can be seen that there is a trade-off: accuracy decreases as followability increases. EBI can break the trade-off seen in EMA. In the case of $K=1$, that is even if only the inverse Bayesian inference is used, the trade-off can be broken, but the performance is improved as the number of components is increased. SDEM can also break the trade-off, but there is no noticeable difference depending on the number of components.

Fig. 6 shows the time evolution of the estimated value of each method, for the case where the number of components and the discounting rate are selected, to achieve the best performance . The correct value is also shown in the figure. In the EBI, it can be seen that there is little fluctuation in the estimated value in the stable period, and the accuracy is high.

In the results shown above, the simulations of EBI were performed for only $m=0$. The results of EBI when m is changed are shown below. Values of α and m were shifted, with a 0.1 increment in the interval [0.0, 1.0] and [-2.0, 2.0], respectively, and simulations were performed to obtain total RMSE

calculated from the entire interval for each pair of parameters. Fig. 7 shows total RMSE for each pair of parameter values. In most α regions, the RMSE is low in the case of $m \leq 0$. When m exceeds 0.0, RMSE increases rapidly.

DISCUSSION AND CONCLUSIONS

In general, if a method such as EMA with the discounting rate is used to improve followability to a sudden change, it is necessary to increase the discounting rate. This means that the estimation is performed weighting the more recent data more heavily. That is, as long as a constant discount rate is used, a trade-off exists: the followability is improved by the higher discounting rate, but the accuracy is lessened.

In this paper, we simulated the task of estimating the data generation distribution in a non-stationary situation, where the data generation distribution changes suddenly. As a result, the EBI proposed in this paper was able to modify the trade-off seen in 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 in order to achieve high performance. In the literature [14], we derived α and m that best fit the causal strength felt by humans, from equation (7) and the eight types of experimental data shown in the literature [15–19]. As a result, the values of α were in the range of 0.25 to 0.6. In other words, they were far from the $\alpha = 0$ value that means conditional probability. On the other hand, 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.

In the EBI, various models for 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 'they.' Assume that someone's facial expression suddenly changed when you are estimating 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 new emotion. On the other hand, it is also possible to consider that it is a new facial expression representing joy.

It is more difficult to detect a sign of a 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. The reason is that in 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, there is a need for a technique for making an appropriate decision using a model while efficiently providing learning to the model from data. In the future, such technique can be expected to be applied to, for example, detection of a disease sign.

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 farther past data is weakened. At the same time, 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, so 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. However, there are some differences. For example, when past history is taken into account in updating the weight of Gaussian mixture distribution, there is a difference between whether the posterior probability is considered or the likelihood is considered, 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 simulation of various tasks.

As limitations, our method cannot estimate variance. In this simulation, only a one-dimensional normal

distribution was handled. We consider variance estimation and extension to multidimensional distribution as future work.

REFERENCES

1. Dehaene, S. *Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts* (Viking, 2014).

2. Chater, N. & Oaksford, M. *The Probabilistic Mind: Prospects for Bayesian Cognitive Science* https://doi.org/10.1093/acprof:oso/9780199216093.003.0001 (Oxford University Press, 2008).

3. Bishop, C. M. *Pattern Recognition and Machine Learning* (Springer, 2011).

4. Blei, D. M., Ng, A. Y. & Jordan, M. I. Latent Dirichlet allocation. *J. Mach. Learn. Res.* **3**, 993–1022, DOI: http://dx.doi.org/10.1162/jmlr.2003.3.4-5.993 (2003).

5. Neal, R. M. & Hinton, G. E. A view of the EM algorithm that justifies incremental, sparse, and other variants in *Learning in Graphical Models* (ed. Jordan, M. I.) 355–368, https://doi.org/10.1007/978-94-011-5014-9_12 (Kluwer Academic Publishers, 1998) .

6. Sato, M. & Ishii, S. On-line EM algorithm for the normalized Gaussian network. *Neural Computation* **12,** 407–432, DOI: https://doi.org/10.1162/089976600300015853 (2000).

7. Yamanishi, K., Takeuchi, J., Williams, G. & Milne, P. On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms. *Data Mining Knowledge Discovery* **8**, 275–300, DOI: https://doi.org/10.1023/b:dami.0000023676.72185.7c (2004).

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

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

10. Dias, J. G. & Cortinhal, M. J. The SKM algorithm: A k-means algorithm for clustering sequential data in *Advances in Artificial Intelligence – IBERAMIA 2008. IBERAMIA 2008. Lecture Notes in Computer Science* (eds. Geffner H., Prada R., Machado Alexandre I. & David N.) 173–182, https://doi.org/10.1007/978-3-540-88309-8_18 (Springer, 2008).

11. Zhong, S. Efficient online spherical k-means clustering. *IEEE International Joint Conference on Neural Networks.* **5**, 3180–3185, DOI: https://doi.org/10.1109/ijcnn.2005.1556436 (2005).

12. Andrews, N. O. & Fox, E. A. Recent developments in document clustering. *Technical Report TR-07-35* https://www.researchgate.net/publication/228350017_Recent_Developments_in_Document_Clustering (2007).

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

14. Shinohara, S. *et al*. Extended Bayesian inference incorporating symmetry bias. Preprint at https://www.biorxiv.org/content/10.1101/698290v1 DOI: https://doi.org/10.1101/698290 (2019).

15. Buehner, M. J., Cheng, P. W. & Clifford, D. From Covariation to Causation: A Test of the Assumption of Causal Power. *J. Exper. Psychol. Learn., Mem., Cogn.* **29,** 1119–1140, DOI: http://dx.doi.org/10.1037/0278-7393.29.6.1119 (2003).

16. Lober, K. & Shanks, D. R. Is causal induction based on causal power? Critique of Cheng (1997). *Psychol. Rev.* **107,** 195–212, DOI: http://dx.doi.org/10.1037/0033-295X.107.1.195 (2000).

17. White, P. A. Making causal judgments from the proportion of confirming instances: The pCI rule. *J. Exper. Psychol. Learn. Mem. Cogn.* **29,** 710–727, DOI: http://dx.doi.org/10.1037/0278-7393.29.4.710 (2003).

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

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

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

21. Arecchi, F. Phenomenology of consciousness: From apprehension to judgment. *Nonlinear Dynamics Psychol. Life Sci.* **15,** 359–375 (2011).

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* **152,** 44–65, DOI: https://doi.org/10.1016/j.biosystems.2016.12.003 (2016)

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

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

ACKNOWLEDGEMENTS

This research is partially supported by the Center of Innovation Program from the Japan Science and Technology Agency, JST. This work was supported by JSPS KAKENHI Grant Numbers JP16K01408 and JP15H03002.

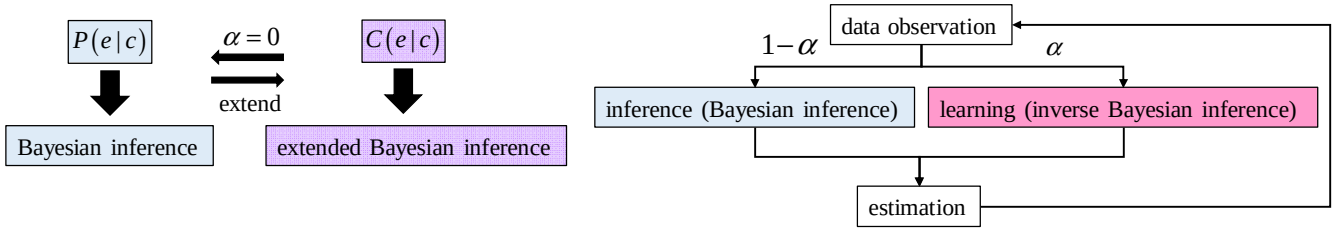
AUTHOR CONTRIBUTIONS STATEMENT

S. S. conceived the extended confidence model and the extended Bayesian inference, conducted the meta-analysis and the simulation, wrote the manuscript. U. C. 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.

ADDITIONAL INFORMATION

Competing interests

The authors report no financial interests or potential conflicts of interest.



(a) Bayesian inference and EBI (b) extended Bayesian Inference (EBI)

Fig. 1. Comparison between Bayesian inference and extended Bayesian inference (EBI). (a) Relationship between Bayesian inference and EBI; and (b) an overview of the processing flow of EBI. In (b), if we suppose $\alpha = 0$, the portion of inverse Bayesian inference disappears, corresponding to the Bayesian inference.

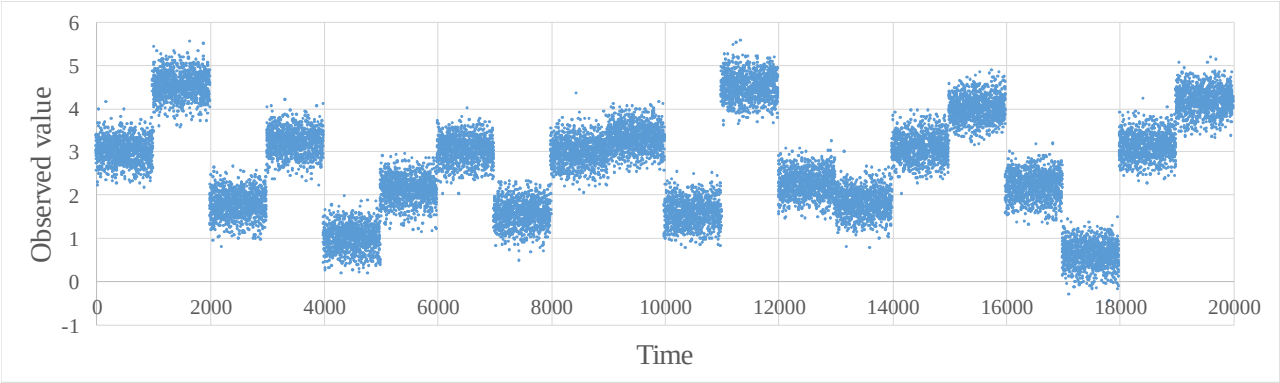
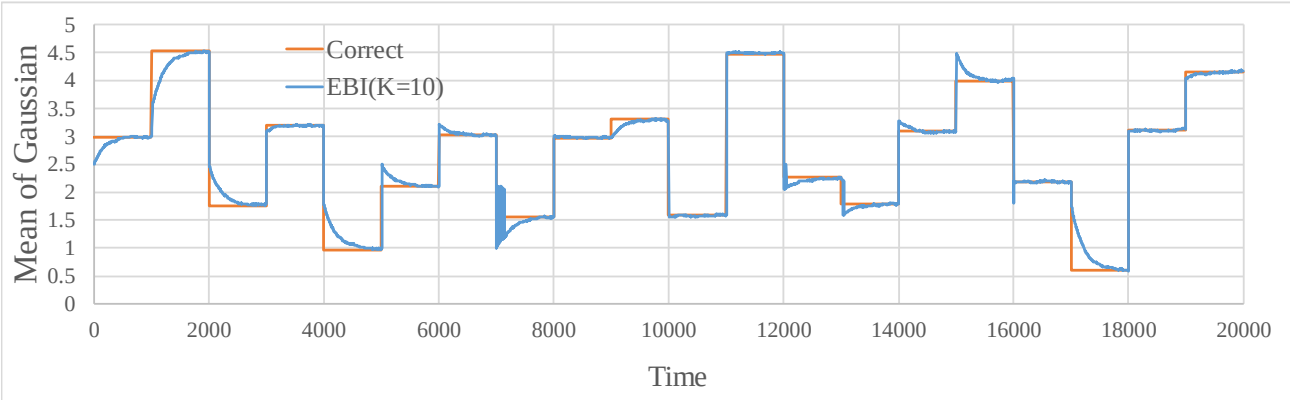
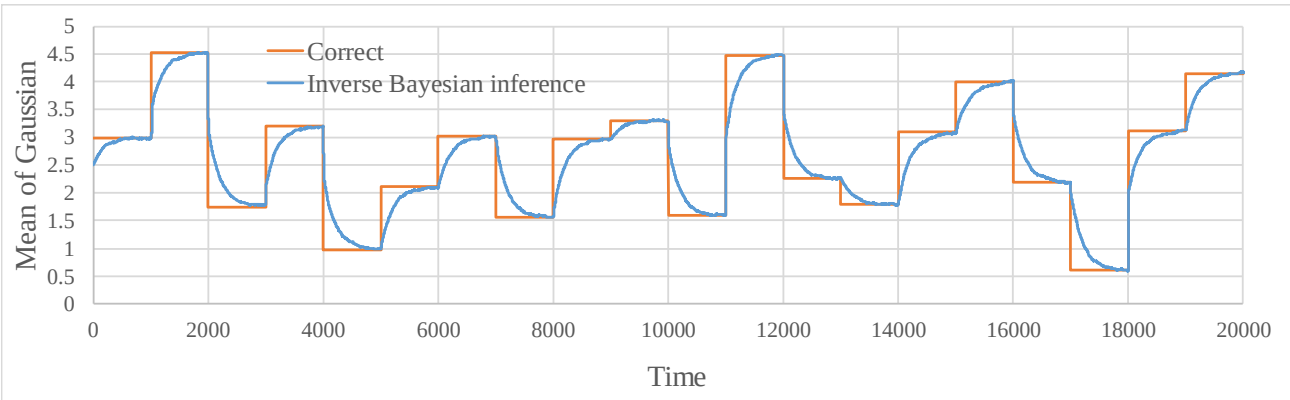


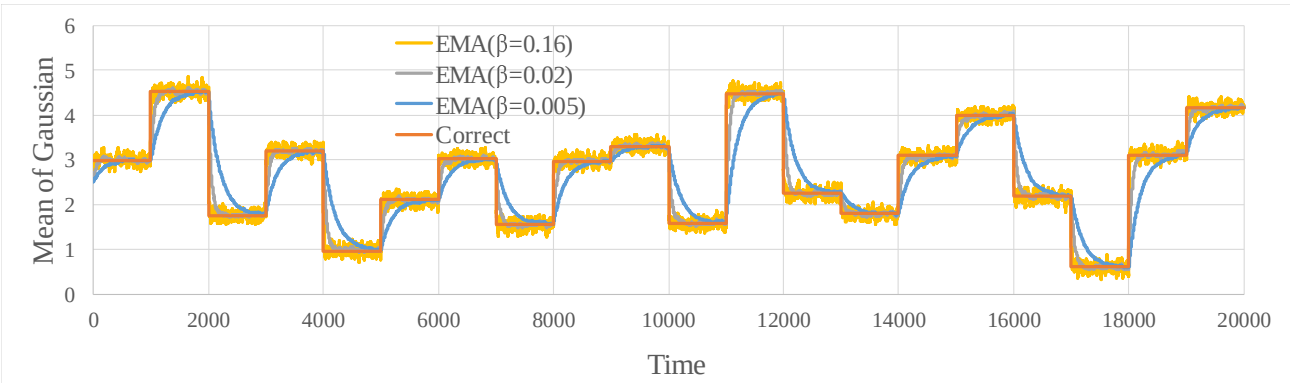
Fig. 2. Time progress of observed values.



(a) Estimated values by EBI. $(\alpha, m, K) = (0.01, 0.0, 10)$

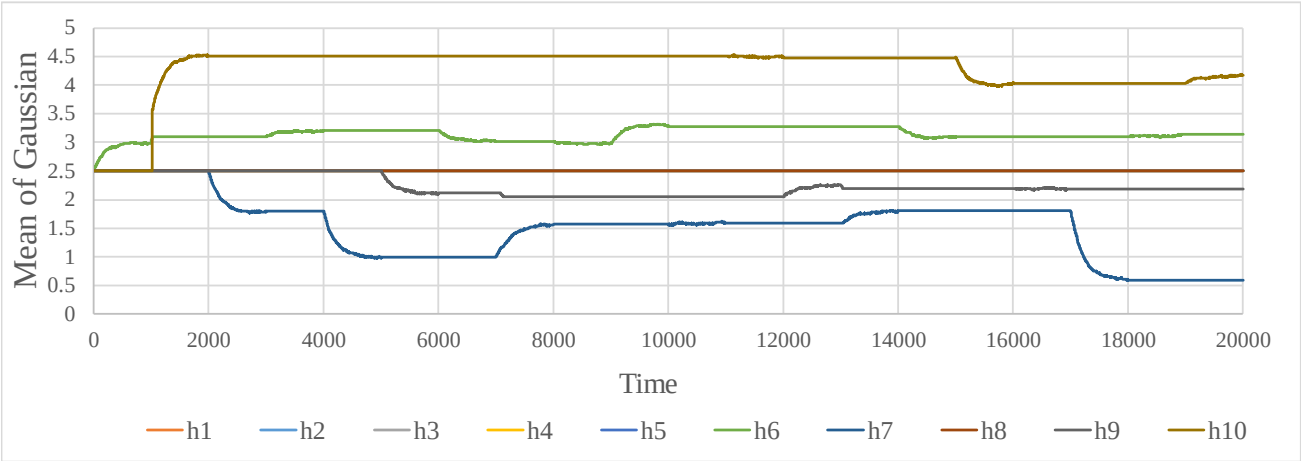


(b) Estimated values by inverse Bayesian inference. $(\alpha, m, K) = (0.01, 0.0, 1)$

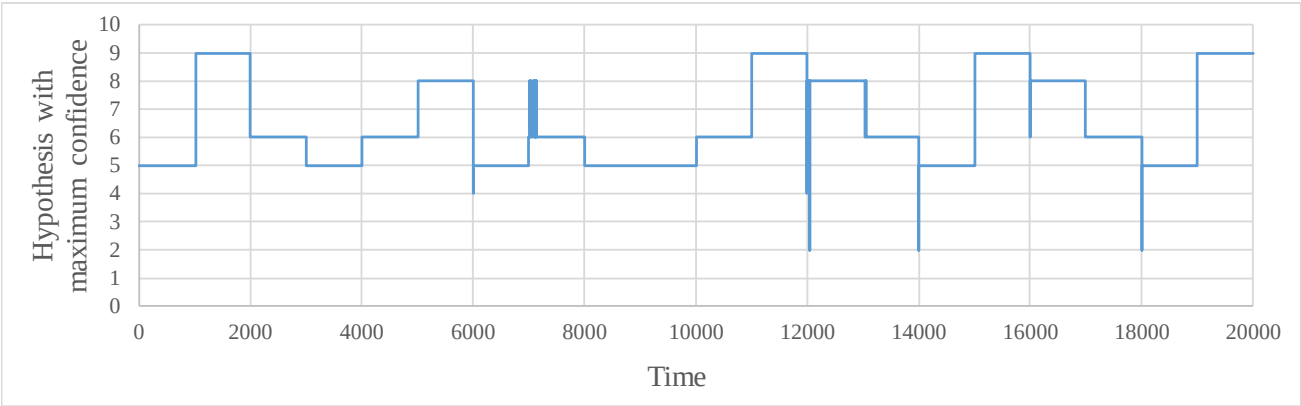


(c) Estimated values by EMA. $\beta = 0.005, 0.02, 0.16$

Fig. 3 Time progress of the estimated values for the mean of Gaussian. The figures include the correct mean.

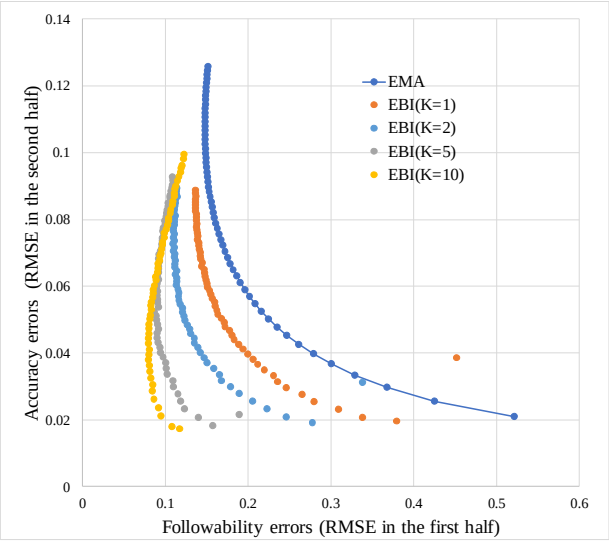


(a) Time progress of the mean of normal distribution for each hypothesis

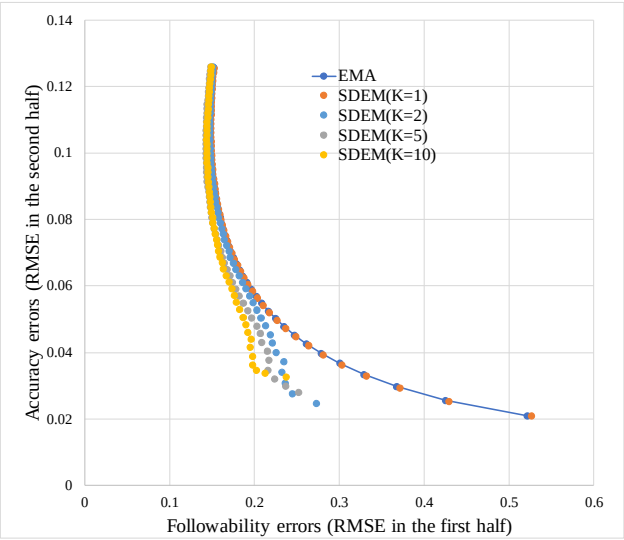


(b) Time progress of the hypothesis with the greatest degree of confidence

Fig. 4 Internal state of extended Bayesian inference

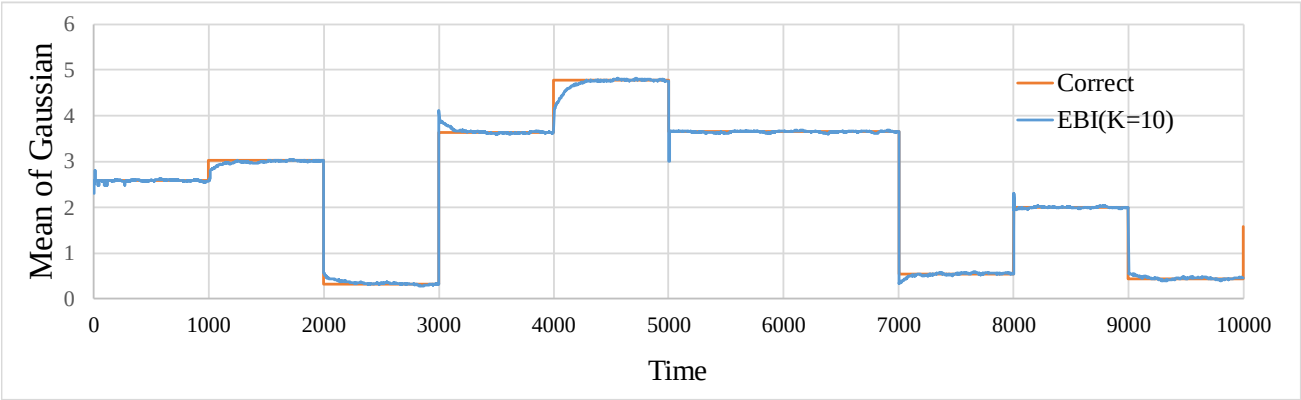


(a) EBI

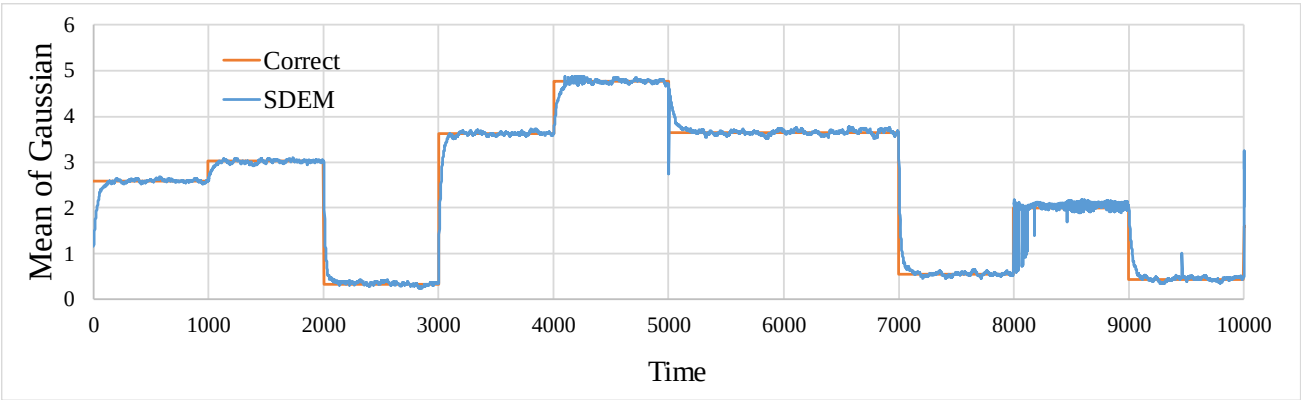


(b) SDEM

Fig. 5 Relationship between the followability errors and the estimation accuracy errors.

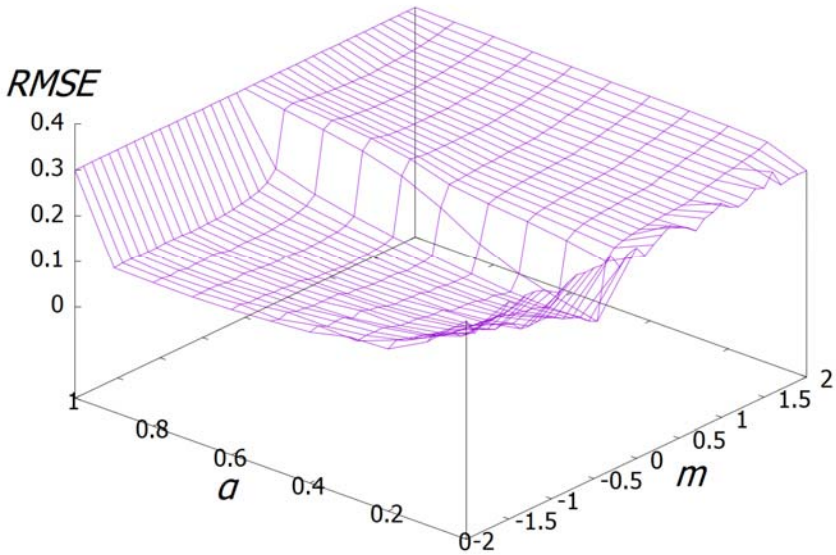


(a) Estimated values by EBI. $(\alpha, m, K) = (0.02, 0.0, 10)$

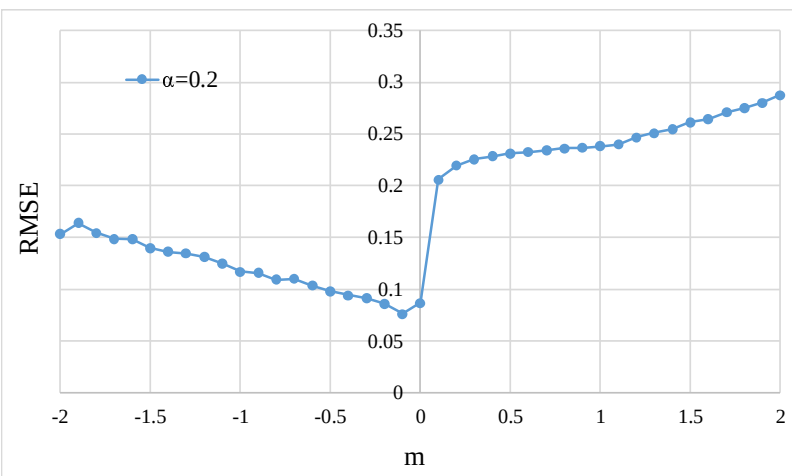


(b) Estimated values by SDEM. $(\beta, K) = (0.03, 10)$

Fig. 6 Time progress of the estimated values for the mean of Gaussian. The figure includes the correct mean.



(a) In all cases.



(a) In the case of $\alpha = 0.2$.

Fig. 7 RMSE for each pair of parameter values.