Development of a variational scheme for model inversion of multi-area model of brain. Part II: VBEM method

Abbas Babajani-Feremi a,*, Hamid Soltanian-Zadeh a,b

a Image Analysis Lab., Radiology Department, Henry Ford Hospital, Detroit, MI 48202, USA
b Control and Intelligent Processing Center of Excellence, Electrical and Computer Engineering Department, University of Tehran, Tehran 14395-515, Iran

A R T I C L E  I N F O

Article history:
Received 23 September 2009
Received in revised form 22 October 2010
Accepted 29 October 2010
Available online 16 November 2010

Keywords:
EEG
MEG
Effective connectivity
Model inversion
Variational Bayesian expectation maximization

A B S T R A C T

In Part I and Part II of these two companion papers (henceforth called Part I and Part II), we develop and evaluate a variational Bayesian expectation maximization (VBEM) method for model inversion of our multi-area extended neural mass model (MEN). In this paper, we develop the VBEM method to estimate posterior distributions of parameters of MEN. We choose suitable prior distributions for the model parameters in order to use properties of a conjugate–exponential model in implementing VBEM. Consequently, VBEM leads to analytically tractable forms. The proposed VBEM algorithm starts with initialization and consists of repeated iterations of a variational Bayesian expectation step (VB E-step) and a variational Bayesian maximization step (VB M-step). Posterior distributions of the model parameters are updated in the VB M-step. Distribution of the hidden state is updated in the VB E-step. We develop a variational extended Kalman smoother (VEKS) to infer the distribution of the hidden state in the VB E-step and derive the forward and backward passes of VEKS, analogous to the Kalman smoother. In Part I, we evaluate and validate the VBEM method using simulation studies.

© 2010 Elsevier Inc. All rights reserved.

1. Introduction

In Part I of these two companion papers (henceforth called Part I), we evaluate and validate performance of a variational Bayesian expectation maximization (VBEM) method for model inversion of a multi-area extended neural mass model (MEN) using simulation studies. Details of the VBEM method are presented in this paper (Part II of these two companion papers henceforth called Part II). The proposed VBEM method is based on a variational Bayesian approach introduced by Beal and Ghahramani for learning posterior distributions of parameters of an ordinary linear state-space model (SSM) [1–3]. Based on their work, we develop a VBEM method for learning posterior distributions of parameters of MEN which is a non-linear SSM. We choose suitable prior distributions for the model parameters in order to use properties of the conjugate–exponential (CE) model in the implementation of the VBEM method. Consequently, the VBEM algorithm leads to analytically tractable forms.

The VBEM algorithm starts with initialization and consists of repeated iterations of a variational Bayesian expectation step (VB E-step) and a variational Bayesian maximization step (VB M-step). Posterior distributions of the model parameters are updated in the VB M-step. Distribution of the hidden state is updated in the VB E-step. We introduce a variational extended Kalman smoother (VEKS) to infer distribution of the hidden state in the VB E-step. We derive the forward and backward passes of VEKS analogous to the Kalman filter and Rauch–Tung–Striebel smoothing algorithms [4]. In this paper, our focus is to establish and develop the VBEM method as an effective connectivity analysis tool for E/MEG data.

The VBEM scheme is formerly similar to Daunizeau et al. [5] with important exceptions that we consider novel contributions. The variational Bayesian (VB) scheme used in [5] is based on the VB-Laplace approximation on the parameters and hidden-states of a non-linear state-space model. Although variances of state and observation noise have conjugate priors in [5] but the overall scheme is not conjugate and thus it does not lead to an analytically tractable form for posterior distributions of the parameters and hidden states. Our proposed VBEM scheme is a full CE model for parameters, hyper parameters, and hidden states of MEN and leads to analytically tractable forms for all posterior distributions. In addition, we derive specific equations required to invert MEN. This is non-trivial given non-linearities and complexity of MEN.

Organization of the paper is as follows. In the variational Bayesian approach section, we introduce theorems required for the VBEM learning method. In the next section, we develop a VBEM method for MEN where prior distributions of the parameters, the VB M-step, and the VB E-step are derived. Finally, conclusions are presented at the end of the paper.

* Corresponding author. Address: Radiology Image Analysis Lab., Henry Ford Health System, One Ford Place, 2F, Detroit, MI 48202, USA. Tel.: +1 313 874 4160; fax: +1 313 874 4494.
E-mail address: abbash@rad.hfh.edu (A. Babajani-Feremi).

0025-5564/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.mbs.2010.11.001
2. Variational Bayesian approach

In this section, we introduce our variational Bayesian approach for model inversion of MEN which is, in principle, a variational extended Kalman smoother. It is variational in that it is based upon variational or ensemble learning of a variational (free-energy) bound on the log-evidence or marginal likelihood for a particular model. It is a Kalman smoother in that it requires both forward and backward passes through the time-series to optimize the posterior density of the hidden states. We will first discuss the conventional expectation maximization (EM) optimization to motivate the particular form of our variational scheme. Then, we will present the variational homolog of EM. A key difference between the two approaches is that the conventional EM scheme ignores uncertainty about the time-invariant model parameters. In contradistinction, the proposed variational scheme optimizes an approximate posterior density for both of the parameters and the hidden states of the model.

2.1. Expectation maximization optimization

Consider a model with hidden variables $x$ and observed variables $y$ where the stochastic dependencies between the variables are described by parameters $\theta$. In particular, consider a model that produces an i.i.d. observation dataset $y = \{y_1, \ldots, y_T\}$ using a set of hidden variables $x = \{x_1, \ldots, x_T\}$ such that logarithm of the likelihood, as a function of $\theta$, is written as

$$ L(\theta) = \ln p(y|\theta) = \ln \prod_{t=1}^{T} p(y_t|\theta) = \sum_{t=1}^{T} \ln \int dx_t p(x_t, y_t|\theta). $$

The maximum likelihood method (ML) seeks to find parameter $\hat{\theta}_{ML}$ that maximizes the likelihood, or equivalently logarithm of the likelihood in Eq. (1). The problem of maximizing $L(\theta)$ with respect to $(w.r.t.) \theta$ can be simplified by introducing an auxiliary distribution over the hidden variables that gives rise to a lower bound on $L(\theta)$. A free distribution $q_{\theta}(x_t)$ over the hidden variables can be used to obtain a lower bound using Jensen’s inequality [6]

$$ L(\theta) = \sum_{t=1}^{T} \ln \int dx_t p(x_t, y_t|\theta) = \sum_{t=1}^{T} \ln \int dx_t q_{\theta}(x_t) \frac{p(x_t, y_t|\theta)}{q_{\theta}(x_t)}, $$

$$ \geq \sum_{t=1}^{T} \int dx_t q_{\theta}(x_t) \ln \frac{p(x_t, y_t|\theta)}{q_{\theta}(x_t)} = \sum_{t=1}^{T} \left[ \int dx_t q_{\theta}(x_t) \ln p(x_t, y_t|\theta) - \int dx_t q_{\theta}(x_t) \ln q_{\theta}(x_t) \right] = \mathcal{F}(q_{\theta}(x_t), \theta), $$

where the negative of $\mathcal{F}(\cdot, \cdot)$ is known as the free energy in statistical physics.

The expectation–maximization (EM) method is an iterative algorithm that alternates between two steps (E-step and M-step) to derive a lower bound between $L(\cdot)$ and $L(\theta)$. At each iteration, the E-step and the M-step maximize $\mathcal{F}(q_{\theta}(x_t), \theta)$ w.r.t. $q_{\theta}(x_t)$ and $\theta$, respectively.

**EM E-step:** $q_{\theta}^{(k+1)}(x_t) = \arg \max_{q_{\theta}(x_t)} \mathcal{F}(q_{\theta}(x_t), \theta^{(k)}), \quad (4)$

**EM M-step:** $\theta^{(k+1)} = \arg \max \mathcal{F}(q_{\theta}^{(k+1)}(x_t), \theta), \quad (5)$

where superscript $k$ represents the iteration number. It can be shown that two steps in Eqs. (4) and (5) are equal to the following steps [2].

**EM E-step:** $q_{\theta}^{(k+1)}(x_t) = p(x_t|y_t, \theta^{(k)}), \quad \forall t = 1, \ldots, T, \quad (6)$

**EM M-step:** $\theta^{(k+1)} = \arg \max_{\theta} \sum_{t=1}^{T} \int dx_t q_{\theta}^{(k+1)}(x_t) \ln p(x_t, y_t|\theta). \quad (7)$

Starting with an initial estimate of parameter $\theta$, the EM algorithm iteratively maximizes the free energy $F(\cdot, \cdot)$ using the above two steps to find an optimal value of parameter $\theta$.

2.2. Variational Bayesian expectation maximization (VBEM) method

Consider a model (say model $m$) with hidden variables $x$ and observed variables $y$ where unknown parameters $\theta$ describe stochastic dependencies between the variables. We can use the Bayesian learning in the model to achieve two main goals. The first goal is to find the posterior distribution over parameters of the model, i.e., $p(\theta|y,m)$. The second goal is to perform model comparison using the approximate marginal likelihood $p(y|m)$. In the Bayesian learning, the marginal likelihood of model $m$, $p(y|m)$, is maximized.

$$ \ln p(y|m) = \ln \int d\theta \exp \left[ \ln p(y, \theta|m) \right]. $$

The marginal likelihood can be lower bounded by introducing free distributions over both of the hidden variables and the parameters ($q_{\theta}(x_t)$ and $q_{\theta}(\theta)$) using Jensen’s inequality.

$$ \ln p(y|m) = \ln \int d\theta \exp \left[ \ln p(y, \theta|m) \right] $$

$$ \geq \int d\theta q_{\theta}(\theta) \left[ \ln p(x_t, y_t|\theta,m) \right] = \int d\theta q_{\theta}(\theta) \left[ \ln p(x_t, y_t|\theta,m) + \ln \frac{p(\theta,m)}{q_{\theta}(\theta)} \right] = \mathcal{F}(q_{\theta}(x_t), q_{\theta}(\theta)), $$

where Jensen’s inequality [7] is used in Eq. (10) and $p(\theta|m)$ is prior distribution of the parameters of model $m$. The variational Bayesian algorithm iteratively maximizes the free energy $F(\cdot, \cdot)$ w.r.t. the free distributions $q_{\theta}(x_t)$ and $q_{\theta}(\theta)$. The following theorem provides update equations for the variational Bayesian learning.

**Theorem 1** (Variational Bayesian expectation maximization (VBEM)). Let $m$ be a model with parameters $\theta$, an i.i.d. observation dataset $y = \{y_1, \ldots, y_T\}$ and the corresponding hidden variables $x = \{x_1, \ldots, x_T\}$. A lower bound on the model log marginal likelihood in Eq. (10) can be iteratively optimized by performing the following updates.

**VB E-step:**

$$ q_{\theta}^{(k+1)}(x_t) = \frac{1}{Z_{x_t}} \exp \left[ \int d\theta q_{\theta}^{(k)}(\theta) \ln p(x_t, y_t|\theta, m) \right], $$

**VB M-step:**

$$ q_{\theta}^{(k+1)}(\theta) = \frac{1}{Z_\theta} \exp \left[ \ln p(x_t, y_t|\theta, m) q_{\theta}^{(k+1)}(\theta) \right], $$

where superscript $k$ denotes the iteration number, $Z_x$ and $Z_{\theta}$ are normalization constants, and $\langle \cdot \rangle$ represent the expected value over the free distributions. In addition, the update rules converge to a local maximum of $F(q_{\theta}(x_t), q_{\theta}(\theta))$.

**Proof.** See [2]. □
By considering a particular class of graphical models, conjugate-exponential (CE) models, the VBEM algorithm will be more similar to the classical EM algorithm. Moreover, it leads to analytically tractable forms of Bayesian integrals in Eqs. (11) and (12). CE models satisfy the following condition.

**Condition 1**: The complete-data likelihood and the parameter prior, as its conjugate, belong to the exponential family if

\[ p(x_t, y_t | \theta) = h_1(\theta) f(x_t, y_t) \exp[\Omega(\theta)^T \cdot z(x_t, y_t)], \]

\[ p(\theta | \omega, \nu) = h_2(\omega, \nu) h_1(\theta)^\nu \exp[\Omega(\theta)^T \cdot \nu], \]

where \(z(\cdot, \cdot)\) and \(f(\cdot, \cdot)\) are two functions that define the exponential family, \(\Omega(\theta)\) is a vector of the natural parameters, \(\omega\) and \(\nu\) are hyperparameters of the prior, superscript ‘tr’ represents the transpose operator, and \(h_1(\cdot)\) and \(h_2(\cdot, \cdot)\) are two normalization constants.

The VBEM algorithm for CE models can be derived using the following theorem that shows how properties of CE family make it especially amenable to the VB approximation.

**Theorem 2.** **VBEM for conjugate–exponential models**

\[ q^{k+1}(x_t) = p(x_t | y_t, T^{(k)}) \propto \int f(x_t, y_t) e^{\bar{z}(x_t, y_t)}, \]

VB E-step: \[ q^{k+1}(x_t) = \prod_{t=1}^{T} q^{k+1}(x_t), \]

VB M-step: \[ \bar{z}(x_t, y_t) = \mathbb{E}[z(x_t, y_t) | q^{k+1}(x_t)]. \]

where \(\bar{z}(x_t, y_t) = \mathbb{E}[z(x_t, y_t) | q^{k+1}(x_t)]\) is the expectation of the sufficient statistic \(z(\cdot, \cdot)\).

**Proof.** See [2].

According to **Theorem 2**, CE formalism provides the following simplification results. The first result is that the posterior over hidden variables calculated in Eq. (16) as the variational Bayesian E-step is similar to that in E-step of the EM algorithm in Eq. (6). The second result is related to the VB M-step in Eq. (17) where the analytical form of the variational posterior \(q^{k+1}(\theta)\) does not change during iterations of VBEM. If it were able to change in the general case of **Theorem 1**, the posterior could quickly become unmanageable and the algorithm became too complicated.

### 3. Variational Bayesian approach for MEN

In this section, we describe our variational scheme, which is effectively a conventional variational scheme with two variational steps. The first step (E-step) optimizes an approximate posterior density on the hidden states, while the second step (M-step) optimizes the posterior density on the parameters. To handle the computational load of dealing with the time-series of the hidden states, we appeal to a Kalman-like formulation to provide a variational Kalman smoother. Furthermore, to finesse the variational scheme per se, we use conjugate priors on the parameters. Finally, to ensure that the true posteriors are approximately Gaussian, we use a locally linear model that uses a local linear approximation of the non-linear sigmoid function.

#### 3.1. Parameters and topology of model

In Part I, we derive the following discrete state-space representation for MEN.

\[ x_t = A_l x_{t-1} + A_2 + (\theta_T \otimes E_0) x_{t-1} + (\theta_{tp} \otimes E_1) u_t + w_t, \]

\[ y_t = \text{diag}(\theta_r) \cdot C x_t + r_t, \]  

where \(t\) is a discrete time point, \(T\) is the total number of samples, \(x_t \in \mathbb{R}^{N_{so}}\) and \(y_t \in \mathbb{R}^{N_{so}}\) are state and observation vectors, \(N\) is the number of cortical areas, \(u_t \in \mathbb{R}^{N_{so}}\) is the input vector that represents the effects of the external stimuli to the cortical areas, and \(A_{1_{so-so}}, A_{2_{so-so}}, E_{so-so}, E_{1_{so-so}}, \) and \(C_{so-so}\) are fixed matrices. Here, the sigmoid function ‘S’ transforms the average membrane potential into an average rate of fired action potentials as described by

\[ S(x) = \frac{2e_0}{1 + e^{-x}} - e_0. \]
A sequence of the state vector $x$ in the state Eq. (18) follows a first-order Markov process. Therefore, a joint probability of the sequence of states and observations of the dynamical system in Eq. (18) is given by

$$p(x, y |0, m) = p(x_0, y_1 |0, m) t \prod_{t=1}^{T} p(x_t |x_{t-1}, 0, m)p(y_t |x_t, 0, m).$$

Here, the prior over the hidden state at time $t = 0$ is Gaussian distributed, i.e., $p(x_0 |m) = \mathcal{N}(\mu_0, \Sigma_0)$ where the mean $\mu_0$ and covariance matrix $\Sigma_0$ can be considered as hyperparameters of the model. Because the state-space representation in Eq. (18) shows dynamics of the multi-area model that generates an event-related signal, we assume that all hidden states are at rest at time $t = 0$ and thus $x_0 = 0$. Therefore, we will neglect the term related to $x_0$. The following linear state-space representation is derived from the dynamical model in Eq. (29) will be a CE model if we assign the model parameters using the E-step in Eq. (6) and the M-step in Eq. (7). For linear dynamical systems with Gaussian state and observation noise, the conditional distribution in the RHS of Eq. (6) is Gaussian. It can be shown that the solution of the Kalman smoother, known as the Rauch–Tung–Striebel smoother [4], is equivalent to the EM E-step for linear systems [1].

One way to reduce complexity of the problem is to approximate including NDS in Eq. (18), exact inference equations are intractable but linear system. The sigmoid function $\text{sigmoid}(x)$ cannot be written down in closed form. How-ever, if we consider a local linearization of the model, it can be shown that the solution of the extended Kalman smoother (EKS) is equivalent to the EM E-step for a non-linear system [8].

The EM method can be applied to estimate parameters of NDS in Eq. (18), i.e., $\theta = \{\theta_0, \theta_1, \theta_2, \theta_3, \rho_0, \rho_1\}$, using a given observation vector $y_{1:T}$ and a known input $u_{1:T}$. Two steps of the EM method in Eqs. (6) and (7) are applied to the locally linear model in Eq. (29) to estimate parameter $\theta = \{\theta_0, \theta_1, \theta_2, \theta_3, \rho_0, \rho_1\}$. For the EM E-step, free distribution over the hidden variable has the following Gaussian distribution.

$$q_{k+1}^{(k)}(x_t) = p(x_t | y_{1:T}, \theta^{(k)}) \sim \mathcal{N} \left( \mu_{k+1}^{(k)}, \Sigma_{k+1}^{(k)} \right),$$

where mean $\mu_{k+1}^{(k)}$ and covariance matrix $\Sigma_{k+1}^{(k)}$ are calculated in the $(k + 1)$th iteration of the EM E-step using solution of EKS. To calculate solution of EKS, parameters of the model are fixed at their estimated values in previous iteration of the EM algorithm. In the EM M-step according to Eq. (7), the calculated mean and covariance of the normal distribution $q(x_{k+1}^{(k)}|y_{1:T})$ is used to find the optimal value of parameter $\theta$.

We now turn to the variational Bayesian treatment of MEN. A point estimate of parameter $\theta$ is calculated in the EM method. However, the parameter $\theta$ has a probability density function in the variational Bayesian analysis method that can be estimated using the VBEM optimization scheme. The complete-data likelihood for the locally linear system in Eq. (29) is Gaussian, which belongs to the exponential family distribution, and thus Eq. (13) of Condition 1 is satisfied. To benefit from using the properties of the CE model and implement the VBEM method based on Theorem 2, we will propose suitable prior distributions for the model parameters.

The EM E-step in Eq. (6) and the VB E-step in Eq. (16) have similar representations. Since solution of EKS is equivalent to the EM E-step, we should be able to use an algorithm similar to EKS for inference of the hidden state sequence’s sufficient statistics. We derive the forward and backward passes analogous to the Kalman filter and Rauch–Tung–Striebel smoothing algorithms [4], and call the proposed method as variational extended Kalman smoother (VEKS). The VB M-step is fairly straightforward. The CE formulation based on Theorem 2 can be used to infer posterior distributions of the model parameters in the VB E-step. However, for the VB M-step, we prefer to employ Eq. (12) based on Theorem 1 to motivate and understand the form of the prior distributions of the model parameters.

3.3. Prior distributions of model parameters

To benefit from simplification of the CE model in the variational Bayesian analysis of MEN, the complete-data likelihood and the parameter prior need to satisfy Condition 1. The prior distribution of the hidden state at $t = 0$ as well as distributions of state and observation noises are assumed to be Gaussian. Consequently, using a locally linear model, conditional distribution of the hidden state at any given history of the outputs and inputs will be Gaussian. Therefore, the complete-data likelihood, i.e., $p(x, y |0)$, is Gaussian and Eq. (13) of Condition 1 is satisfied. On the other hand, we need to assign appropriate prior distributions to the model parameters and satisfy Eq. (14) of Condition 1.
In Appendix B, we show that if rows of matrices $\theta'$ and $\theta''$ as well as elements of vector $\theta''$ have Gaussian distributions and precision vectors $\rho'$ and $\rho''$ have Gamma distributions, Eq. (14) of Condition I will be satisfied and thus the dynamical system in Eq. (29) will be a CE model. In Appendix A, we derive appropriate prior distributions for the parameters of the multi-area model to generate a CE model.

3.4. VB M-step: parameter distributions

Free distributions over the model parameters are updated in the VB M-step of the VBEM algorithm using Theorem 1. In the VB M-step, based on Theorem 1 and Eq. (12), the posterior distributions of the model parameters are derived from two terms: the prior distributions of the parameters, i.e., $p(\theta|m)$, and an exponential term.

$$q_n(\theta) = \frac{1}{Z}p(\theta|m) \cdot \exp\left(\ln p(x, y|0, m)\right)q_0(\theta(x)).$$

In Appendix B, we calculate the exponential term. Then, we show that if Eqs. (A-1)–(A-6) represent the prior distributions of the parameters, the VB M-step leads to an analytically tractable form of Bayesian learning. Finally, the posterior distributions of the model parameters are extracted in the following form using Eqs. (B-14)–(B-19).

$$q_n(\theta) = \prod_{n=1}^{N} q(n_0(\theta_n|\theta_n^0, \mu_n^0, \rho_n^0, \rho_n'^0))$$

$$= \prod_{n=1}^{N} q(n_0(\theta_n|\theta_n^0, \mu_n^0, \rho_n^0) \cdot q_0(\mu_n^0) \cdot q_0(\rho_n^0)\cdot q_0(\rho_n'^0)$$

It is notable that we assume that the state noise $w_t \sim N(0, A_t)$ and the observation noise $\gamma_t \sim N(0, A_t)$ are Gaussian in our model. Based on this assumption and to satisfy the condition of conjugacy, we show in Appendix B that the precision vectors $\rho'$ and $\rho''$ have Gamma distributions and long range connection (LRC) parameter $\theta'$, input parameter $\theta''$, and output parameter $\theta'''$ have Gaussian distributions.

3.5. VB E-step: variational extended Kalman smoother (VEKS)

After updating posterior distributions of the parameters of the dynamical system in Eq. (10) using the VB M-step, the free distribution over the hidden state, i.e., $q_n(x_t|1)$, is calculated in the VB E-step based on Eq. (16) in Theorem 2. To provide a tractable form for distribution of the hidden state of the non-linear multi-area model, we use a locally linear approximation of the state Eq. (10). It generates a Gaussian distribution for the hidden state assuming a Gaussian noise in the model. Let $\mu_{t-1}$ be the mean of the state vector $x_{t-1}$ at time point $t-1$. We use a locally linear approximation of the sigmoid function $S(\cdot)$.

$$S(x_{t-1}) \approx S(\mu_{t-1}) + S'_{t-1}(x_{t-1} - \mu_{t-1}) = S(\mu_{t-1}) + \epsilon_{t-1}.$$  

$$S'_{t-1} = \text{diag}\left(\frac{\partial S(x)}{\partial x} \mid x = \mu_{t-1}\right), \quad \epsilon_{t-1} = S(\mu_{t-1}) - S_{t-1} \cdot \mu_{t-1}.$$

Using the above equations, the non-linear multi-area model in Eq. (10) can be approximated by the following linear but time-variant system.

$$X_t = A_t(\theta)x_{t-1} + B_t(\theta)U_t + W_t,$$

$$y_t = C_t(\theta)x_t + D_t(\theta)U_t + \nu_t,$$

where

$$A_t(\theta) = A_t + (\theta' \odot E_0)S_{t-1}$$

$$B_t(\theta) = [A_2 + (\theta' \odot E_0)] \cdot \epsilon_{t-1} \cdot (\theta'' \odot F_t),$$

$$U_t = \begin{bmatrix} 1 \\ \epsilon_{t-1} \end{bmatrix}$$

$$C_t(\theta) = \text{diag}(\theta''C_t),$$

$$D_t(\theta) = 0.$$

Here, representation of the dynamical model in Eq. (34) is similar to an ordinary linear system considering the above forms for $B_t(\theta)$ and $U_t$. Note that we keep $D_t(\theta)$ in Eq. (34) although its value is zero to derive formulation of VEKS in a more general form where the observation equation can be non-linear and depend on the input. Thus, the methodology introduced in this study to derive the VBEM method can be used for estimation of parameters of various non-linear systems in other medical or non-medical applications.

The linear system in Eq. (34) is a CE model if the proposed prior distributions of the model parameters in Eqs. (A-1)–(A-6) are exploited. Consequently, we can use Theorem 1 and Eq. (16) to implement the VB E-step for the dynamical system in Eq. (34). Furthermore, the dynamical system in Eq. (34) is a first-order Markov process. It can be shown that the VB E-step for a model based on Eq. (34) is equivalent to the following distribution for the hidden state $x_t$.

$$q_n(x_t) = p(x_t|x_1, (\Omega(\theta))) = N(x_t|x_t^1, \Gamma_t),$$

where $(\Omega(\theta)) = (\Omega(\theta)|x_0)$ is the expected natural parameters over the posterior distribution of the parameters and the mean $x_t$ as well as the covariance matrix $\Gamma_t$ are used to calculate sufficient statistics of the hidden state. A purpose of the standard Kalman smoother is to compute a marginal posterior distribution of state $x_t$, which is recursively calculated from the forward and backward recursions of VEKS, respectively. Algorithms 2 and 3 show a full implementation for the forward and backward recursions of VEKS, respectively.

For the local linear dynamical system in Eq. (34), as a first-order Markov process, it can be shown that

$$p(x_t|x_1, (\Omega(\theta))) \propto p(x_t|x_1, (\Omega(\theta))) \cdot p(y_1, \ldots, y_t|(\Omega(\theta))).$$

The VEKS algorithm is based on Eq. (40). First, a marginal posterior distribution $p(x_t|x_1, (\Omega(\theta)))$ is computed in the forward recursion of the VEKS algorithm where $\mu_t$ and $\Sigma_t$ are recursively calculated from $t = 1$ to $t = T$. Then backward recursion of the VEKS algorithm uses the computed values of $\mu_t$ and $\Sigma_t$ to recursively calculate $\eta_t$ and $\Psi_t$ from $t = T$ to $t = 1$. Because the proposed VEKS algorithm has a lot of mathematical formulations, details of the VEKS algorithm are presented in Appendix D. Flowcharts of the forward and backward recursions of the VEKS algorithm are illustrated in Figs. 1 and 2, respectively. Algorithms 2 and 3 show a full implementation for the forward and backward recursions of VEKS, respectively.

The computed $(\epsilon_t, \Gamma_t, \epsilon_{t-1}, \Gamma_{t-1}, \ldots)$, as output of Algorithm 3, are used in the VB M-step (Algorithm 1) to calculate the sufficient statistics of the hidden state and infer the posterior distributions of the parameters. Logarithm of probability of the data $\ln Z = \ln p(y_1, \ldots, y_T|x_0)$ is computed in the forward recursion of VEKS (Algorithm 2) and used to calculate a lower bound on the marginal likelihood of the model, $F(\cdot)$, in Appendix E.
### Algorithm 1. VB M-step for updating posterior distributions of parameters of a multi-area model.

- **Fixed parameters used in the algorithm:**
  - $y_{1:T}$ and $u_{1:T}$ as known observation and input of the multi-area model.
  - Fixed matrices $A_k$, $B_k$, $E_k$, and $C_k$ which specify the multi-area model in Eq. (18).
  - Fixed part of covariance matrix of state noise according to Eq. (20): $\Sigma^0_{t|t-1} = \text{diag}(\{1 \quad 0 \quad 0 \quad 0 \quad 0 \})$.
  - Hyperparameters of the prior according to Eqs. (A-1)-(A-6):
    $$
    \begin{align*}
    \Sigma^0_{u_{1:T} | y_{1:T}} & = \Sigma^0_{u_{1:T}} \Sigma^0_{y_{1:T}, u_{1:T}} \Sigma^0_{y_{1:T}} \Sigma^0_{u_{1:T}}, \\
    \mu^0_{u_{1:T} | y_{1:T}} & = \Sigma^0_{u_{1:T} | y_{1:T}}^{-1} \mu^0_{u_{1:T}}, \\
    b^0_{u_{1:T} | y_{1:T}} & = \mu^0_{u_{1:T}} \Sigma^0_{u_{1:T} | y_{1:T}}^{-1}, \\
    \end{align*}
    $$
    
- **Input of the algorithm:**
  - Mean $\gamma$, covariance matrix $\Gamma_{i,0}$, and cross-covariance matrix $\Gamma_{i,t}$, for all time steps. $[\gamma, \Gamma_{0}, \Gamma_{1,0}, \Gamma_{1,1}]_{t=1}^{T}$, are output of Algorithm 3 and computed in the VB E-step using the solution of VEKS.

- **Output of the algorithm:**
  - Hyperparameters of the posterior distributions of the parameters:
    $$
    \begin{align*}
    \Sigma^0_{u_{1:T} | y_{1:T}} & = \Sigma^0_{u_{1:T}} \Sigma^0_{y_{1:T}, u_{1:T}} \Sigma^0_{y_{1:T}} \Sigma^0_{u_{1:T}}, \\
    \mu^0_{u_{1:T} | y_{1:T}} & = \Sigma^0_{u_{1:T} | y_{1:T}}^{-1} \mu^0_{u_{1:T}}, \\
    b^0_{u_{1:T} | y_{1:T}} & = \mu^0_{u_{1:T}} \Sigma^0_{u_{1:T} | y_{1:T}}^{-1}, \\
    \end{align*}
    $$
  - $\Sigma^0_{u_{1:T} | y_{1:T}}$ and $\Sigma^0_{x_{1:T} | y_{1:T}} = \text{diag}(\{1 \quad 0 \quad 0 \quad 0 \quad 0 \})$ for $t = 1 : T$.

1. Compute matrices related to the local linearization of the sigmoid function in Eq. (C-4):
   - $\Sigma^0_{x_{1:T} | y_{1:T}}$ and $\Sigma^0_{x_{1:T} | y_{1:T}} = \text{diag}(\{1 \quad 0 \quad 0 \quad 0 \quad 0 \})$ for $t = 1 : T$.

2. Compute matrices $\Phi_{u_{1:T} | y_{1:T}}$, $\Phi_{y_{1:T} | x_{1:T}}$, $\Phi_{u_{1:T} | x_{1:T}}$, and $\Phi_{y_{1:T} | y_{1:T}}$ using Eqs. (C-11)-(C-16) to calculate the sufficient statistics of the hidden state, i.e., matrices $\varphi_{u_{1:T}}$, $\varphi_{y_{1:T}}$, $\varphi_{u_{1:T}}$, and $\varphi_{y_{1:T}}$, as well as vector $\varphi_{y_{1:T}}$ using Eqs. (C-18)-(C-23).

3. Compute matrices $R_{u_{1:T}}$, $R_{y_{1:T}}$, and $R_{u_{1:T}}$ using Eqs. (C-28)-(C-30).

4. Compute hyperparameters of distributions in Eqs. (B-6)-(B-11):
   - $\Sigma^0_{u_{1:T} | y_{1:T}}$, $\mu^0_{u_{1:T} | y_{1:T}}$, $b^0_{u_{1:T} | y_{1:T}}$, $\mu^0_{u_{1:T} | y_{1:T}}$, $b^0_{u_{1:T} | y_{1:T}}$, $\mu^0_{u_{1:T} | y_{1:T}}$, $b^0_{u_{1:T} | y_{1:T}}$, and $\mu^0_{u_{1:T} | y_{1:T}}$ for $n = 1 : N$.

5. Compute hyperparameters of posterior distributions of the parameters using the computed hyperparameters in the previous step and the hyperparameters of the prior distributions according to Eqs. (B-14)-(B-19):
   - $\Sigma^0_{u_{1:T} | y_{1:T}}$, $\mu^0_{u_{1:T} | y_{1:T}}$, $b^0_{u_{1:T} | y_{1:T}}$, $\mu^0_{u_{1:T} | y_{1:T}}$, $b^0_{u_{1:T} | y_{1:T}}$, and $\mu^0_{u_{1:T} | y_{1:T}}$ for $n = 1 : N$.

6. Output the computed hyperparameters of the posterior distributions of the parameters in the previous step.

3.6. Complete VBEM learning algorithm for multi-area model

A complete learning algorithm of the VBEM method for a multi-area E/MEG model is presented in Fig. 3 and Algorithm 4. The algorithm starts with initialization and consists of repeated iterations of a VB E-step, calculation of the lower bound $\mathcal{L}$, and a VB M-step. In initialization of the algorithm, the first step is to construct the multi-area model using activation detection of E/MEG data and establish a state-space representation of the model in Eq. (18). Consequently, matrices $A_k$, $B_k$, $E_k$, and $C_k$ are computed and the detected E/MEG signals in the cortical area of the model are used as the observation signals $y_{1:T}$ in the state-space model.

The next step in the initialization of Algorithm 4 is to generate an input vector $u_{1:T}$ using the profiles of the external stimuli. The final step is to initialize hyperparameters of the prior distributions of the model parameters, introduced in A-1-A-6, and hyperparameters $\mu_0$ and $\Sigma_0$ related to the prior distribution of the hidden state $x_0 \sim N(x_0 | \mu_0, \Sigma_0)$ at time $t = 0$. Further details about initialization of the algorithm are presented in Discussion section.

In the VB E-step of Algorithm 4, the free distributions over the hidden states $q_0(x_t) = p(x_t | y_{1:T}, (\Omega(t))) = N(x_t | y_{1:T}, \Gamma_{t,t})$, $t = 1, \ldots, T$.

### Algorithm 2. Forward recursion of VEKS for a multi-area model.

- **Fixed parameters used in the algorithm:**
  - $y_{1:T}$ and $u_{1:T}$ as known observation and input of the multi-area model.
  - Fixed matrices $A_k$, $B_k$, $E_k$, and $C_k$ which specify the multi-area model in Eq. (18).
  - Fixed part of covariance matrix of state noise according to Eq. (20): $\Sigma^0_{t|t-1} = \text{diag}(\{1 \quad 0 \quad 0 \quad 0 \quad 0 \})$.

- **Input of the algorithm:**
  - Hyperparameters of the posterior distributions of the parameters which are calculated in previous iteration of VB M-step according to Algorithm 1: $\Sigma_{t|t-1}$, $\mu_{t|t-1}$, $\alpha^0$, $\alpha^0$, and $\mu_{t|t-1}$, $\mu_{t|t-1}$, $\mu_{t|t-1}$, $\mu_{t|t-1}$, $\mu_{t|t-1}$, $\mu_{t|t-1}$, $\mu_{t|t-1}$, $\mu_{t|t-1}$, $\mu_{t|t-1}$, and $\mu_{t|t-1}$.

- **Output of the algorithm:**
  - Values of $(\mu_t, \Sigma_t)$, where $\mu_t$ and $\Sigma_t$ are mean and covariance matrix of Gaussian distribution $p(x_t | y_{1:t}, (\Omega(t))) = N(x_t | \mu_t, \Sigma_t)$, values of covariance matrix $(\Sigma_{t-1} | t-1)$ defined in Eq. (46), and logarithm of probability of the data
  $$
  \ln Z = \ln p(y_{1:T} | \pi_{0:t})
  $$

1. Initialize hyperparameters $\mu_0$ and $\Sigma_0$ as mean and variance of the hidden state $x_0$.

2. Compute time independent parameters in Eqs. (D-19)-(D-25) which will be used to calculate the expected natural parameters in Step 3(b).

3. For $t = 1$ to $T$
   - (a) Compute $\Sigma_{t-1} = \text{diag}(\{0 \quad 0 \quad 0 \quad 0 \quad 0 \})$ and $\gamma_t = \gamma_{t-1} - \gamma_t - \gamma_t$,
     $\mu_t$, according to Eq. (33) using the calculated mean value of the state, i.e., $\mu_{t-1}$, in previous time step of this algorithm according to Eq. (48).

   (b) Compute expected natural parameters in Eqs. (D-2)-(D-18) using the calculated values of parameters in Step 2 and the calculated values of $\gamma_t$, as well as $c_{t-1}$ in Step 3(a).

   (c) Compute input vector $U_t = \{ u_{t|t} \}$ according to Eq. (37).

4. Compute $\mu_t$ and covariance matrix $\Sigma_t$ for the marginal posterior distribution of the hidden state: $p(x_t | y_{1:T}, (\Omega(t))) = N(x_t | \mu_t, \Sigma_t)$
   $$
   \begin{align*}
   \Sigma_{t-1} & = \Sigma_{t-1} - \left( A_{t-1}^T A_{t-1} \right)^{-1}, \\
   \Sigma_t & = \left( A_{t-1}^T C_{t-1} \Sigma_{t-1} A_{t-1}^T \right) - \left( A_{t-1}^T A_{t-1} \Sigma_{t-1} A_{t-1}^T \right)^{-1}, \\
   \mu_t & = \left( A_{t-1}^T C_{t-1} \Sigma_{t-1} A_{t-1}^T \right) U_t + \left( A_{t-1}^T A_{t-1} \Sigma_{t-1} A_{t-1}^T \right) \mu_{t-1}, \\
   \end{align*}
   $$

5. Compute $\ln Z = \ln p(y_{1:T} | \pi_{0:t})$ and also
   $$
   \ln Z = \sum_{t=1}^{T} \ln z_t(\gamma_t).
   $$

End For.
are inferred. To this end, the forward and backward recursions of VEKS using Algorithms 2 and 3 are performed to \( \gamma_t \), \( \Gamma_{1-t} \). Here, the mean \( \gamma_t \) and the covariance matrix \( \Gamma_{1-t} \) represent the Gaussian distribution of the state vector \( x_t \) and \( x_{1-t} \) represents the cross-covariance matrix between \( x_{t-1} \) and \( x_t \). Computed \( \gamma_t \), \( \Gamma_{1-t} \), \( \Gamma_{1-t-1} \), are used to calculate the sufficient statistics of the hidden state in the VB-M-step of Algorithm 4 to infer the posterior distributions of the parameters. In addition, logarithm of probability of the data \( \ln Z = \ln p(y_{1:T}(0))_{t=0}^T \), which will be used to calculate the lower bound \( F(t) \), in Step 3 of Algorithm 4, is computed in the forward recursion of VEKS.

**Algorithm 3.** Backward recursion of VEKS for a multi-area model.

- **Fixed parameters used in the algorithm:**
  - \( y_{1:T} \) and \( u_{1:T} \) as known observation and input of the multi-area model
  - Fixed matrices \( A_1, A_2, E_1, E_2, C_1 \) which specify the multi-area model in Eq. (18)
  - Fixed part of covariance matrix of state noise according to Eq. (20): \( \Sigma_0 = \text{diag} \{ 10 1 0 1 1 0 0 \} \)

- **Input of the algorithm:**
  - Hyperparameters of posterior distributions of the parameters which are calculated in previous iteration of VB M-step according to Algorithm 1: \( \hat{\Sigma}_{\theta} \), \( \hat{\Sigma}_p \), \( \hat{\alpha}_t \), \( \hat{\beta}_t \), \( \hat{\gamma}_t \), \( \hat{\eta}_t \), \( \hat{\nu}_t \), \( \hat{b}_t \), \( \hat{\omega}_t \), and \( \hat{\epsilon}_t \).
  - Computed values of \( \mu_t, \Sigma_{t-1}, \Sigma_{t} \) from Algorithm 2

- **Output of the algorithm:**
  - Mean \( \gamma_t \) and covariance matrix \( \Gamma_t \) related to marginal posterior distribution \( p(x_t|y_{1:T}, (\Omega(\theta))) = N(x_t|\gamma_t, \Gamma_t) \) as well as cross-covariance matrix \( \Gamma_{1-t} \) for all time steps. Output values \( \gamma_t, \Gamma_{1-t}, \Gamma_{1-t-1} \), are used in Algorithm 1 to calculate sufficient statistics of the hidden state.

1. Initialize \( \Psi_1 = 0 \) to satisfy the end condition \( p(y_{1:T} | x_T, (\Omega(\theta))) = 1 \)
2. Compute time independent parameters in Eqs. (D-19)-(D-25) which will be used to calculate expected natural parameters in Step 3-(b)
3. For \( t = T \) to 1
   (a) Compute \( S_{t-1} = \text{diag} \{ \mu_{t-1} \} \) and \( e_{t-1} = S_{t-1} \mu_{t-1} \), \( \mu_{t-1} \), according to Eq. (33) using mean value of state, i.e., \( \mu^\theta, \) which is computed in Algorithm 2
   (b) Compute expected natural parameters in Eqs. (D-2)-(D-18) using the calculated values of parameters in Step 3-(a) and the calculated values of \( S_{t-1} \) as well as \( e_{t-1} \) in Step 3-(a)
   (c) Compute input vector \( U_t = | 1 | u_{t-1} | m \) according to Eq. (37)
   (d) Perform backward recursion to compute \( \eta_{t-1} \) and covariance matrix \( \Psi_{t-1} \) of Gaussian distribution \( p(y_{1:T} | x_{t-1}, (\Omega(\theta))) = N(x_{t-1}|\eta_{t-1}, \Psi_{t-1}) \)

\[
\Psi_t = ((A_t^0 - C_0^0 A_1 C_1) + \Psi_{t-1})^{-1},
\]
\[
\Psi_{t-1} = \left( (A_t^0 - C_0^0 A_1 C_1) - (A_t^0 - C_0^0 A_1 C_1) \Psi_{t-1} (A_t^0 - C_0^0 A_1 C_1) \right)^{-1},
\]
\[
\eta_{t-1} = \Psi_{t-1} \left[ - (A_t^0 - C_0^0 A_1 C_1) U_t + (A_t^0 - C_0^0 A_1 C_1) \Psi_{t-1} \right]
\times \left[ (A_t^0 - C_0^0 A_1 C_1) U_t + (A_t^0 - C_0^0 A_1 C_1) \Psi_{t-1} \right],
\]
(e) Compute mean \( \gamma_t \) and covariance matrix \( \Gamma_t \) of Gaussian distribution \( p(x_t|y_{1:T}, (\Omega(\theta))) = N(x_t|\gamma_t, \Gamma_t) \)
\[
\Gamma_t = \left( \Sigma_t^{-1} + \Psi_t^{-1} \right)^{-1},
\]
\[
\gamma_t = \Gamma_t \left( \Sigma_t^{-1} \mu_t + \Psi_t^{-1} \eta_t \right).
\]
(f) Compute cross-covariance matrix \( \Gamma_{1-t} \)
\[
\Gamma_{1-t} = \Sigma_{t-1} \left( A_t^{-1} \Gamma_t A_t^{-1} \right)^T \Gamma_{1-t} \]

End For.
4. Output \( \gamma_t, \Gamma_{1-t}, \) and \( \Gamma_{1-t-1} \) for \( t = 1, \ldots, T \)

**Algorithm 4.** Complete learning algorithm of the VBEM method for a multi-area E/MEG model.

1. Initialization
   (a) Activation detection of the E/MEG data to construct the multi-area model
   I. Determine number of active cortical areas
   II. Assign the detected E/MEG signals in the cortical areas to observation vector \( y_{1:T} \)
   III. Compute matrices \( A_1, A_2, E_1, E_2, C_1 \) using a lead field matrix and the intrinsic parameters of the multi-area model
   (b) Assign appropriate signal to input vector \( u_{1:T} \) based on profiles of the external stimuli
   (c) Initialize hyperparameters of the prior distributions of the model parameters in Eqs. (A-1)-(A-6): \( \Sigma_{\theta}^{0}, \Sigma_p^{0}, \alpha_\theta^{0}, \alpha_\theta^{0} \), and \( \mu_0, \mu_{0,t}, b_0^1, b_0^2, \mu_0, \mu_{0,t}, b_0^1, b_0^2 \)
   (d) Initialize hyperparameters \( \mu_0 \) and \( \Sigma_0 \) as mean and covariance matrix of prior distribution of the hidden state \( x_0 \sim N(x_0|0, \Sigma_0) \) at time \( t = 0 \)
2. Perform VB E-step using variational extended Kalman smoother (VEKS) to infer \( \{ \gamma_t, \Gamma_{1-t}, \Gamma_{1-t-1} \} \) and logarithm of probability of data \( \ln Z = \ln p(y_{1:T}(0))_{t=0}^T \)
   (a) Perform forward recursion of VEKS using Algorithm 2
   (b) Perform backward recursion of VEKS using Algorithm 3
3. Compute lower bound on the marginal likelihood of the model, \( F(t) \)
   (a) Compute right hand sides of Eqs. (E-10)-(E-14) using the hyperparameters of the posterior and prior distributions of the model parameters which are calculated in Step 2 and initialized in Step 1, respectively
   (b) Compute \( F(t) \) based on Eq. (E-9) using the calculated terms in Step 3-(a) and the calculated \( \ln Z \) in Step 2
4. Perform VB M-step using Algorithm 1 to update hyperparameters of the posterior distributions of model parameters
5. While the lower bound \( F(t) \) is increasing, go to Step 2
6. Output the following computed quantities
   (a) Hyperparameters of the posterior distributions of the parameters of the model
   (b) Lower bound on the marginal likelihood of the model, \( F(t) \)
   (c) Mean \( \gamma_t \) and covariance matrix \( \Gamma_t \) of the hidden state vector for all time steps, i.e., \( \{ \gamma_t, \Gamma_{1-t}, \Gamma_{1-t-1} \} \), which represent the distribution over the hidden state: \( q_t(x_t) = p(x_t|y_{1:T}, (\Omega(\theta))) = N(x_t|\gamma_t, \Gamma_t), \ t = 1, \ldots, T \)
The lower bound on the marginal likelihood of the model, $F(\cdot)$, is inferred in the next step of Algorithm 4. Hyperparameters of the prior and posterior distributions of the parameters, from Step 1 and Step 4, as well as the computed $\ln Z$ in Step 2 are used to compute the lower bound $F(\cdot)$ according to Eqs. E-9, E-10, E-11, E-12, E-13, E-14. In the VB M-step of Algorithm 4, hyperparameters of the posterior distributions of the model parameters are updated using Algorithm 1. Repeated iterations of the VB E-step, calculation of $F(\cdot)$, and the VB M-step continue until change in $F(\cdot)$ between successive iterations of VBEM is less than a small threshold. After this iteration, $F(\cdot)$ saturates, meaning that the lower bound on the model likelihood will not increase much if we continue the iterations for further learning of the posterior distributions of the model parameters.

4. Summary and conclusion

We have developed a variational Bayesian expectation maximization (VBEM) method for model inversion of our multi-area extended neural mass model (MEN) using EEG and MEG signals. To benefit from simplification of the CE model, prior distributions of the model parameters have to satisfy Condition 1. If rows of matrices $\theta^d$ and $\theta^o$ as well as elements of vector $\theta^v$ have Gaussian distributions and precision vectors $\rho^d$ as well as $\rho^v$ have Gamma distributions, Eq. (14) of Condition 1 will be satisfied. We use the following prior distributions of the model parameters that generate a CE model.

$$p(\theta|m) = p(\theta^d, \theta^o, \theta^v; \rho^d, \rho^v|m) = \prod_{n=1}^{N} p(\theta^d_n|m, \rho^d_n)p(\theta^o_n|m, \rho^o_n)p(\theta^v_n|m, \rho^v_n),$$

where

$$p(\theta^d_n|m, \rho^d_n) = N\left(\theta^d_n|m^0_{\theta^d_n|m, \rho^d_n}, \rho^{-1}_{\theta^d_n} \Sigma^0_{\theta^d_n}\right), \quad n = 1, \ldots, N,$$

$$p(\theta^o_n|m, \rho^o_n) = N\left(\theta^o_n|m^0_{\theta^o_n|m, \rho^o_n}, \rho^{-1}_{\theta^o_n} \Sigma^0_{\theta^o_n}\right), \quad n = 1, \ldots, N,$$

$$p(\theta^v_n|m, \rho^v_n) = \text{Gamma}\left(\theta^v_n|\alpha^0_{\theta^v_n|m, \rho^v_n}, \beta^0_{\theta^v_n|m, \rho^v_n}\right), \quad n = 1, \ldots, N,$$

$$p(\rho^d_n) = \text{Gamma}\left(\rho^d_n|\alpha^0_{\rho^d_n}, \beta^0_{\rho^d_n}\right), \quad n = 1, \ldots, N,$$

$$p(\rho^o_n) = \text{Gamma}\left(\rho^o_n|\alpha^0_{\rho^o_n}, \beta^0_{\rho^o_n}\right), \quad n = 1, \ldots, N,$$

$$p(\rho^v_n) = \text{Gamma}\left(\rho^v_n|\alpha^0_{\rho^v_n}, \beta^0_{\rho^v_n}\right), \quad n = 1, \ldots, N.$$
and the Gamma distribution in the above equations is defined as
\[
\text{Gamma}(\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot x^{\alpha-1} \cdot e^{-\beta x}, \tag{A-7}
\]
where the shape parameter $\alpha$ and the inverse scale parameter $\beta$ represent the Gamma distribution. Here, $N$ is the number of cortical areas in the multi-area model, row vectors $\beta_n^t$ and $\rho_n^t$ are $n$th rows of the corresponding matrices, scalars $\alpha_n^t$, $\phi_n^t$, and $\rho_n^t$ are the $n$th elements of the vectors $\beta^t$, $\phi^t$ and $\rho^t$, respectively. Here, the prior distributions are stated using the hyperparameters $\mu, \Sigma, a, b$ with the specified superscripts and subscripts.

**Appendix B. Distributions of model parameters using VB M-step**

Free distributions over the model parameters are updated in the VB M-step of the VBEM algorithm. We derive the VB M-step formulation based on Theorem 1. We illustrate that if we use the proposed prior distributions of the parameters in Eqs. (A-1)-(A-6), the multi-area model will be a CE model. In the VB M-step based on Theorem 1 and Eq. (12), posterior distributions of the model parameters are constructed from two terms: the prior distributions of the parameters, i.e., $p(\theta|m)$, and an exponential term. We start to calculate the exponential term which can be written in the following form using the joint probability in Eq. (26).

**Algorithm 3**

1. **Initialize** $\Psi_t^{-1} = 0$
2. **Compute the time independent parameters** in Eqs. (D-19)-(D-25)
3. For $t = T$ to 1
   - **Compute** $S^t_{ai} = \text{diag}(\partial S(x)/\partial x|_{x=\mu_{ai}})$
   - $\sigma_{ai} = S(\mu_{ai}) - S^t_{ai} \cdot \mu_{ai}$
   - **Compute** the expected natural parameters using Eqs. (D)-18)
   - **Compute** $\eta_{ai}$ and $\Psi_{ai}$ using Eqs. (51)-53)
   - **Compute** $\Gamma_{ai}, \gamma_{ai}$, and $\Gamma_{aii}$ using Eqs. (54)-(56)
4. If $t = T - 1$, go to 5.
5. Output: $\{\gamma_{ai}, \Gamma_{ai}, \Gamma_{aii}\}_{ai=1}^n$

**Fig. 2.** Flowchart of the backward recursion of VEKS for a multi-area model. Algorithm 3 provides further details for this figure. The dashed green boxes show inputs of the algorithm. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

\[
\exp(\ln p(x_i, y_i|\theta, m) q_{\theta}^{k+1}(\theta)) = p(x_i|m) \exp \left( \sum_{j=1}^T \ln p(x_{i,j-1}, \theta, m) q_{\theta}^{k+1}(\theta) \right) \\
\quad \cdot \exp \left( \sum_{j=1}^T \ln p(y_i|x_{i,j}, \theta, m) q_{\theta}^{k+1}(\theta) \right), \tag{B-1}
\]

where $p(x_i|m) = 1$ and the free distribution over the hidden state is Gaussian distributed.

\[
q_{\theta}^{k+1}(x_i) = N(x_i|\gamma_{ai}, \Gamma_{ai}), \quad t = 1, \ldots, T \tag{B-2}
\]

and sufficient statistics of the hidden state are calculated using the mean $\gamma_{ai}$, the covariance matrix $\Gamma_{ai}$, and the cross-covariance matrix $\Gamma_{aii}$, estimated in the ($k+1$)th iteration of the VB E-step using VEKS. In Appendix C, we use a local linearization of the model to calculate RHS of Eq. (11) and show that

\[
\sum_{i=1}^T \ln p(x_{i,j-1}, \theta, m)_{q_{\theta}^{k}(\theta)} = -\frac{1}{2} \sum_{n=1}^N p_n^t \left( \phi_n^t \psi_n^t - 2(\phi_n^t \psi_n^t + \psi_n(n-1)) \phi_n^t \\
+ \phi_n^t \psi_n^t + 2\phi_n(n-1) \phi_n^t + \psi_n(n) \right) \\
+ \frac{T}{2} \sum_{n=1}^N \ln \rho_n^t - \frac{T}{2} \ln \left( |A_n| \right) - \frac{8TN}{2} \ln(2\pi). \tag{B-3}
\]
Initialization
- Construct the multi-area model
- Initialize hyperparameters of the prior distributions

```
VB-E step (VEKS)
- Compute \( \{y_t, \Gamma_{t;}, \Gamma_{t;+1;} \}^T \)
- Compute \( \ln Z = \langle \ln p(y_t|\theta) \rangle_{q(\theta)} \)

Algorithm 2: Forward recursion of VEKS
(See Fig. 1)

Algorithm 3: Backward recursion of VEKS
(See Fig. 2)
```

Compute lower bound \( F(\cdot,\cdot) \)
using Eqs. (E-9)-(E-14)

```
VB-M step
Algorithm 1: Update hyperparameters of the posterior distributions of parameters
```

- Hyperparameters of the posterior distributions of the parameters
- Lower bound \( F(\cdot,\cdot) \)
- Mean \( y_t \) and covariance matrix \( \Gamma_{t;} \) of the hidden state vector for all time steps, \( \{y_t, \Gamma_{t;} \}^T \)

**Output of the VBEM algorithm:**
- Hyperparameters of the posterior distributions of the parameters
- Lower bound \( F(\cdot,\cdot) \)
- Mean \( y_t \) and covariance matrix \( \Gamma_{t;} \) of the hidden state vector for all time steps, \( \{y_t, \Gamma_{t;} \}^T \)

Fig. 3. Flowchart of the complete learning algorithm of the VBEM method for a multi-area E/MEG model. Algorithm 4 provides further details for this figure.

\[
\left\langle \sum_{k=1}^{T} \ln p(y_t|X, \theta, m) \right\rangle_{q_{\theta}(\theta)} = -\frac{1}{2} \sum_{n=1}^{N} \left( \rho_{n}^{2}(\theta_{n}^{2} R_{xx}(n, n) - 2 \theta_{n}^{2} R_{xy}(n, n) + R_{yy}(n, n) \right) \right. 
- T \ln \rho_{n}^{2} + \frac{N \cdot T}{2} \ln(2 \pi). 
\]

(B-4)

Here, for the sake of simplicity, we use the notation \( q_{\theta}(X) \) instead of \( q_{\theta}^{k+1}(X) \) in the above equations. The matrices \( R_{xx}, R_{xy}, R_{yy} \) as well as the vector \( \psi_{\alpha} \) are sufficient statistics of the hidden state calculated using \( \{y_t, \Gamma_{t;}, \Gamma_{t;+1;} \}^T \). In Appendix C, the sufficient statistics of the hidden state are calculated in Eqs. (C-18)-(C-23) and (C-28)-(C-30).

Considering Eqs. (B-1)-(B-4), we have:
\[
\exp(\ln p(x|y, \theta, m)_{q_{\theta}(\theta)}) = Z_{0} q_{\theta}^1(\theta) = q_{\theta}^1(\theta^{*}, \rho^*) q_{\theta}^1(\theta^{*}, \rho^*) q_{\theta}^1(\theta^{*}, \rho^*) q_{\theta}^1(\theta^{*}, \rho^*)
\]
\[
\times Z_{0} \prod_{n=1}^{N} q_{\theta}^1 \left( d_{n} | \rho_{n}^{*} \right) q_{\theta}^1 \left( \psi_{\alpha}^{n} | \rho_{n}^{*} \right) q_{\theta}^1 \left( \psi_{\alpha}^{n} | \rho_{n}^{*} \right) q_{\theta}^1 \left( \psi_{\alpha}^{n} | \rho_{n}^{*} \right).
\]

(B-5)

where \( Z_{0} \) is a normalization constant and the distributions in Eq. (B-5) are:
\[
\left\{ q_{\theta}^1(\theta^{*}, \rho^*) = N(0, \mu_{\theta}^1 | \mu_{\theta}^1, \rho^*) \right. 
\times \Sigma_{\theta}^{-1}(\theta^{*}, \rho^*) \}
\]
\[
\left. \Sigma_{\theta}^{-1}(\theta^{*}, \rho^*) = \psi_{\alpha}^1 \right\}
\]
\[
\left\{ h_{\theta}^{1}(\theta^{*}, \rho^*) = \left[ \psi_{\alpha}^1 + \psi_{\alpha}^1(n, :) \right] \Sigma_{\theta}^{-1}(\theta^{*}, \rho^*) \right\}.
\]

(B-6)
\begin{equation}
\begin{aligned}
q^1_b(p_b^2) & = N\left(p_b^2, \mu_{\omega_{B_b^2}}^{-1}, \Sigma_{\omega_{B_b^2}}\right), \\
\Sigma_{\omega_{B_b^2}} & = \left(\psi_{0} - \psi_{\omega_{B_b^2}}\right)^{-1}, \\
\mu_{\omega_{B_b^2}} & = \left[\psi_{\omega_{B_b^2}} \psi_{\omega_{B_b^2}} + \psi_{\omega_{B_b^2}}\right]^{-1}.
\end{aligned}
\end{equation}

(B-7)

\begin{equation}
\begin{aligned}
q^1_f(q^2_s) & = \text{Gamma}\left(p_s^{a^1}, b_s^{a^1}\right), \\
a^1 & = a^0 + 1, \\
b_s^{a^1} & = b_s^{a^0} + 1.
\end{aligned}
\end{equation}

(B-8)

\begin{equation}
\begin{aligned}
q^1_c(q^2_k) & = \text{Gamma}\left(p_k^{a^2}, b_k^{a^2}\right), \\
a^2 & = a^0 + 1, \\
b_k^{a^2} & = b_k^{a^0} + 1.
\end{aligned}
\end{equation}

(B-9)

\begin{equation}
\begin{aligned}
q^1_0(q^2) & = N\left(q^2, \mu_{\omega_{B_0^2}}^{-1}, \Sigma_{\omega_{B_0^2}}\right), \\
\Sigma_{\omega_{B_0^2}} & = \psi_{\omega_{B_0^2}}^{-1}, \\
\mu_{\omega_{B_0^2}} & = \left[\psi_{\omega_{B_0^2}} + \psi_{\omega_{B_0^2}}\right]^{-1}.
\end{aligned}
\end{equation}

(B-10)

\begin{equation}
\begin{aligned}
q^1_0(q^2_1) & = \text{Gamma}\left(p_1^{a^3}, b_1^{a^3}\right), \\
a^3 & = a^0 + 1, \\
b_1^{a^3} & = b_1^{a^0} + 1.
\end{aligned}
\end{equation}

(B-11)

The marginal posterior distribution for \( \theta_s \) is given by:

\begin{equation}
\begin{aligned}
q^1_\theta(q^2_\theta) & = N\left(q^2_\theta, \mu_{\omega_{B_\theta^2}}^{-1}, \Sigma_{\omega_{B_\theta^2}}\right), \\
\Sigma_{\omega_{B_\theta^2}} & = \psi_{\omega_{B_\theta^2}}^{-1}, \\
\mu_{\omega_{B_\theta^2}} & = \left[\psi_{\omega_{B_\theta^2}} + \psi_{\omega_{B_\theta^2}}\right]^{-1}.
\end{aligned}
\end{equation}

(B-12)

Appendix C. Calculation of exponential terms in VB M-step

The sequence of the state vector \( x \) in the state Eq. (18) follows a first-order Markov process. Logarithm of the joint probability of the sequence of states and observations of dynamical system in Eq. (18) is therefore given by:

\begin{equation}
\begin{aligned}
\log p(x, y; \theta, \varphi, \omega) & = \log p(x_0, y_1; \theta, \varphi, \omega) \\
& + \sum_{t=1}^{T} \log p(x_t | x_{t-1}, \theta, \varphi, \omega) + \log p(y_t | x_t, \theta, \varphi, \omega).
\end{aligned}
\end{equation}

(C-1)

Note that the covariance matrices of each row of \( \theta, \varphi, \omega \) matrices are constructed from two parts: the first part, stated by \( \Sigma \) with different subscripts and superscripts in the above equations, is the same for all rows and only the second part, stated by \( \psi \) varies w.r.t. the row numbers. Algorithm 1 illustrates the VB M-step for updating the posterior distributions of the parameters of the multi-area model.

The following linear state-space representation is derived from NDS in Eq. (18) using a linear approximation of the sigmoid function.

\begin{equation}
\begin{aligned}
x_t & = A x_{t-1} + \beta (\omega \otimes E) S_t - 1 + (\omega \otimes E) u_t + w_t, \\
y_t & = \text{diag}(\varphi) \cdot C_j x_t + v_t \quad t = 1, \ldots, T.
\end{aligned}
\end{equation}

(C-5)

Because the state noise and the prior over the hidden state at time \( t = 0 \) in the state Eq. (C-5) are Gaussian distributed, \( p(x_t | x_{t-1}, \theta, \varphi, \omega) \) is Gaussian distributed and we have

\begin{equation}
\begin{aligned}
q^1_\theta(q^2_\theta) & = \text{Gamma}\left(p_\theta^{a^4}, b\theta^{a^4}\right), \\
as & = a^0 + 1, \\
b\theta^{a^4} & = b\theta^{a^0} + 1.
\end{aligned}
\end{equation}

(B-13)

\begin{equation}
\begin{aligned}
q^1_s(q^2_s) & = \text{Gamma}\left(p_s^{a^5}, b_s^{a^5}\right), \\
a^5 & = a^0 + 1, \\
b_s^{a^5} & = b_s^{a^0} + 1.
\end{aligned}
\end{equation}

(B-14)

\begin{equation}
\begin{aligned}
q^1_c(q^2_k) & = \text{Gamma}\left(p_k^{a^6}, b_k^{a^6}\right), \\
a^6 & = a^0 + 1, \\
b_k^{a^6} & = b_k^{a^0} + 1.
\end{aligned}
\end{equation}

(B-15)

\begin{equation}
\begin{aligned}
q^1_\omega(q^2_\omega) & = \text{Gamma}\left(p_\omega^{a^7}, b_\omega^{a^7}\right), \\
a^7 & = a^0 + 1, \\
b_\omega^{a^7} & = b_\omega^{a^0} + 1.
\end{aligned}
\end{equation}

(B-16)
\[
\ln p(x_t | x_{t-1}, \theta, m) = -\frac{1}{2} \left( A_t x_{t-1} + (\theta^T \otimes E_0) \hat{S}_{t-1} + (\theta^T \otimes E_1) u_t - x_t \right)^T \times A_t^{-1} \left( A_t x_{t-1} + (\theta^T \otimes E_0) \hat{S}_{t-1} + (\theta^T \otimes E_1) u_t - x_t \right) - \frac{1}{2} \ln |A_t| - \frac{8N}{2} \ln (2\pi).
\]

(C-6)

Considering definition \( F_t \equiv A_t x_{t-1} + A_t \hat{S}_{t-1} - x_t \), using cyclic permutations of the trace operator, and considering decomposition \( A_t = A_t^T A_t^T \) based on Eq. (20), we have

\[
\ln p(x_t | x_{t-1}, \theta, m) = -\frac{1}{2} \text{trace} \left\{ \left( (\theta^T \otimes E_0) \hat{S}_{t-1} + (\theta^T \otimes E_1) u_t + F_t \right) A_t^{-1} \times \left( (\theta^T \otimes E_0) \hat{S}_{t-1} + (\theta^T \otimes E_1) u_t + F_t \right) \right\} - \frac{1}{2} \ln |A_t| - \frac{8N}{2} \ln (2\pi).
\]

(C-7)

\[
\ln p(x_t | x_{t-1}, \theta, m) = -\frac{1}{2} \text{trace} \left\{ \left( (\theta^T \otimes E_0^T) A_t^{-1} (\theta^T \otimes E_0), \hat{S}_{t-1}, \hat{S}_{t-1}^T \right) \right\} + 2 \left( (\theta^T \otimes E_0^T) A_t^{-1} (\theta^T \otimes E_1), u_t^T \right) - \frac{1}{2} \text{trace} \left\{ \left( (\theta^T \otimes E_0^T) A_t^{-1} (\theta^T \otimes E_1), u_t^T \right) \right\} - \frac{1}{2} \ln |A_t|^T - \frac{8N}{2} \ln (2\pi).
\]

(C-8)

where the mixed-product property of the Kronecker product, i.e., \((A \otimes B) \cdot (C \otimes D) = (A \otimes C) \cdot (B \otimes D)\), is used in the above equations.

We can now calculate the following expectation over the distribution of the hidden state calculated in the previous iteration of the VB E-step, i.e., \( q_{\theta^T}(x) \). For the sake of simplicity, we use the notation \( q_{\theta^T}(x) \) instead of \( q_{\theta^T}(x) \) in the following equations.

\[
\left\{ \sum_{t=1}^{N} \ln p(x_t | x_{t-1}, \theta, m) \right\}_{q_{\theta^T}} = -\frac{1}{2} \text{trace} \left\{ \left( (\theta^T \otimes E_0^T) (\theta^T A_t^{-1} (\theta^T \otimes E_1)) \right) \right\} + 2 \left( (\theta^T \otimes E_0^T) A_t^{-1} (\theta^T \otimes E_1), u_t^T \right) - \frac{1}{2} \ln |A_t|^T - \frac{8N}{2} \ln (2\pi).
\]

(C-9)

Based on properties of the Kronecker product in Appendix F, the following equation can be derived from Eq. (C-10).

\[
\left\{ \sum_{t=1}^{N} \ln p(x_t | x_{t-1}, \theta, m) \right\}_{q_{\theta^T}} = -\frac{1}{2} \sum_{n=1}^{N} \rho_n^T \left( \theta_n^T \varphi_n \theta_n^T - 2 (\theta_n^T \varphi_n + \phi_n (n))) \theta_n^T \right) + \frac{1}{2} \sum_{n=1}^{N} \ln \rho_n^T - \frac{1}{2} \ln \left( |A_t|^T \right) - \frac{8N}{2} \ln (2\pi).
\]

(C-17)
where

\[ \psi_{ij}(i,j) = \text{trace} \left( E_{ij}^T \Phi_{ij}(i,j) \right), \quad i = 1, \ldots, N; \quad j = 1, \ldots, N. \]  

(C-18)

\[ \psi_{ii}(i,j) = \text{trace} \left( E_{jj}^T \Phi_{ii}(i,i) \right), \quad i = 1, \ldots, M; \quad j = 1, \ldots, N. \]  

(C-19)

\[ \psi_{ij}(i,j) = \text{trace} \left( E_{ij}^T \Phi_{ij}(i,i) \right), \quad i = 1, \ldots, N; \quad j = 1, \ldots, N. \]  

(C-20)

\[ \psi_{ij}(i,j) = \text{trace} \left( E_{ij}^T \Phi_{ij}(i,j) \right), \quad i = 1, \ldots, M; \quad j = 1, \ldots, M. \]  

(C-21)

\[ \psi_{ii}(i,j) = \text{trace} \left( E_{ij}^T \Phi_{ii}(i,i) \right), \quad i = 1, \ldots, M; \quad j = 1, \ldots, M. \]  

(C-22)

\[ \psi_i(i,i) = \text{trace} \left( \Phi_{ii}(i,i) \right), \quad i = 1, \ldots, N. \]  

(C-23)

Here, the scalar \( \rho_i^n \) is the \( n \)th entry of the precision vector \( \rho \) of the state noise according to Eq. (20) and the row vectors \( \delta_i^n \) and \( \theta_i^n \) are the \( n \)th rows of \( \delta \) and \( \theta \) matrices, respectively. The 8-by-1 vectors \( l \) and \( j \) are defined as \( l = 8(i - 1) + 1:8i \) and \( j = 8(j - 1) + 1:8j \), respectively.

For the observation equation of the dynamical system in Eq. (18), we can derive the following equations for the hidden state vector.

\[ \ln p(y_t|x_t, \theta, m) = -\frac{1}{2} \left[ y_t - \text{diag}(\theta) \cdot C_t \cdot x_t \right]^T \left[ y_t - \text{diag}(\theta) \cdot C_t \cdot x_t \right] \]

\[ -\frac{1}{2} \ln |A_t| - \frac{N}{2} \ln (2\pi), \]

\[ \text{diagonal}(\theta) = \frac{1}{2} \left[ \text{diag}(\theta^T) \cdot C_t \cdot x_t \right]^T \left[ \text{diag}(\theta^T) \cdot C_t \cdot x_t \right] \]

\[ +2\theta_t^2 \cdot x_t^2 - 2\theta_t \cdot \text{diag}(\theta) \cdot C_t \cdot x_t \]

\[ -\frac{1}{2} \ln |A_t| - \frac{N}{2} \ln (2\pi), \]

\[ \text{trace}(\theta) = \frac{1}{2} \text{trace} \left( \text{diag}(\theta^T) \cdot C_t \cdot x_t \cdot C_t^T \right) \]

\[ +2\theta_t \cdot y_t^2 - 2\theta_t \cdot \text{diag}(\theta) \cdot C_t \cdot x_t \]

\[ -\frac{1}{2} \ln |A_t| - \frac{N}{2} \ln (2\pi). \]

The following expectation over distribution of the hidden state can be derived from Eq. C-26.

\[ \langle \ln p(y_t|x_t, \theta, m) \rangle_{q(\theta)} = -\frac{1}{2} \text{trace} \left( \text{diag}(\theta^T) \cdot A_t^{-1} \right) \left[ \text{diag}(\theta^T) \cdot R_{\theta_x} \right] \]

\[ +\text{trace} \left( \text{diag}(\theta^T) \cdot A_t^{-1} \right) \left[ \text{diag}(\theta^T) \cdot R_{\theta_y} \right] \]

\[ +\frac{T}{2} \sum_{n=1}^{N} \ln \rho_n^2 - \frac{T}{2} \ln (2\pi). \]

\[ \text{Eq. C-27} \]

where

\[ R_{\theta_x} = \sum_{t=1}^{T} C_t x_t^2 \]

\[ R_{\theta_y} = \sum_{t=1}^{T} C_t y_t^2 \]

and Eq. C-27 can be written as

\[ \langle \ln p(y_t|x_t, \theta, m) \rangle_{q(\theta)} = -\frac{1}{2} \sum_{n=1}^{N} \left( \rho_n^2 \cdot \text{trace} \left( \text{diag}(\theta^T) \cdot R_{\theta_x} \right) \right) \]

\[ -2\theta_n^2 \cdot y_t^2 (n, n) + R_{\theta_y} \cdot y_t^2 (n, n) \]

\[ -T \ln \rho_n^2 \]

\[ -\frac{N}{2} \ln (2\pi). \]

(C-31)

where scalars \( \theta_n^2 \) and \( \rho_n^2 \) are the \( n \)th entries of the vector \( \theta \) and the precision vector \( \rho \) in Eq. (20), respectively.

**Appendix D. Forward and backward recursions of VEKS for multi-area model**

Before dealing with the forward and backward recursions of the VEKS algorithm, we first calculate the expected natural parameters \( \Omega(\theta) \) which are computed at the end of the VB M-step. The relevant natural parameters of the dynamical system in Eq. (34) are

\[ \Omega(\theta) = \Omega(A_t), B_t, C_t, D_t, A_t \]

\[ \Omega(A_t) = A_t + \gamma_t \cdot B_t, \]

\[ \Omega(B_t) = (A_t^{-1} + \gamma_t), \]

\[ \Omega(C_t) = C_t \]

\[ \Omega(D_t) = D_t \]

The expected natural parameters \( \Omega(\theta) \) contain the following terms.

\[ \langle A_t \rangle = A_t + (A_{t+1} + \gamma_{t+1}) \]

\[ \langle B_t \rangle = (A_t^{-1} + \gamma_t) \]

\[ \langle C_t \rangle = C_t \]

\[ \langle D_t \rangle = D_t \]

\[ \langle A_t \rangle = A_t + \gamma_t \cdot B_t, \]

\[ \langle B_t \rangle = (A_t^{-1} + \gamma_t), \]

\[ \langle C_t \rangle = C_t \]

\[ \langle D_t \rangle = D_t \]

\[ \langle A_t \rangle = A_t + \gamma_t \cdot B_t, \]

\[ \langle B_t \rangle = (A_t^{-1} + \gamma_t), \]

\[ \langle C_t \rangle = C_t \]

\[ \langle D_t \rangle = D_t \]

\[ \langle A_t \rangle = A_t + \gamma_t \cdot B_t, \]

\[ \langle B_t \rangle = (A_t^{-1} + \gamma_t), \]

\[ \langle C_t \rangle = C_t \]

\[ \langle D_t \rangle = D_t \]

\[ \langle A_t \rangle = A_t + \gamma_t \cdot B_t, \]

\[ \langle B_t \rangle = (A_t^{-1} + \gamma_t), \]

\[ \langle C_t \rangle = C_t \]

\[ \langle D_t \rangle = D_t \]

\[ \langle A_t \rangle = A_t + \gamma_t \cdot B_t, \]

\[ \langle B_t \rangle = (A_t^{-1} + \gamma_t), \]

\[ \langle C_t \rangle = C_t \]

\[ \langle D_t \rangle = D_t \]

\[ \langle A_t \rangle = A_t + \gamma_t \cdot B_t, \]

\[ \langle B_t \rangle = (A_t^{-1} + \gamma_t), \]

\[ \langle C_t \rangle = C_t \]

\[ \langle D_t \rangle = D_t \]

\[ \langle A_t \rangle = A_t + \gamma_t \cdot B_t, \]

\[ \langle B_t \rangle = (A_t^{-1} + \gamma_t), \]

\[ \langle C_t \rangle = C_t \]

\[ \langle D_t \rangle = D_t \]

\[ \langle A_t \rangle = A_t + \gamma_t \cdot B_t, \]

\[ \langle B_t \rangle = (A_t^{-1} + \gamma_t), \]

\[ \langle C_t \rangle = C_t \]

\[ \langle D_t \rangle = D_t \]

\[ \langle A_t \rangle = A_t + \gamma_t \cdot B_t, \]

\[ \langle B_t \rangle = (A_t^{-1} + \gamma_t), \]

\[ \langle C_t \rangle = C_t \]

\[ \langle D_t \rangle = D_t \]

\[ \langle A_t \rangle = A_t + \gamma_t \cdot B_t, \]

\[ \langle B_t \rangle = (A_t^{-1} + \gamma_t), \]

\[ \langle C_t \rangle = C_t \]

\[ \langle D_t \rangle = D_t \]

\[ \langle A_t \rangle = A_t + \gamma_t \cdot B_t, \]

\[ \langle B_t \rangle = (A_t^{-1} + \gamma_t), \]

\[ \langle C_t \rangle = C_t \]

\[ \langle D_t \rangle = D_t \]

\[ \langle A_t \rangle = A_t + \gamma_t \cdot B_t, \]

\[ \langle B_t \rangle = (A_t^{-1} + \gamma_t), \]

\[ \langle C_t \rangle = C_t \]

\[ \langle D_t \rangle = D_t \]

\[ \langle A_t \rangle = A_t + \gamma_t \cdot B_t, \]

\[ \langle B_t \rangle = (A_t^{-1} + \gamma_t), \]

\[ \langle C_t \rangle = C_t \]

\[ \langle D_t \rangle = D_t \]

\[ \langle A_t \rangle = A_t + \gamma_t \cdot B_t, \]

\[ \langle B_t \rangle = (A_t^{-1} + \gamma_t), \]

\[ \langle C_t \rangle = C_t \]

\[ \langle D_t \rangle = D_t \]
After performing the VB M-step, the mean vectors and the covariance matrices in Eqs. (D19)–(D25) are calculated and used to compute the expected natural parameters in Eqs. (D2)–(D18). The calculated expected natural parameters are used in the forward and backward recursions of the VEKS algorithm to infer the distribution $q_k(x_{1:t})$ over the hidden state.

D.1. Forward recursion of VEKS

A purpose of the forward recursion in the Kalman filtering is to compute the marginal posterior distribution of the state $x_t$ at the time step $t$ given the history of the measurements up to the time step $t$. For the local linear system in Eq. (34), the marginal posterior distribution is Gaussian distributed.

$$p(x_t | y_{1:t}) = N(x_t | \mu_t, \Sigma_t).$$

The dynamical system in Eq. (34) is a first-order Markov process and we have

$$p(x_t | y_{1:t}) = \frac{p(x_t | y_{1:t-1})p(X_{1:t-1})}{p(y_{1:t})}$$

$$p(y_{1:t}) = p(y_t | x_{1:t})p(x_{1:t}) = \frac{p(y_t | x_{1:t})}{p(y_t | x_{1:t-1})} \exp$$

$$\exp \left\{ -\frac{1}{2} \left[ (x_t - \mu_{t-1})^T \Sigma_{t-1}^{-1} (x_t - \mu_{t-1}) \right. \right.$$}

$$\left. + (x_t - \mu_{t-1})^T \Sigma_{t-1}^{-1} (x_t - \mu_{t-1}) \right\}$$

$$\left[ \frac{1}{2} \right] \frac{1}{\sigma_t(y_t)} \int dx_{t-1} p(x_{t-1} | y_{1:t-1})$$

$$= \frac{1}{\sigma_t(y_t)} \int dx_{t-1} N(x_{t-1} | \mu_{t-1}, \Sigma_{t-1}) p(x_{t-1} | A, x_{t-1}),$$

$$= \frac{1}{\sigma_t(y_t)} \int dx_{t-1} N(x_{t-1} | \mu_{t-1}, \Sigma_{t-1}) p(x_{t-1} | A, x_{t-1}),$$

$$= \frac{1}{\sigma_t(y_t)} \int dx_{t-1} N(x_{t-1} | \mu_{t-1}, \Sigma_{t-1}) N(x_{t-1} | A, x_{t-1}),$$

$$= \frac{1}{\sigma_t(y_t)} \int dx_{t-1} N(x_{t-1} | \mu_{t-1}, \Sigma_{t-1}) N(x_{t-1} | A, x_{t-1}),$$

$$= \frac{1}{\sigma_t(y_t)} \int dx_{t-1}$$

The forward recursion for VEKS is derived based on Eq. (D39) and presented in Algorithm 2. Fig. 1 illustrates a flowchart of the VEKS forward recursion.

D.2. Backward recursion of VEKS

A purpose of an optimal Kalman smoother is to compute the marginal posterior distribution of the state $x_t$ after receiving all measurements from $t = 1$ up to the final time $t = T$, i.e., $p(x_t | y_{1:T})$. Therefore, information about future observations is incorporated in the backward pass of the Kalman filtering to update the posterior distribution of the hidden state at the current time step. For the local linear dynamical system in Eq. (34), as a first-order Markov process, it can be shown that

$$p(x_t | y_{1:T}) = \frac{p(x_t | y_{1:T})}{p(y_{1:T} | x_{1:t})} \propto \frac{p(x_t | y_{1:T})}{p(y_{1:T} | x_{1:t})}$$

$$= \frac{1}{\sigma_t(y_t)} \int dx_{t-1} N(x_{t-1} | \mu_{t-1}, \Sigma_{t-1})$$

$$= \frac{1}{\sigma_t(y_t)} \int dx_{t-1} N(x_{t-1} | \mu_{t-1}, \Sigma_{t-1})$$

$$= \frac{1}{\sigma_t(y_t)} \int dx_{t-1} N(x_{t-1} | \mu_{t-1}, \Sigma_{t-1})$$

Noise in the dynamical system in Eq. (34) is assumed to be Gaussian. Therefore, the conditional distributions in Eq. (D40) are Gaussian distributed and we define the following distributions for them.

$$p(x_t | y_{1:T}) = N(x_t | \mu_t, \Sigma_t),$$

$$p(x_t | y_{1:T}) = N(x_t | \mu_t, \Sigma_t),$$

$$p(x_t | y_{1:T}) = N(x_t | \mu_t, \Sigma_t),$$

Here, distributions $p(x_t | y_{1:T})$ and $p(y_{1:T} | x_{1:t})$ are calculated in the forward and backward recursions of the Kalman smoother, respectively. Based on Eqs. (D40)–(D43), it is straightforward to derive the following distribution for $p(x_t | y_{1:T})$ as the solution of the Kalman smoother.
\[
\Gamma_{t+1} = \left( \Sigma_{t}^{-1} + \Psi_t^{-1} \right)^{-1},
\]
\[
\gamma_t = \Gamma_{t} \left( \Sigma_{t}^{-1} \mu_t + \Psi_t^{-1} \eta_t \right)^{-1}.
\]

We have previously derived the distribution \( p(x_t|y_{1:t}) \) based on the forward recursion of the Kalman filtering. We now turn to the backward recursion in order to calculate distribution \( p(y_{1:t}|x_t) \). For a first-order Markov process, such as the dynamical system in Eq. (34), we have
\[
p(y_{1:t}|x_t) = \int dx_t p(y_{1:t}|x_t, x_{|1:t-1}) \]
\[
= \int dx_t p(x_{|1:t-1}|y_{1:t}) p(x_t|y_{1:t}, x_{|1:t-1}).
\]

\[\text{We define } \Gamma_{t-1,t} \text{ as the cross-covariance matrix between the hidden states at times } t-1 \text{ and } t, \text{ given all the observations } y_{1:t}:\]
\[
\Gamma_{t-1,t} = \langle (x_{t-1} - \langle x_{t-1} \rangle)(x_t - \langle x_t \rangle)^T \rangle.
\]

where \( \langle \cdot \rangle \) denotes expectation w.r.t. the posterior distribution over the hidden state sequence given all the observations. It is not difficult to obtain the variational version of the cross-covariance matrix using the joint distribution in Eq. (D-52):
\[
\Gamma_{t-1,t} = \Sigma_{t-1} \langle A_{t-1}^T A_t \rangle^T \Gamma_{t-1,t}.
\]

### Appendix E. Calculation of lower bound on model marginal likelihood

Before dealing with the calculation of a lower bound on the model marginal likelihood, it is helpful to introduce and use the Kullback–Leibler (KL) divergence. For two distributions \( q(\theta) \) and \( p(\theta) \), the KL divergence of \( p(\theta) \) from \( q(\theta) \) is defined as:
\[
\text{KL}[q(\theta)||p(\theta)] = \int dq(\theta) \ln \frac{q(\theta)}{p(\theta)}.
\]

Here, \( q(\theta) \) and \( p(\theta) \) are the posterior and the prior distributions of the model parameters. A lower bound can be concisely written using the following definition for the expected conditional Kullback–Leibler divergence.
\[
\text{KL}[q_\alpha(x|\beta)||p(x|\beta)]_{q_\alpha} = \int dq_\alpha(x) \int dq_\alpha(x) \ln \frac{q_\alpha(x|\beta)}{p(x|\beta)}.
\]

Here, \( \alpha \) and \( \beta \) are a pair of parameters of the model, \( q_\alpha(\beta) \) is the posterior distribution of \( \beta \), and \( q_\alpha(x|\beta) \) and \( p(x|\beta) \) are the conditional posterior and prior distributions, respectively.

The lower bound on the marginal likelihood of a model is defined in Eq. (10).
\[
F(q_\alpha(x), q_\alpha(\theta)) = - \int dq_\alpha(x) \ln q_\alpha(x) + \int dq_\alpha(x) \ln \frac{p(x, y|\theta, m)}{q_\alpha(x)} + \int dq_\alpha(\theta) \ln \frac{p(\theta|m)}{q_\alpha(\theta)}.
\]

For a CE model based on Eq. (13) of Condition 1, we have
\[
\langle \ln p(x,y|\theta, m) \rangle_{q_\theta(b)} = \ln p(x,y|\overline{\theta}, m) \\
= \ln p(x,y|\overline{\theta}, m) + \ln p(y|\overline{\theta}, m) \\
= \ln p(x,y|\overline{\theta}, m) + \ln Z, \ln Z \\
= \langle \ln p(y|x|\theta, m) \rangle_{q_\theta(b)}. \tag{E-4}
\]

where \( \overline{\theta} = (\Omega(\theta))_{q_\theta(b)} \) is the expected natural parameters and \( Z \) is a normalization constant whose logarithm is calculated in Eq. (50) presented in Algorithm 2. On the other hand, based on Theorem 2 and Eq. (16) for the VB E-step, we have

\[
q_\theta(x) = p(x|y, \overline{\theta}, m). \tag{E-5}
\]

Straight after the VB E-step based on Eqs. (E-3)-(E-5), the lower bound \( F(\cdot) \) is obtained as

\[
F(q_\theta(x), q_\theta(b)) = \ln Z + \int d\theta q_\theta(b) \ln \frac{p(\theta|x, m)}{q_\theta(b)} = \int d\theta q_\theta(b) \ln \frac{p(\theta|x, m)}{q_\theta(b)}. \tag{E-6}
\]

As shown before, the free distribution over the parameters of the dynamical model in Eq. (18) under the constraint of the CM model is

\[
q_\theta(\theta) = q_\theta, \theta, \mu, \rho, q_\theta(\theta), q_\theta(\mu), q_\theta(\rho) \tag{E-7}
\]

Inserting the prior and posterior distributions of the model parameters in Eqs. (A-1) and (E-7) into Eq. (E-10), the lower bound on the log marginal likelihood of the dynamical model in Eq. (18) is obtained.

\[
F(q_\theta(x_0:t), q_\theta(b)) = \int d\theta q_\theta(b) \ln \frac{p(\theta|x_0:t, m)}{q_\theta(b)} = \int d\theta q_\theta(b) \ln \frac{p(\theta|x_0:t, m)}{q_\theta(b)} \tag{E-8}
\]

Using the Kullback-Leibler (KL) divergence, the lower bound in Eq. (E-8) is written as

\[
F(q_\theta(x_0:t), q_\theta(b)) = -KL[q_\theta(b)||p(\theta|x_0:t)] \\
= \langle KL[q_\theta(b)||p(\theta|x_0:t)] \rangle_{q_\theta(b)} \\
= \langle KL[q_\theta(b)||p(\theta|x_0:t)] \rangle_{q_\theta(\theta)} \\
= \langle KL[q_\theta(b)||p(\theta|x_0:t)] \rangle_{q_\theta(\mu)} \\
= \langle KL[q_\theta(b)||p(\theta|x_0:t)] \rangle_{q_\theta(\rho)} + \ln Z. \tag{E-9}
\]

where \( q_\theta(\cdot) \) and \( p(\cdot|x_0:t) \) are the conditional posterior and prior distributions of the model parameters and \( \ln Z \) is calculated using Eq. (50). To calculate the lower bound \( F(\cdot) \) in the above equation, we use the prior distributions in Eqs. (A-1)-(A-6) and the estimated posterior distributions of the parameters, updated in the VB M-step according to Eqs. (B-14)-(B-19).

In the following equations, we derive all required terms to calculate the lower bound \( F(\cdot) \) in Eq. (E-13) (see Appendix C3 in [2] for more details).
Here, $\Gamma(\cdot)$ and $\text{Digamma}(\cdot)$ in Eqs. (E-14) and (F-3) are Gamma and Digamma functions, respectively. Finally, the lower bound on the marginal likelihood of the model in Eq. (E-13) is calculated using the terms calculated in Eqs. (E-14), (F-1)-(F-4) and Eq. (50) in Algorithm 2.

Appendix F. Properties of Kronecker product

Let $U$ be an $N$-by-$M$ matrix, $V$ be an $N$-by-$K$ matrix, $A = \text{diag}(I)$ be an $N$-by-$N$ diagonal matrix, $B$ be an $M$-by-$1$ matrix, and $M$ be a matrix whose dimensions are stated by appropriate superscripts in the following equations. We have the following properties.

1: \[
\text{trace}(\{A \otimes A\} B_{MK-MM}) = \sum_{n=1}^{N} \lambda(n) \cdot S(n,:) \text{ and } S(i,j) = \text{trace}(A B(L,J)) \quad \text{ for } i = 1, \ldots, N.
\]

2: \[
\text{trace}(\{[U^i \cdot A \cdot V] \otimes A\} B_{MK-MM}) = \sum_{n=1}^{N} \lambda(n) \cdot U(n,:) \cdot S(U(n,:)) \text{ and } S(i,j) = \text{trace}(A B(L,J)) \quad \text{ for } i = 1, \ldots, M.
\]

3: \[
\text{trace}(\{[U^i \cdot A \cdot V] \otimes A\} B_{MK-MM}) = \sum_{n=1}^{N} \lambda(n) \cdot V(n,:) \cdot S(U(n,:)) \text{ and } S(i,j) = \text{trace}(A B(L,J)) \quad \text{ for } i = 1, \ldots, M.
\]

Here, $U(n,:)$ and $V(n,:)$ are the $n$th rows of the corresponding matrices and the $L$-by-$1$ vectors $I$ and $J$ contain numbers from $(i-1)L+1$ to $iL$ and from $(j-1)L+1$ to $jL$, respectively.

Proof. We prove the third property and other properties can be straightforwardly proved based on it. Let $e_i = [0 \ldots 0 \ 1 \ 0 \ldots 0]$ and $f_j = [0 \ldots 0 \ 1 \ 0 \ldots 0]$ be 1-by-$M$ and 1-by-$K$ vectors whose entries are zero except the $i$th entry for $e_i$ and the $j$th entry for $f_j$ which are one. The matrix $B$ can be written as

\[
B = \sum_{i=1}^{N} \sum_{j=1}^{M} \{f_j \cdot e_i\} \otimes B(I,J), \quad \text{ for } \{i = (i-1)L+1 : iL \land j = (j-1)L+1 : jL\}.
\]

Using the decomposition in Eq. (44), the left hand side of Eq. (33) is

\[
\text{trace}(\{[U^i \cdot A \cdot V] \otimes A\} B) = \sum_{i=1}^{N} \sum_{j=1}^{M} \text{trace}(\{[U^i \cdot A \cdot V] \otimes A\} \{f_j \cdot e_i\} \otimes B(I,J)).
\]