Journal of Mathematical Psychology 60 (2014) 1–19

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

journal homepage: www.elsevier.com/locate/jmp

Tutorial

# A tutorial on variational Bayes for latent linear stochastic time-series models

Dirk Ostwald [a,b,*], Evgeniya Kirilina [c], Ludger Starke [a], Felix Blankenburg [a,b,c]

[a] Max-Planck-Institute for Human Development, Center for Adaptive Rationality (ARC), Berlin, Germany

[b] Bernstein Center for Computational Neuroscience, Berlin, Germany

[c] Department of Education and Psychology, Neurocomputation and Neuroimaging Unit, Freie Universität, Berlin, Germany

## HIGHLIGHTS

- Stochastic time-series modeling.
- Linear Gaussian state space models.
- Variational Bayes.

## ARTICLE INFO

*Article history:*
Received 21 March 2013
Received in revised form
14 April 2014

## ABSTRACT

Variational Bayesian methods for the identification of latent stochastic time-series models comprising both observed and unobserved random variables have recently gained momentum in machine learning, theoretical neuroscience, and neuroimaging methods development. Despite their established use as a computationally efficient alternative to sampling-based methods, their practical application in mathematical psychology has so far been limited. In this tutorial we attempt to provide an introductory overview of the theoretical underpinnings that the variational Bayesian approach to latent stochastic time-series models rests on by discussing its application in the linear case.



**Contents**



* Correspondence to: Center for Adaptive Rationality, Max-Planck-Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany.
*E-mail address:* dirk.ostwald@mpib-berlin.mpg.de (D. Ostwald).

http://dx.doi.org/10.1016/j.jmp.2014.04.003
0022-2496/

## 1. Introduction

Imagine the following scenario: you have obtained a reaction time data set of a human participant in a perceptual discrimination task over the course of a perceptual training regime (Fig. 1).

The data-points of this curve may, for example, represent median reaction times on a set of test trials, each set collected after completion of a training block. Based on these results, you conclude that the perceptual discrimination process underlying the observed decrease in reaction times has improved with training. In other words, the unobserved, underlying process increased in terms of its quality (which is what you had hoped to achieve by the training procedure), and the observed reaction time decrease can be conceived as an indirect observation of this cognitive process. Somehow, you think, it would be nice to directly access this underlying process and observe its increase.

Inspecting the data further, you realize that obviously, in addition to the decrease in reaction time, there is considerable variation from data-point to data-point. Based on your earlier considerations, this variation may stem from two different sources: either the underlying process is somewhat volatile, or the variation arises at some point of the transformation from the underlying cognitive process to the measured data. Sources for this variability arising in the transformation could for example lie in the participant's motor system or some technical component of your experimental setup. Most likely, you conclude, both "process volatility" and "transformation noise" make some contribution to the observed data variability.

You thus deal with the following data-analytical problem: you would like to infer the state of a latent learning process $x$ from observed data $y$. The underlying process exhibits volatile saturation behavior, that is, it gets better with training, but exhibits random fluctuations. Let us assume that you aim to quantify the increase with a rate constant $\alpha$ and its volatility with a number $\sigma > 0$. For simplicity, let us further assume you would like to quantify the mapping from the underlying process to the reaction time data by a linear function with slope $b$ and quantify the transformation noise by a number $\sigma_y^2 > 0$. In summary, you thus would like to estimate a set of parameters $\alpha, \sigma, b$ and $\sigma_y^2$ in a model that governs the dynamic and stochastic relationships between the underlying process and the observed data.

Thinking further about the data, you start to wonder, what actually happened between data-points. Clearly, the perceptual learning process did not stop when the reaction time measurement was made and stayed constant until the next measurement. In a way, you become convinced that the discrete reaction time data-points $y_1, y_2, \ldots y_{50}$ are more or less an artifact of your measurement procedure, and the actual perceptual learning process evolved continuously during training. You thus also face the challenge of inferring a latent continuous time process from discrete time observation points.

Finally, imagine you have become convinced of the conceptually more intuitive approach to statistical inference offered by the Bayesian paradigm as compared to classical statistics (for example, because you studied (Wagenmakers, 2007)). You might actually have some prior assumptions about the parameter values based on similar studies you ran in the past, which you believe should inform your current inferences. Or you might firmly believe that your training procedure will not decrease the observer's perceptual discrimination ability and thus have a clear prediction about the sign of the rate parameter $\alpha$. Further, you might also be aware that any mathematical formulation you will use to explain the observed data corresponds to only one of a plethora of possible data models (Burnham & Anderson, 2004) and that the Bayesian paradigm offers a principled and straightforward method to compare models based on their relative plausibility under Occam's razor heuristic (Domingos, 1999).

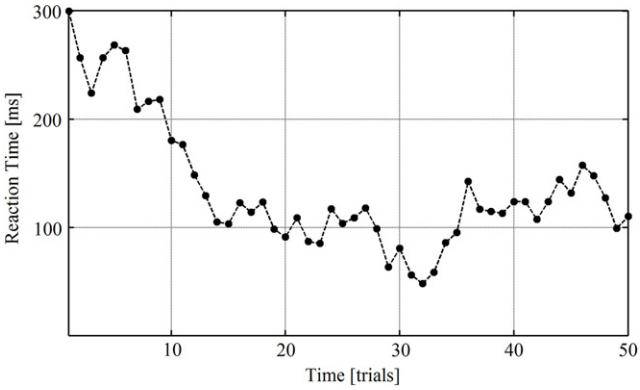


**Fig. 1. A hypothetical reaction time data set.** The figure depicts a simulated observed variable trajectory from a latent linear stochastic time-series model as discussed in Section 2. For demonstrative purposes the trajectory is interpreted as the decrease in reaction times of a single participant in a perceptual discrimination task over the course of a training regime. For example, the data-points may correspond to median reaction times, each collected from a set of perceptual discrimination task trials. Between these test sets, the participant is imagined to have undergone a perceptual training procedure. The data shown correspond to the data that is subject to a variational Bayesian analysis in Section 5 of this tutorial.

### 1.1. Overview

Collectively, the scenario just described is an instance of the following problem, commonly faced in scientific inquiry: one would like to obtain Bayesian estimates of an unobserved, volatile, continuous time process that was observed at discrete time-points under some kind of transformation and under the addition of measurement noise. A mathematical framework that allows one to tackle this kind of problem is variational Bayes (VB) for latent stochastic time-series models. In brief, VB for latent stochastic time-series models is a computationally efficient alternative to numerical sampling approaches (as discussed for example in Lunn (2013)) for Bayesian model identification. By "Bayesian model identification" we understand the combination of Bayesian posterior distribution inference *and* Bayesian model evidence approximation. In contrast to commonly employed model evidence approximations such as the Bayesian information criterion (BIC) (Schwarz, 1978), the model comparison criterion in the VB context – the variational free energy – is not applied post-hoc, but is an integral part of the inference procedure. The VB approach originated from statistical physics (Tuckerman, 2010) and has been introduced into the statistical literature in the context of ensemble learning (MacKay, 1995) and as a generalization of the Expectation–Maximization (EM) algorithm (Neal & Hinton, 1998). It has been popularized in the machine learning, theoretical neuroscience, and neuroimaging communities due to the seminal works of Beal (2003), Dayan and Abbott (2005), Roberts and Penny

(2002), Friston, Mattout, Trujillo-Barreto, Ashburner, and Penny (2007) and Friston et al. (2002). At present, VB underlies a number of computational approaches for Bayesian parameter inference and model comparison commonly used in the neuroimaging community (Friston, 2007). The focus of this tutorial is on the formal development of VB as a Bayesian parameter inference framework, and we only occasionally remark on the variational free energy as a means for model comparison.

VB for latent stochastic time-series models can be viewed as a unifying account of two broad fields of contemporary mathematics, namely stochastic differential equation modeling (Arnold, 1998; Honerkamp, 1993; Kloeden & Platen, 1999) and approximate-deterministic Bayesian inference (Barber, 2012; Bishop, 2007; Murphy, 2012). Its analytical foundations are hence multifaceted. The aim of this tutorial can thus merely be to provide a high-level theoretical introduction to this broad topic. For the mathematically inclined reader this tutorial is accompanied by a set of supplementary materials, available at https://archive.org/details/OstwaldEtAlR2Supplement which we will refer to as "Supplement" in the following. The Supplement covers the mathematical underpinnings of the concepts discussed in some depth and provides a number of examples for the analytical derivation of VB algorithms. Additionally, the Supplement contains a collection of MATLAB functions (The MathWorks, Natick, MA) which provide examples for the implementation of VB algorithms and were used to generate the technical figures of this tutorial.

The outline of the tutorial is as follows: In Section 2, we discuss the mathematical formulation of latent stochastic time-series models. We are here only concerned with a special case, namely the linear one, which upon discretization simplifies to a linear Gaussian state space model (LGSSM). In Section 3, we introduce the general VB framework with a focus on factorized approximations and provide an example by discussing how it can be used to infer the mean and variance of a univariate Gaussian. In Section 4, we then apply VB to the model class introduced in Section 2. As we are dealing with LGSSMs, we will touch upon classical algorithmic approaches to obtain probabilistic statements about their latent states. Finally, in Section 5 we will come back to the reaction time example discussed above, and show how the theoretical developments of Sections 2–4 come to life in this scenario. Of course, this tutorial example is just one very special case of the multitude of possibilities in which the VB approach for latent stochastic time-series models can be effectively applied. In Section 6, we outline some of the generalizations of the approach discussed here and, eventually, review some of the domains of interest to mathematical psychology in which the VB approach has been and may be used in future work.

Before proceeding, a word on notation: In this tutorial we are dealing with mathematical models that comprise random variables and non-random quantities. We also often require statements of conditional probability. To keep track of which quantities are random variables, and thus can be conditioned on, and which quantities are non-random quantities, and thus cannot be conditioned on, we will write $p(x|a)$ for a conditional probability distribution over $x$, conditioned on the random variable $a$, and $p_a(x)$ for an unconditional probability distribution over $x$, parameterized by the non-random quantity $a$. By $N(x;\, \mu, \sigma^2)$, we understand a place-holder for the probability density function of a normal distribution over $x$ and by $G(x;\, a, b)$ a place-holder for the probability density function of a Gamma distribution over $x$. For example, we thus write $p\left(x|\mu, \sigma^2\right) = N(x;\, \mu, \sigma^2)$, if we understand $\mu$ and $\sigma^2$ as random variables, and treat them on the right-hand side of the expression as such. Likewise, we write $p_{\mu,\sigma^2}(x) = N(x;\, \mu, \sigma^2)$, if we understand $\mu$ and $\sigma^2$ as non-random (known or unknown) quantities. The notation "$N(x|\mu, \sigma^2)$" often encountered in machine learning textbooks will not be used. We opted for denoting random variables by lower case letters and denote specific realizations of random variables using an asterisk. For example, if $y$ denotes a random variable, then $y = y^*$ implies that the specific value $y^*$ of the random variable $y$ has been observed. A more comprehensive discussion of the notational conventions used in this tutorial, as well as an overview of a number of important properties of Gaussian and Gamma distributions is provided in Supplement Sections 1 and 2.

## 2. Linear Gaussian state space models

In this section we introduce the central probabilistic model of this tutorial, the linear Gaussian state space model (LGSSM) as an approximation to a latent linear stochastic differential equation (SDE) (Fig. 2). This unifying view of LGSSMs and latent linear SDEs is motivated as follows: the VB approach discussed in Sections 3 and 4 may, more generally, be viewed as a Bayesian scheme for the estimation of continuous time, latent SDE models (Daunizeau, Friston, & Kiebel, 2009; Daunizeau, Stephan, & Friston, 2012; Havlicek, Friston, Jan, Brazdil, & Calhoun, 2011; Li et al., 2011). Specifically, posterior distributions over the discrete time LGSSM parameter variables may be transferred to posterior distributions over the corresponding continuous time latent SDE system using the integral transformation theorem for probability density functions (Casella & Berger, 2002) as demonstrated in Section 5. The unifying view of latent SDEs and LGSSM by means of a discretization approach may thus be regarded as an attempt to bridge the gap between the theoretical focus of stochastic analysis and the applied approach of statistics and machine learning. In addition, other VB estimation schemes for latent random systems (for example Friston (2008a,b), Friston, Stephan, Li, and Daunizeau (2010); Friston, Trujillo-Barreto, and Daunizeau (2008)) are derived directly in the context of hierarchical continuous time random processes, as discussed in Section 6. In summary, establishing the continuous time origin of the LGSSM enables us to view the VB approach to LGSSMs in the more general context of Bayesian random dynamical system identification (Arnold, 1998).

### 2.1. LGSSMs as discretized latent linear SDEs

The model considered in this tutorial is based on two equations: a "dynamics equation" governing the stochastic temporal evolution of a latent (also referred to as "unobserved", "unobservable", "hidden" or "state") random variable and an "observation equation" describing the mapping from the latent to the observed (also referred to as "visible" or "observable") random variable. Empirical measurements of observed variables are usually discrete. This discrete time nature of observed time-series data may be envisaged as observing the continuous time latent variable at discrete time-points under the transformation of the observation process and under the addition of measurement noise. In its most general form, the model may thus be formulated as the following pair of equations, describing an augmented diffusion process:

$$\begin{aligned} dX(t) &= \mu\left(t, X\left(t\right)\right) dt + \sigma\left(t, X\left(t\right)\right) dW(t), \quad t \in [0, s] \\ y_{t_i} &= f\left(t_i, X\left(t_i\right)\right) + \eta_{t_i}, \quad t_i \in [0, s], i = 0, 1, \ldots, n. \end{aligned} \tag{1}$$

In (1), $X$ denotes the latent variable as a function of continuous time $t$, $\mu$ denotes a drift function which describes the structural aspects of the evolution of $X$, $\sigma$ denotes a volatility function which describes the stochastic aspects of the evolution of $X$, $W(t)$ denotes a Wiener process, $y_{t_i}$ denotes the observable variable at time-point $t_i$, $f$ denotes a function that maps the state of the latent variable $X$ at time $t_i$ onto the deterministic component of $y_{t_i}$, and $\eta_{t_i}$ denotes an additional error term distributed according to a Gaussian distribution with expectation 0 and variance $\sigma_y^2$. In the context of the introductory example, $X$ would refer to the unobserved perceptual

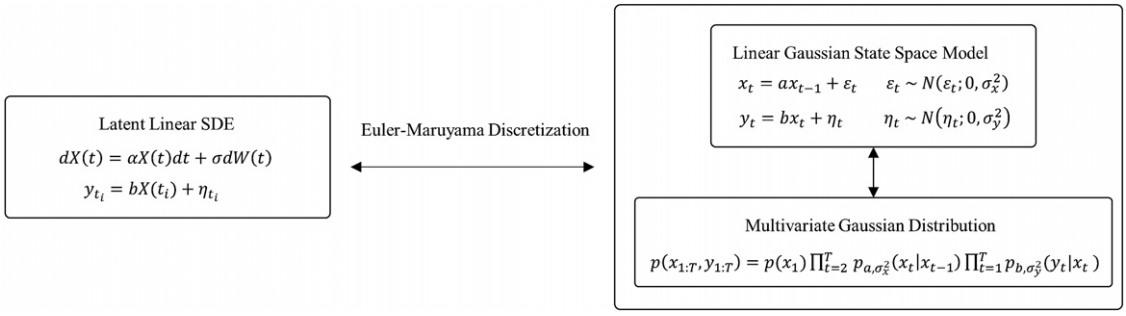


**Fig. 2. Latent linear stochastic time-series models.** The figure sketches the model formulation problem discussed in the tutorial. From a stochastic analysis viewpoint, linear Gaussian state space models (right) may be conceived as discretized, latent, autonomous, linear SDEs in the narrow sense (left). The Ito-calculus based Euler–Maruyama discretization scheme linking both viewpoints provides a bridge between the "theoretical" approach of stochastic analysis and the "applied" approach of machine learning. Additionally, an algorithmic treatment of LGSSMs is best achieved when viewed as multivariate Gaussian distributions (lower right).

learning process, while $y_{t_i}$ would refer to the observed median reaction time data-points. More common exemplary models that are formulated as SDEs in the mathematical psychology literature are models of reaction times (e.g. Ratcliff, 1978; Ratcliff & Van Dongen, 2011; Smith & Ratcliff, 2004) and value-based decision making (e.g. Busemeyer, Jessup, Johnson, & Townsend, 2006; Busemeyer & Townsend, 1993; Huang, Sen, & Szidarovszky, 2012). There exist a number of excellent introductory articles to stochastic differential equations (e.g. Brown, Ratcliff, & Smith, 2006; Huang et al., 2012; Smith, 2000; Zhang, Bogacz, & Holmes, 2009). For readers interested in the mathematical details of the formulation used here, we have included an introductory review of the most important concepts from stochastic analysis in Supplement Section 3. Note that we use a capital letter for the unobserved variable $X$ in its continuous formulation, and a lower case letter for $x$ in its discrete time formulation below.

To simplify proceedings, from now on, we only consider those augmented diffusions for which the dynamics equation may be expressed as a linear SDE under pure additive noise with constant, i.e., time-invariant, coefficients. In this case, the functions $\mu$ and $\sigma$ above take the general forms

$$\mu\left(t, X\left(t\right)\right) := \alpha X(t) \quad \text{and} \quad \sigma\left(t, X\left(t\right)\right) := \sigma \tag{2}$$

for fixed constants $\alpha$ and $\sigma$. Further, only linear and time-invariant observation equations are considered, that is, the function $f$ takes the form

$$f\left(t_i, X\left(t_i\right)\right) := bX(t_i) \tag{3}$$

for a fixed constant $b$. We thus arrive at the "augmented linear diffusion process"

$$\begin{aligned} dX(t) &= \alpha X\left(t\right) dt + \sigma dW\left(t\right), \quad t \in [0, s] \\ y_{t_i} &= bX(t_i) + \eta_{t_i}, \quad t_i \in [0, s], i = 0, 1, \ldots, n. \end{aligned} \tag{4}$$

The latent stochastic differential equation in (4) is generally known as the Langevin-Equation and represents an autonomous linear SDE in the narrow sense (Kloeden & Platen, 1999).

A discrete time version of the system described in Eq. (4) can be obtained by Euler–Maruyama discretization (Mil'shtein, 2011). Informally, Euler–Maruyama discretization corresponds to the combination of Euler's method for the numerical evaluation of ordinary differential equations (Press, 2007) and Ito's stochastic integration (Øksendal, 2003). Specifically, because the volatility function $\sigma$ is constant in the current scenario, Ito integration simplifies to Riemann–Stieltjes integration and the volatility parameter $\sigma$ becomes a "scaling" parameter for the additive noise in the familiar LGSSM form of (4)

$$\begin{aligned} x_t &= ax_{t-1} + \varepsilon_t, \quad t = 2, \ldots, T \\ y_t &= bx_t + \eta_t, \quad t = 1, \ldots, T. \end{aligned} \tag{5}$$

In (5), $a := (1 + \alpha \Delta t)$ and $\varepsilon_t$ is distributed according to a univariate Gaussian with expectation 0 and variance $\sigma_x^2 = \sigma^2 \Delta t$, where $\Delta t$ denotes the discretization time interval of the Euler–Maruyama approach. In other words, the Euler–Maruyama discretization scheme allows for deriving closed form solutions for the parameters of a latent autoregressive process of order 1 (AR(1)-process (Shumway & Stoffer, 2011)) in terms of the latent stochastic differential equation parameters $\alpha$ and $\sigma$ and the distance between the discrete time-points. In addition, in (5) we have set $y_t := y_{t_i}$ and $\eta_t := \eta_{t_i}$, rendering $t$ the discrete time counting index, which runs from 1 to $T := n+1$, as opposed to $i$ in (1), which ran from $t_0 := 0$ to $t_n := s$. It should be noted that the approach of deriving an AR(1)-LGSSM representation based on a linear latent SDE is straightforward only in the univariate case. In this case, it provides an intuitive way to obtain Bayesian posterior estimates of the parameters of the latent stochastic dynamical system (1): If posterior distributions for $a$ and $\sigma_x^2$ can be derived at the LGSSM level, these may be transferred to the latent SDE level by means of the integral transform for probability density functions. However, it should be noted that this approach is formally correct only in the limiting case $n \to \infty$ (see e.g. Ozaki (1992) for an in-depth discussion of this problem). A more comprehensive treatment of the discretization approach introduced above and the integral transform for probability density functions is provided in Supplement Section 3.4.

### 2.2. LGSSMs as Gaussian distributions

In the previous section, we discussed how the LGSSM formulated in Eq. (5) may be viewed as a discrete time approximation of a continuous time augmented stochastic dynamical system. The aim of the current section is to study the standard form of a univariate LGSSM in further detail. By "univariate LGSSM" we understand that both the latent state variable $x$ and the observed variable $y$ are one-dimensional. For general, multivariate LGSSMs $x$ and $y$ correspond to random vectors (see Supplement Section 6).

It is central for the understanding of LGSSMs that Eq. (5) specifies a multivariate joint Gaussian distribution over the latent state variables $x_1, \ldots, x_T$ and the observed variables $y_1, \ldots, y_T$. Specifically, due to the Gaussian properties of $\varepsilon_t$ and $\eta_t$, expression (5) specifies the following conditional probability distributions over state and observed variables $x_t$ and $y_t$, respectively,

$$p_{a,\sigma_x^2}\left(x_t|x_{t-1}\right) = N\left(x_t; ax_{t-1}, \sigma_x^2\right), \quad t = 2, \ldots, T \tag{6}$$

and

$$p_{b,\sigma_y^2}\left(y_t|x_t\right) = N\left(y_t; bx_t, \sigma_y^2\right), \quad t = 1, \ldots, T. \tag{7}$$

The specification of the joint distribution for the LGSSM is completed by a further Gaussian distribution over $x_1$ with expectation $\mu_1$ and variance $\sigma_1^2$

$$p_{\mu_1,\sigma_1^2}\left(x_1\right) = N(x_1; \mu_1, \sigma_1^2) \tag{8}$$

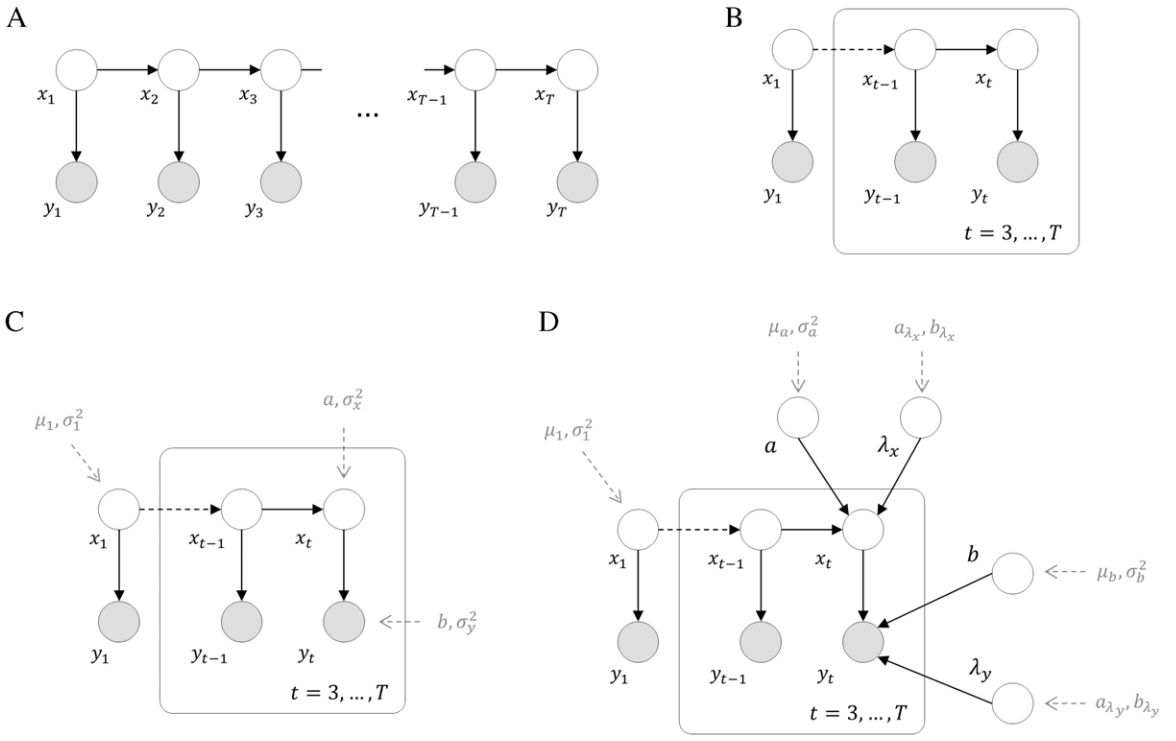


**Fig. 3. Graphical depiction of the LGSSM** (A) The LGSSM comprises a set of unobserved variables, here denoted by white circles $x_{1:T}$, which exhibit Markovian conditional dependencies, denoted by the arrows between $x_{t-1}$ and $x_t$, $t = 2, \ldots, T$. The LGSSM also comprises a set of observed variables, here denoted by shaded circles, which are mutually independent, but each conditionally dependent on a specific unobserved variable, denoted by the arrows from $x_t$ to $y_t$. (B) The repetitive structure of (A) is more conveniently condensed using the plate notation, i.e. the rectangle over the recurring motif in (A). (C) The figure depicts the LGSSM in a classical, non-Bayesian scenario using the plate notation of Fig. 3(B). Here, the quantities comprising $\theta$ do not correspond to random variables and are thus not shown as circles. $a$ and $\sigma_x^2$ govern the evolution of the unobserved random variables $x_{1:T}$ and $b$ and $\sigma_y^2$ govern the emission of $y_{1:T}$ based on the states of $x_{1:T}$. (D) The figure depicts the LGSSM in the Bayesian scenario of interest in the tutorial. Because the quantities comprising $\theta$ are now treated as random variables, they are grouped with the unobserved variables $x_{1:T}$ into the set of unobserved variables $\vartheta := \{x_{1:T}, \theta\}$ in Section 3. Note that in the Bayesian scenario, the variance parameters $\sigma_x^2$ and $\sigma_y^2$ are replaced by their reciprocals, $\lambda_x$ and $\lambda_y$, respectively. The parameters parameterizing the prior distributions over the quantities in $\theta$ are included for completeness.

(note that from the viewpoint of the augmented linear diffusion process, the distribution over $x_1$ corresponds to the specification of a (stochastic) initial condition $X(t_0) = X(0)$). To write the above more succinctly, we now introduce the "colon notation" abbreviations $x_{1:T} := (x_1, \ldots, x_T)$ and $y_{1:T} := (y_1, \ldots, y_T)$. Using this convention and summarizing the parameters in the vector $\theta := \left(\mu_1, \sigma_1^2, a, \sigma_x^2, b, \sigma_y^2\right)$ then allows for writing the joint distribution over observed and latent variables as

$$p_\theta\left(y_{1:T}, x_{1:T}\right) = p_{\mu_1,\sigma_1^2}\left(x_1\right) \prod_{t=2}^{T} p_{a,\sigma_x^2}\left(x_t|x_{t-1}\right) \prod_{t=1}^{T} p_{b,\sigma_y^2}(y_t|x_t). \quad (9)$$

Eq. (9) states that the joint distribution over all variables $x_{1:T}$ and $y_{1:T}$ of the LGSSM is given by the product of Gaussian marginal and conditional distributions over $x_{1:T}$ and $y_{1:T}$, respectively. Because the product of Gaussian probability density functions is again a Gaussian probability density function, (9) specifies a Gaussian joint distribution by means of its factorization properties. From a data analytical viewpoint, an observed time-series may be viewed as a partially observed realization $\left(x_{1:T}, y_{1:T} = y_{1:T}^*\right)$ sampled from the probability distribution on the left-hand side of Eq. (9). In other words, although the latent variables $x_{1:T}$ have not been observed, they assume a specific, but unknown, state for each realization sampled from (9). Time-series modeling by means of the LGSSM then faces two questions: (1) What is the most likely unobserved sequence of unobservable states $x_{1:T}$ that has given rise to the observed series of values for $y_{1:T}$, and (2) which values did the quantity $\theta$ most likely assume?

At this point it is helpful to establish some linguistic conventions with respect to variables involved in a Bayesian treatment of (9). In a non-Bayesian setting, i.e. the case where $\theta$ is not governed by probability distributions, $y_{1:T}$ are classically referred to as "observed variables", and $x_{1:T}$ are referred to as "latent (or "unobserved" or "hidden") variables". In this context, $\theta$ is referred to as "parameter" (Shumway & Stoffer, 2011). Deriving probabilistic statements about the state of the latent variables based on the observed variables and the (known) parameters is often referred to as "inference" and estimating the parameter value based on the observed data and inferred or augmented states of the unobserved variables is referred to as "(parameter) estimation" or "learning" (e.g. Barber, 2012; Barber & Chiappa, 2007; Bishop, 2007). If, in a Bayesian treatment, the quantity $\theta$ becomes itself governed by probability distributions, the categorical distinction between "latent variables" $x_{1:T}$ and "parameter" $\theta$ becomes somewhat arbitrary. Often, the random variable $\theta$ is nevertheless referred to as parameter, while the parameters (sic!) governing the distribution of $\theta$ are referred to as "hyperparameters". To diminish the risk of confusion, we opted for the language of probabilistic graphical models (see e.g. Jordan (1999) and below). In our Bayesian treatment of (9) we will thus refer to $y_{1:T}$ as "observed variables" and to $x_{1:T}$ and $\theta$ as "unobserved variables". The parameters governing the distribution of $\theta$, which are in this tutorial not treated as random variables, are from now on referred to as "parameters". While this linguistic treatment might appear pedantic at this point, it becomes crucial in the introduction of VB in the next section. In its general version, VB only distinguishes between observed random variables $y$ (i.e., $y_{1:T}$ with respect to the LGSSM) and unobserved random variables $\vartheta$ (i.e., $x_{1:T}$ and $\theta$ with respect to the LGSSM).

The LGSSM of Eq. (5) (or, equivalently, Eq. (9)) is visualized in Fig. 3, using the intuitive notation afforded by probabilistic graphical models. As noted in Lodewyckx et al. (2011), graphical

models are a standard high-level language for representing probabilistic models, widely used in statistics, machine learning, and, increasingly, psychological and neuroimaging modeling (e.g. Daunizeau, David, & Stephan, 2011; David et al., 2006; Friston, Harrison, & Penny, 2003; Lee, 2008; Shiffrin, Lee, Kim, & Wagenmakers, 2008). In Fig. 3(A) nodes (circles) correspond to random variables and edges (arrows) are used to indicate conditional dependencies between variables. Observed variables are shaded and unobserved variables are not shaded. In Fig. 3(B) additional use is made of the "plate" notation which allows for the simplified graphical depiction of recurring conditional dependency motifs. Fig. 3(C) visualizes the LGSSM in a non-Bayesian scenario, while Fig. 3(D) visualizes the LGSSM in the scenario of the tutorial. Here, the quantities comprising $\theta$ correspond to unobserved variables, which are themselves governed by parameterized probability distributions. It is the parameters of these probability distributions that the VB approach allows to derive numerical update equations for, as will be discussed in subsequent sections.

## 3. Variational Bayes

Variational Bayes (VB) is a statistical framework for probabilistic models comprising unobserved variables and has received increasing attention in the machine learning, theoretical neuroscience, and neuroimaging literature since the late 1990s (e.g. Barber, 2012; Bishop, 2007; Friston, 2010; Murphy, 2012; Penny, Kiebel, & Friston, 2003). In this tutorial, we are concerned with the application of VB to LGSSMs as formulated in Section 2, Eq. (9). In Section 4, we discuss how VB for LGSSMs reduces to the combination of parameter update equations with an augmented Kalman–Rauch–Tung–Striebel (KRTS) smoothing algorithm (Briers, Doucet, & Maskell, 2004; Rauch, Striebel, & Tung, 1965). However, to help the understanding of this derivation, it is necessary to consider a more general viewpoint first. Below we outline the VB approach for factorized approximations and discuss its application to a simple example. Based on these intuitions we are then in the position to study the implications of VB for LGSSM in Section 4.

### 3.1. VB using factorized approximations

The general starting point of a Bayesian approach is a joint distribution over observed variables $y$ and unobserved variables $\vartheta$

$$p\left(y, \vartheta\right) = p(\vartheta)p(y|\vartheta) \tag{10}$$

where $p(\vartheta)$ is usually referred to as the prior distribution and $p(y|\vartheta)$ as the likelihood. In the context of Bayesian treatments of LGSSMs (cf. Eq. (9)), $y$ comprises the set of observed variables $y_{1:T}$ and $\vartheta$ comprises the set of unobserved variables $x_{1:T}$, $\mu_1$, $\sigma_1^2$, $a$, $\sigma_x^2$, $b$ and $\sigma_y^2$. Joint distributions over observed and unobserved variables are sometimes referred to as "generative models" (Friston, 2008a,b), a convention we will follow here. Given an observed value $y^*$ of $y$, the first aim of a Bayesian approach is to determine the conditional distribution of $\vartheta$ given $y^*$, referred to as the posterior distribution. The second aim of a Bayesian approach is to evaluate the logarithm of the marginal probability of the observed data $y$ denoted by

$$\ln p\left(y\right) = \ln \int p(y, \vartheta) d\vartheta. \tag{11}$$

If a model only comprises non-random quantities classically referred to as "parameters", the left hand side of (11) is referred to as "log likelihood" (Lehmann & Casella, 1998; Myung, 2003) and no integration as on the right-hand side of (11) is required. However, if a model comprises unobserved random variables, which are integrated out as on the right hand side of (11), (11) is referred to as the "log marginal likelihood" or "log model evidence" (Beal, 2003; MacKay, 2003). The log model evidence allows for comparing different models in their plausibility to explain observed data. It thus forms the necessary prerequisite for Bayesian model comparison (Kass & Raftery, 1995; Penny, 2012; Penny, Stephan, Mechelli, & Friston, 2004; Wasserman, 2000). In the VB framework it is not the log model evidence itself which is evaluated, but rather a lower bound approximation to it. This is due to the fact that if a model comprises many unobserved variables $\vartheta$, the integration of the right-hand side of Eq. (11) can become analytically burdensome or even intractable. To nevertheless achieve the two aims of a Bayesian approach, VB in effect replaces an integration problem with an optimization problem. To this end, VB exploits a set of information theoretic quantities, which we will introduce below. We start with a brief note on information theory.

Information theory can be viewed as a collection of information theoretic quantities (Cover & Thomas, 1991; Shannon, 1948). Well known examples of information theoretic quantities are information entropy and mutual information. The common and defining feature of information theoretic quantities is that they are functions that map probability distributions onto scalar numbers. Because probability distributions (or, more precisely, probability density functions and probability mass functions) are themselves functions, information theoretic quantities are also referred to as functionals (functions of functions). We will highlight this special mathematical nature of information theoretic quantities by using non-serif symbols for them, for example, $\mathsf{F}$ instead of $F$. In the following, we first introduce a decomposition of the log model evidence into the sum of two information theoretic quantities and then explain how these quantities are used to achieve posterior distribution inference and log model evidence approximation.

The following log model evidence composition forms the core of the VB approach (Fig. 4(A)):

$$\ln p\left(y\right) = \mathsf{F}\left(q(\vartheta)\right) + \mathsf{KL}\left(q(\vartheta)||p\left(\vartheta|y\right)\right) \tag{12}$$

where $q(\vartheta)$ denotes an arbitrary probability distribution over the unobserved variables, which is used as an approximation of the posterior distribution $p\left(\vartheta|y\right)$. In the following, $q(\vartheta)$ is referred to as "variational distribution". In words, Eq. (12) states that for an arbitrary variational distribution $q(\vartheta)$ over the unobserved variables, the log model evidence comprises the sum of two information theoretic quantities: $\mathsf{F}(q(\vartheta))$, referred to as the "variational free energy", and $\mathsf{KL}(q(\vartheta)||p(\vartheta|y))$, referred to as the Kullback–Leibler (KL) divergence between the true posterior distribution $p\left(\vartheta|y\right)$ and the variational distribution $q(\vartheta)$ (Kullback & Leibler, 1951). In the VB framework, the variational free energy $\mathsf{F}(q(\vartheta))$ is defined as

$$\mathsf{F}(q(\vartheta)) \coloneqq \int q\left(\vartheta\right) \ln\left(\frac{p\left(y, \vartheta\right)}{q\left(\vartheta\right)}\right) d\vartheta. \tag{13}$$

For the LGSSM, the functional (13) corresponds to a high dimensional integral over $x_{1:T}$ and $\theta$. The KL-divergence term is defined as

$$\mathsf{KL}(q(\vartheta)||p\left(\vartheta|y\right)) \coloneqq \int q\left(\vartheta\right) \ln\left(\frac{q\left(\vartheta\right)}{p\left(\vartheta|y\right)}\right) d\vartheta. \tag{14}$$

Again, for the LGSSM, the KL-divergence functional corresponds to a high dimensional integral over $x_{1:T}$ and $\theta$. The KL-divergence can intuitively be understood as a distance function between two probability distributions: if the two probability distributions $q\left(\vartheta\right)$ and $p\left(\vartheta|y\right)$ are very dissimilar, $\mathsf{KL}\left(q\left(\vartheta\right)||p\left(\vartheta|y\right)\right)$ returns a high positive value. If, and only if, the two probability distributions $q\left(\vartheta\right)$ and $p\left(\vartheta|y\right)$ are identical, it returns 0. In other words, the KL-divergence is always larger or equal to 0, but never negative. This non-negativity property is a consequence of the definition of the KL-divergence and Jensen's inequality for convex functions, as shown in Appendix A. The non-negativity of the KL-divergence

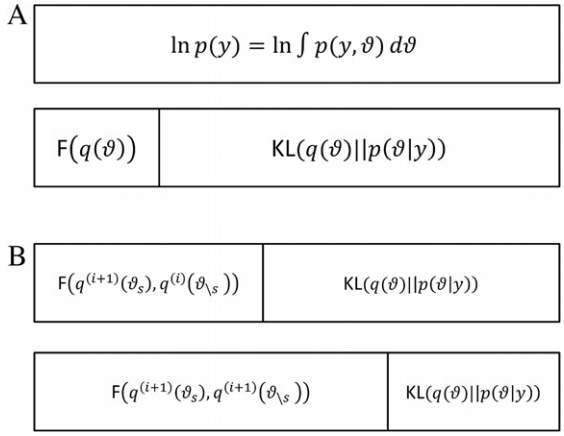


**Fig. 4. Principles of VB for mean-field approximations.** (A) Fig. 4(A) visualizes the log model evidence decomposition that lies at the heart of the VB approach. The upper vertical bar is meant to represent the log model evidence, which is a function of the generative model $p(y, \vartheta)$ and is constant for any observation $y^*$ of $y$. As shown in Appendix B, the log model evidence can be rewritten into the sum of the variational free energy term $\mathsf{F}(q(\vartheta))$ and a KL-divergence term $\mathsf{KL}(q(\vartheta)||p(\vartheta|y))$, if one introduces an arbitrary variational distribution over the unobserved variables $\vartheta$. (B) The log model evidence decomposition of Fig. 4(A) is exploited in numerical algorithms for VB inference as depicted in Fig. 4(B): based on a mean-field approximation $q(\vartheta) = q(\vartheta_s)q(\vartheta_{\setminus s})$, the variational free energy can be maximized in a coordinate-wise fashion. Maximizing the variational free energy in turn has two implications: it decreases the KL-divergence between $q(\vartheta)$ and the true posterior $p(\vartheta|y)$ and renders the variational free energy a closer approximation to the log model evidence. This holds true, because the log model evidence for a given observation $y^*$ is constant (represented by the constant length of the vertical bar) and the KL-divergence is non-negative.

has the immediate consequence, that the variational free energy $\mathsf{F}(q(\vartheta))$ is always smaller or equal to the log model evidence

$$\mathsf{F}(q(\vartheta)) \leqslant \ln p(y). \tag{15}$$

This fact is exploited in the numerical application of the VB approach to probabilistic models: Because the log model evidence is a fixed quantity, which only depends on the choice of $p(y, \vartheta)$, manipulating the variational distribution $q(\vartheta)$ in such a manner that the variational free energy increases has two consequences: first, the lower bound to the log model evidence becomes tighter, and the variational free energy a better approximation to the log model evidence. Second, because the left hand side of (12) remains constant, the KL-divergence between the true posterior and its variational approximation decreases, which renders the variational distribution $q(\vartheta)$ a better approximation to the true posterior distribution $p(\vartheta|y)$ (Fig. 4(B)). Some readers may wonder why the decomposition of the log model evidence stated in (12) holds. In fact, it is a mere consequence of the definitions of the variational free energy (13) and the KL-divergence (14) and can be readily verified as shown in Appendix B.

A common choice for the variational distribution $q(\vartheta)$ over the unobserved variables is a factorization over sets of variables $s$, often referred to as "mean-field approximation"

$$q(\vartheta) = q(\vartheta_s)q(\vartheta_{\setminus s}). \tag{16}$$

In (16) $\vartheta_{\setminus s}$ denotes all unobserved variables not in the $s$th group. For the LGSSM such a factorization is for example $q(x_{1:T})q(\theta)$. In this case, the variational free energy becomes a function of two arguments, namely $q(\vartheta_s)$ and $q\left(\vartheta_{\setminus s}\right)$. Due to the complexity of the integrals involved, a simultaneous analytical maximization of the variational free energy with respect to both its arguments is often difficult to achieve, and a "coordinate-wise" approach, i.e. maximizing first with respect to $q(\vartheta_s)$ and second with respect to $q(\vartheta_{\setminus s})$, is preferred (Beal, 2003). Notably, the assumed factorization over sets of variables corresponds to the assumption, that the respective variables form stochastically independent contributions to the multivariate posterior, which, depending on the true form of the generative model, may have weak or strong implications for the validity of the ensuing posterior inference.

The question is thus how to obtain the arguments $q(\vartheta_s)$ and $q\left(\vartheta_{\setminus s}\right)$ that maximize the variational free energy. It turns out that this challenge corresponds to a well-known problem in statistical physics, which has long been solved in a general fashion using variational calculus (Hinton & van Camp, 1993; Tuckerman, 2010; Van Brunt, 2004). In contrast to ordinary calculus, which deals with the optimization of functions with respect to real numbers, variational calculus deals with the optimization of functions (in this context also referred to as "functionals") with respect to functions. Using variational calculus, it can be shown that the variational free energy is maximized with respect to the unobserved variable partition $\vartheta_s$, if $q(\vartheta_s)$ is set proportional (i.e. equal up to a scaling factor) to the exponential of the expected log joint probability of $y$ and $\vartheta$ under the variational distribution over $\vartheta_{\setminus s}$. Formally, this can be written as

$$q(\vartheta_s) \propto \exp\left(\int q\left(\vartheta_{\setminus s}\right) \ln p(y, \vartheta)\, d\vartheta_{\setminus s}\right). \tag{17}$$

The result stated in (17) is fundamental. It represents the general VB strategy to obtain variational distributions over unobserved variables in light of data and maximizing the lower bound to the log model evidence. We thus refer to (17) as the "VB inference theorem for mean-field approximations". For some foundations of variational calculus and how these enable the derivation of the general solution (17), we refer the interested reader to Supplement Section 4.

Based on the VB inference theorem for mean-field approximations, algorithmic implementations of the VB approach can use an iterative coordinate-wise variational free energy ascent. For iterations $i = 0, 1, 2, \ldots$, this strategy proceeds as follows. The ascent starts by initializing $q^{(0)}(\vartheta_s)$ and $q^{(0)}\left(\vartheta_{\setminus s}\right)$, commonly by equating them to the prior distributions over $\vartheta_s$ and $\vartheta_{\setminus s}$, respectively. Based on (17) it then continues by maximizing the variational free energy $\mathsf{F}(q^{(i)}(\vartheta_s), q^{(i)}(\vartheta_{\setminus s}))$, first with respect to the distribution $q^{(i)}(\vartheta_s)$ given $q^{(i)}\left(\vartheta_{\setminus s}\right)$, yielding the updated distribution $q^{(i+1)}(\vartheta_s)$. Then, by exchanging the labeling of $\vartheta_s$ and $\vartheta_{\setminus s}$ in (17), the ascent continues by maximizing the variational free energy with respect to the distribution $q^{(i)}\left(\vartheta_{\setminus s}\right)$ given $q^{(i+1)}(\vartheta_s)$, yielding $q^{(i+1)}\left(\vartheta_{\setminus s}\right)$. This procedure is then iterated until convergence.

Commonly, the initialization step sets the variational distribution $q^{(0)}(\vartheta)$ to the prior distribution $p(\vartheta)$. This defines the starting point of the iterative procedure as representative of the knowledge about the unknown variables before observed data is taken into account. Further, this choice often enables the use of the well-known benefits of parameterized conjugate prior distributions in the context of VB. The initialization of the variational distribution in terms of the prior distribution, and the subsequent optimization of the variational distributions should not be confused with an empirical Bayesian approach, in which the priors themselves are learned from the data (Efron, 2013): On each iteration of the VB algorithm, the variational distribution corresponds to the approximate posterior distribution, not an updated prior distribution. An empirical Bayesian extension of the VB algorithm on the other hand would correspond to a variation of the prior distribution (specifying the VB algorithm starting conditions) after convergence with the aim of increasing the log model evidence *per se*. A VB algorithm as described here merely increases the lower bound to the fixed log model evidence, which is determined by the choice of the prior $p(\vartheta)$ and likelihood $p(y|\vartheta)$, i.e, the generative model $p(y, \vartheta)$. To summarize the above, a general iterative algorithm for VB inference is outlined in Table 1.

The iterative scheme discussed above shares many similarities with the well-known Expectation–Maximization (EM) algorithm for models comprising unobserved variables (Dempster, Laird, &

**Table 1**
General VB algorithm for inferring a variational mean-field approximation to a posterior distribution over unobserved variables in a generative model.

| | |
|---|---|
| Step (0) | Initialize $q^{(0)}(\vartheta_s)$ and $q^{(0)}(\vartheta_{\setminus s})$ appropriately |
| for $i = 0, 1, 2, \ldots$ until convergence | |
| Step (1) | Set $q^{(i+1)}(\vartheta_s)$ proportional to $\exp\left(\int q^{(i)}(\vartheta_{\setminus s}) \ln p(y, \vartheta)\, d\vartheta_{\setminus s}\right)$ |
| Step (2) | Set $q^{(i+1)}(\vartheta_{\setminus s})$ proportional to $\exp\left(\int q^{(i+1)}(\vartheta_s) \ln p(y, \vartheta)\, d\vartheta_s\right)$ |

Rubin, 1977; McLachlan & Krishnan, 2008; Wu, 1983). In fact, the VB approach discussed here can be viewed as a generalization of the EM-algorithm for maximum likelihood estimation to Bayesian inference (Beal, 2003). In the classical approach to maximum likelihood estimation of LGSSMs, step (1) in Table 1 corresponds to the inference or E-step, while step (2) corresponds to the estimation or M-Step (under the additional assumption of a Dirac point measure over $\theta$).

### 3.2. Example: inferring the mean and variance of a univariate Gaussian

To get an intuition of the VB approach, we next demonstrate how it is made concrete in the context of a well-known example: inferring the mean $\mu$ and precision $\lambda$ (inverse variance) of a univariate Gaussian distribution. For introductory VB applications to linear regression and Gaussian mixture models, see for example Tzikas, Likas, and Galatsanos (2008). For a more comprehensive discussion of the example, we refer the reader to Supplement Section 5. For the example, we assume that a set of $N$ i.i.d. realizations $y^* = \left(y_1^*, \ldots, y_N^*\right)$ of variables $y = \{y_n\}_{n=1,\ldots,N}$ has been obtained. In this scenario, a classical ML approach would result in point estimates for $\mu$ and $\lambda^{-1}$ based on the sample mean and the (biased) sample variance, respectively. As outlined above, starting from appropriately chosen prior distributions, the aim of the Bayesian paradigm is to obtain posterior distributions over unobserved variables. The generative model for the current example constitutes a joint probability distribution over the observed variables $y$ and both unobserved variables $\mu$ and $\lambda^{-1}$, $\lambda > 0$. Specifically, the generative model is given by

$$p(y, \mu, \lambda) = p(\mu, \lambda)\, p(y|\mu, \lambda) = p(\mu, \lambda) \prod_{n=1}^{N} p(y_n|\mu, \lambda)\,. \quad (18)$$

A possible choice for the prior distribution $p(\mu, \lambda)$ over the unobserved random variables is given by the product of a univariate Gaussian distribution over $\mu$ and a Gamma distribution over $\lambda$ (Chappell, Groves, & Woolrich, 2008; Penny, 2000) according to

$$\begin{aligned} p(y, \mu, \lambda) &= p(\mu)\, p(\lambda) p(y|\mu, \lambda) \\ &= N(\mu; m_\mu, s_\mu^2) G(\lambda; a_\lambda, b_\lambda) \prod_{n=1}^{N} N\left(y_n; \mu, \lambda^{-1}\right). \end{aligned} \quad (19)$$

Note that the choice of an independent prior distribution of the form $p(\mu)\, p(\lambda)$ is not a necessary prerequisite for the application of the VB approach. For the use of a dependent prior of the form $p(\mu|\lambda)\, p(\lambda)$ in the context of VB, see for example (Murphy, 2012, p. 742). We next consider a mean-field approximation to the posterior distribution over the unobserved variables, i.e., we set

$$p(\mu, \lambda|y) \approx q(\mu)\, q(\lambda)\,. \quad (20)$$

Thus, in this case, the VB mean-field approximation approximates the true posterior distribution over $\mu$ and $\lambda$ with a distribution over $\mu$ that is independent of the distribution over $\lambda$. We set $q(\mu)$ to a Gaussian distribution with variational parameters $m_\mu^q$ and $s_\mu^{2q}$ and we set $q(\lambda)$ to a Gamma distribution with variational parameters $a_\lambda^q$ and $b_\lambda^q$, resulting in

$$q(\mu)\, q(\lambda) = N\left(\mu; m_\mu^q, s_\mu^{2^q}\right) G\left(\lambda; a_\lambda^q, b_\lambda^q\right). \quad (21)$$

In (21), $\left\{m_\mu^q, s_\mu^{2^q}, a_\lambda^q, b_\lambda^q\right\}$ represents a set of "variational" parameters. For convenience and to keep the notational overhead at bay, we will denote the values of the variational parameters on the $i$th iteration of the VB algorithm introduced below as $m_\mu^{(i)}$, $s_\mu^{2^{(i)}}$, $a_\lambda^{(i)}$ and $b_\lambda^{(i)}$, dropping the "$q$" superscript. To reiterate, we have introduced a set of parameters $\left\{m_\mu, s_\mu^2, a_\lambda, b_\lambda\right\}$ governing the prior distribution $p(\mu, \lambda)$ of the generative model in (19). The set of prior parameters, in conjunction with the likelihood specification in (19) determines the log model evidence $\ln p(y)$. Additionally, we have introduced a set of variational parameters $\left\{m_\mu^{(i)}, s_\mu^{2^{(i)}}, a_\lambda^{(i)}, b_\lambda^{(i)}\right\}$ that govern the $i$th variational approximation $q^{(i)}(\mu) q^{(i)}(\lambda)$ to the posterior distribution $p(\mu, \lambda|y)$ and the $i$th variational free energy lower bound approximation $\mathsf{F}(q^{(i)}(\mu) q^{(i)}(\lambda))$ to the log model evidence $\ln p(y)$. Application of the iterative procedure of Table 1 for the current example then corresponds to an iterative procedure for updating $q^{(i)}(\mu)$ and $q^{(i)}(\lambda)$. Due to the conjugacy properties of the involved distributions, Steps (1) and (2) in Table 1 reduce to a parameter updating scheme for $m_\mu^{(i)}$, $s_\mu^{2^{(i)}}$, $a_\lambda^{(i)}$, and $b_\lambda^{(i)}$. The resulting update equations are stated and derived in Supplement Section 5. To obtain the lower bound variational free energy approximation to the log model evidence and to monitor the evolution of the VB algorithm, the defining integral of the variational free energy of Eq. (12) needs to be evaluated based on the example-specific definitions of $p(y, \vartheta) = p(y, \mu, \lambda)$ and $q(\vartheta) = q(\mu)\, q(\lambda)$. This derivation is also documented in Supplement Section 5. Fig. 5 visualizes the application of VB to the univariate Gaussian example.

## 4. VB for LGSSMs

In this and the following section we discuss how the VB theory sketched in Section 3 can be applied to the model class described in Section 2. To this end, we will first introduce the approach for a general univariate LGSSM in Sections 4.1–4.3 and subsequently exemplify the framework using the introductory perceptual learning example in Section 5.

We start by noting that if we are to apply the VB framework to an LGSSM as specified in Eq. (9), we have to render this equation a generative model of the form of the right-hand side of Eq. (10) by introducing marginal distributions over $\theta$ and $x_1$. We thus set $\theta := (a, \sigma_x^2, b, \sigma_y^2)$ and specify the following joint distribution

$$\begin{aligned} p(y_{1:T}, x_{1:T}, \theta) &= p(y_{1:T}, x_{1:T}|\theta)\, p(\theta) \\ &= p(y_{1:T}|\theta, x_{1:T})\, p(x_{2:T}|\theta, x_1)\, p(x_1) p(\theta)\,. \end{aligned} \quad (22)$$

For the remaining components of $\theta := (a, \sigma_x^2, b, \sigma_y^2)$, we here opted for a fully factorized prior distribution of the form

$$p(\theta) = p(a, \sigma_x^2, b, \sigma_y^2) := p(a) p(\sigma_x^2) p(b) p(\sigma_y^2) \quad (23)$$

corresponding to the assumption of full stochastic independence between the variables $a$, $\sigma_x^2$, $b$ and $\sigma_y^2$ in their marginal distribution. Specifically, we use univariate Gaussian distributions for $a$ and $b$ and univariate Gamma distributions for the inverse variances (precisions) $\lambda_x = \sigma_x^{-2}$ and $\lambda_y = \sigma_y^{-2}$ (see Fig. 3(D)). To apply the VB inference theorem for mean-field approximations (17) to the generative model of Eq. (22), we have to decide on a suitable partition for the unobserved variable set. Here, we choose

$$p(\theta, x_{1:T}|y_{1:T}) \approx q(\theta)\, q(x_{1:T})\,. \quad (24)$$

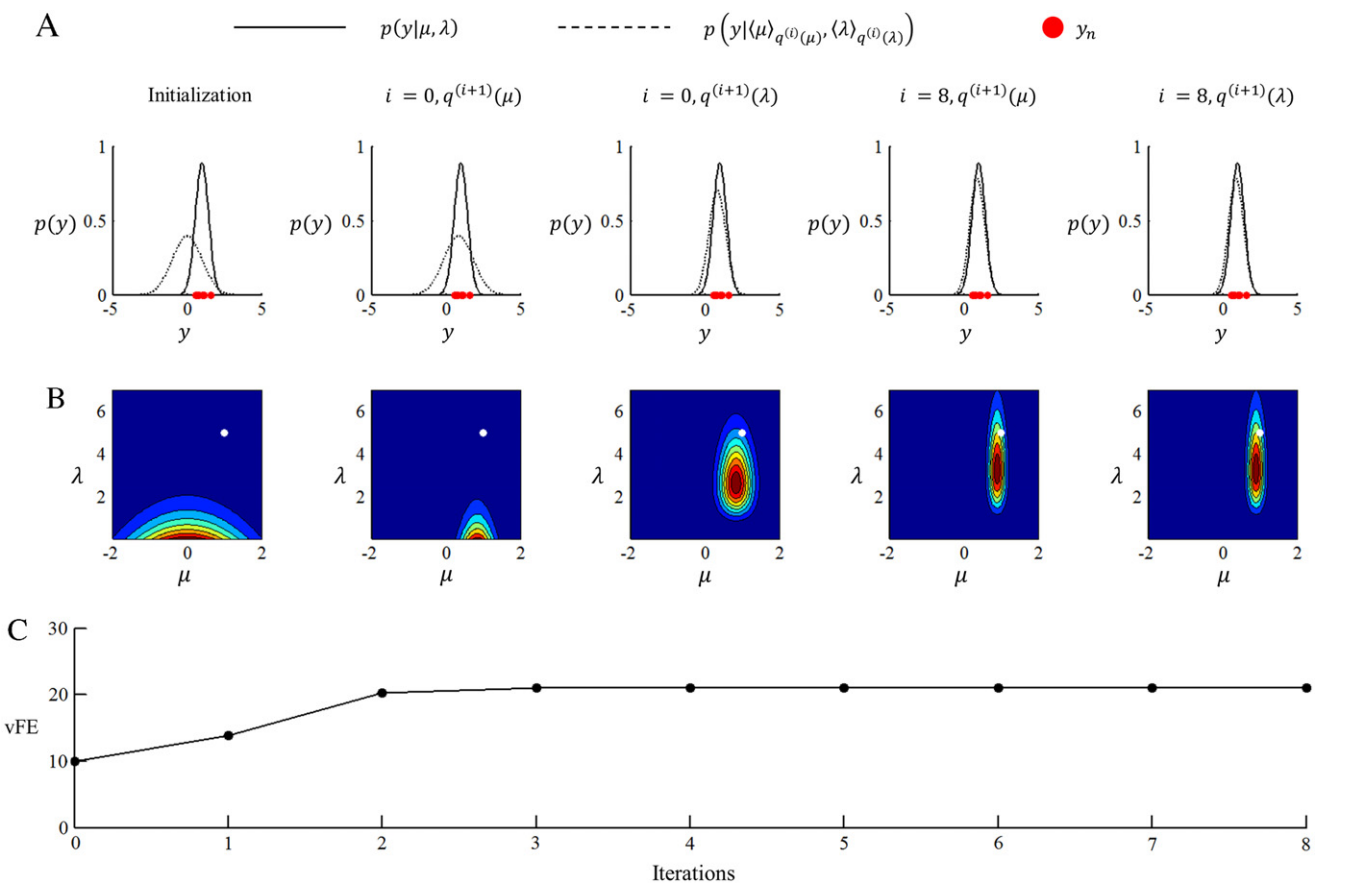


**Fig. 5. VB for the Gaussian.** (A) The panels of Fig. 5(A) depict the true underlying data model $p(y|\mu, \lambda)$, for $\mu = 1$ and $\lambda = 5$ as solid line and $N = 10$ samples $y_n$ from this model on the abscissa as red dots. Based on these samples, on each iteration of the VB algorithm, a variational approximation $q(\mu)q(\lambda)$ is updated. The first panel of (A) shows the univariate Gaussian model as approximated by the expectations over $q(\mu)$ and $q(\lambda)$ as dashed line. The second panel of (A) shows the effect of the update of the distribution $q(\mu)$ on the first iteration of the algorithm. As $q(\mu)$ governs the mean of the univariate Gaussian, the dashed Gaussian is now centered on the mean of the data-points. The third panel of (A) shows the effect of the update of the distribution $q(\lambda)$ on the first iteration of the algorithm. As $q(\lambda)$ governs the precision of the univariate Gaussian model, the dashed Gaussian updates its variance based on the data variability. The fourth and fifth panels of (A) show the corresponding two steps on the 8th iteration. (B) The panels of (B) show the factorized variational distribution $q(\mu)q(\lambda)$ over VB algorithm iterations. The white dot in each panel indicates the true underlying values that gave rise to the observed data. Note that these values were not sampled from the prior distribution, but that the prior distribution embeds the initial uncertainty about this true, but unknown, value before the observation of any data. The ordering of the panels is as in (A). (C) Fig. 5(C) shows the evolution of the variational free energy over iterations of the VB algorithm. For the current model and data set, the variational free energy levels off from approximately 4 iterations onwards. In the VB framework, the final value of the variational free energy after convergence of the algorithm corresponds to the approximation to the log model evidence $\ln p(y)$.

The partition above allows for developing an algorithmic VB approach for the derivation of the right-hand side of (24) in two steps: We first consider inferring $q(\theta)$ in Section 4.1. In Section 4.2 we then discuss inferring $q(x_{1:T})$, classically referred to as latent variable or state "inference" in the context of time-series modeling (Barber & Chiappa, 2007; Bishop, 2007). Finally, in Section 4.3 we discuss the variational free energy as an approximation to the log model evidence under the generative model of Eqs. (22) and (23) and the variational approximation of Eq. (24).

### 4.1. Evaluating $q(\theta)$

The VB inference theorem for mean-field approximations (Eq. (17)) provides a starting point for the derivation of the functional form and parameters of the updated variational distribution over $\theta$. The exact form that an algorithm implementing this theorem takes depends on the functional form chosen for $q(\theta)$. For the current generative model, any approximation to the posterior distribution $p(\theta|y_{1:T})$ naturally factors according to

$$q(\theta) = q\left(a, \sigma_x^2\right) q\left(b, \sigma_y^2\right) \tag{25}$$

due to the factorization properties of the LGSSM (see e.g. Beal (2003)). This simplifies the derivations, because the probability distribution over unobserved variables that govern the temporal evolution of unobserved variables $x_{1:T}$ (i.e., the distributions over $a, \sigma_x^2$), as well as the emission of observed variables $y_{1:T}$, (i.e., the distributions over $b, \sigma_y^2$) can be treated in isolation. The further specifics of the algorithmic implementation is then dependent on the form and parameterization of the variational distributions $q\left(a, \sigma_x^2\right)$ and $q\left(b, \sigma_y^2\right)$. For example, in Barber and Chiappa (2007) and Beal (2003), the distributions are treated as "true bivariate distributions" of the form

$$q\left(a, \sigma_x^2\right) = q(a|\sigma_x^2)q(\sigma_x^2) \quad \text{and} \quad q\left(b, \sigma_y^2\right) = q(b|\sigma_y^2)q(\sigma_y^2). \tag{26}$$

In our application, we additionally assume complete factorization of the form

$$q(\theta) = q(a)\, q(\sigma_x^2) q(b)\, q(\sigma_y^2) \tag{27}$$

i.e. the stochastic independence of the variables $a, \sigma_x^2, b$ and $\sigma_y^2$ in the approximate posterior distribution. As the prior distributions over $a$ and $b$ are set to Gaussian distributions $N\left(a; \mu_a, \sigma_a^2\right)$ and $N\left(b; \mu_b, \sigma_b^2\right)$, we exploit the properties of the exponential-conjugate class by specifying their variational counterparts as $N\left(b; \mu_b^{(i)}, \sigma_b^{2^{(i)}}\right)$ and $N\left(a; \mu_a^{(i)}, \sigma_a^{2^{(i)}}\right)$ for iterations $i = 1, 2, \ldots$, respectively (note that, as in Section 3, parameters of the

variational distributions are distinguished from the parameters of the prior distributions by the iteration superscript $(i)$). Likewise, as the prior distributions over the inverse variances $\lambda_x = \sigma_x^{-2}$ and $\lambda_y = \sigma_y^{-2}$ are set to Gamma distributions, their variational counterparts are specified as $G\left(\lambda_x; a_{\lambda_x}^{(i)}, b_{\lambda_x}^{(i)}\right)$ and $G\left(\lambda_y; a_{\lambda_y}^{(i)}, b_{\lambda_y}^{(i)}\right)$, respectively. Application of the VB inference theorem for mean-field approximations then allows for the derivation of a set of update equations for the variational parameters based on their initialization to the prior parameters. For the full details of this derivation and the resulting update equations, the reader is referred to Supplement Sections 6.1 and 6.2.

### 4.2. Evaluating $q(x_{1:T})$

Application of the VB inference theorem for mean-field approximations (17) to the generative model of Eq. (22) under the mean-field approximation of Eq. (24) implies that setting

$$q\left(x_{1:T}\right) \propto \exp\left(\int q\left(\theta\right) \ln p\left(y_{1:T}, x_{1:T}, \theta\right) d\theta\right) \tag{28}$$

maximizes the variational free energy. In Supplement Section 6.3 we show that the above implies that

$$q\left(x_{1:T}\right) \propto \exp\left(\int q\left(\theta\right) \ln p\left(x_{1:T}|y_{1:T}, \theta\right) d\theta\right). \tag{29}$$

If we assume for the moment that we know the values of the unobserved variables $\theta$, or in other words, that $p\left(\theta = \theta^*\right) = 1$ and $q\left(\theta = \theta^*\right) = 1$ for some value $\theta^*$ and that we set the proportionality constant, rendering $q\left(x_{1:T}\right)$ a probability distribution, to 1, then Eq. (29) simplifies to

$$q\left(x_{1:T}\right) = \exp\left(\ln p_\theta\left(x_{1:T}|y_{1:T}\right)\right) = p_\theta\left(x_{1:T}|y_{1:T}\right). \tag{30}$$

Note that the assumptions $q\left(\theta = \theta^*\right) = 1$ for some $\theta^*$ makes the integration operation redundant, while the assumption that $p\left(\theta = \theta^*\right) = 1$ allows for reformulating the conditional distribution as parameterized by, rather than conditional on, the joint distribution parameters $\theta$. (Also, note that, formally, the parameters of the conditional distribution $p\left(x_{1:T}|y_{1:T}\right)$ are implied by, but not identically to, the parameters of the joint distribution $\theta$. In the following, we will leave the parameterization by $\theta$ implicit.)

Eq. (30) implies that, under the assumption of non-random, known $\theta$, the evaluation of the variational distribution $q\left(x_{1:T}\right)$ is equivalent to the evaluation of the conditional distribution over the variables $x_{1:T}$ given an observed data sequence $y_{1:T}$. This corresponds to the classic "state space model inference" problem. For this problem, a variety of solutions exist in the literature: for a comprehensive review see e.g., Briers et al. (2004); for a review in the context of discrete-space, discrete-time models, i.e. Hidden Markov Models, see e.g. Rabiner (1989) and Visser (2011). It is one of the complicating features of the VB approach to LGSSMs that in contrast to these classic approaches, in the generative model and its ensuing variational approximation the values of $\theta$ are not fixed, but governed by probability distributions. This fact renders standard algorithms inappropriate for evaluating $q(x_{1:T})$, as will be discussed below. However, Barber and Chiappa (2007) have demonstrated a work-around, which allows to nevertheless rely on standard inference algorithms also in the context of VB for LGSSMs (see also Fujimoto, Satoh, and Fukunaga (2011)). In the following, we will thus first introduce algorithms for inferring $q(x_{1:T})$ for the case that the value of $\theta$ is known and non-random, and then explain, how these algorithms can be adapted if only probabilistic statements about the value of $\theta$ are available.

#### 4.2.1. Inference of $q(x_{1:T})$ for non-random $\theta$

As discussed in Section 2.2, $p\left(y_{1:T}, x_{1:T}\right)$ is a $2T$-dimensional Gaussian distribution over the observed and unobserved variables $y_{1:T}$ and $x_{1:T}$, respectively. Because conditional distributions of Gaussian distributions are again Gaussian distributions, $p\left(x_{1:T}|y_{1:T}\right)$ could in principle be determined by finding its $T$-dimensional expectation and its $T \times T$-dimensional covariance parameter based on the parameters of $p\left(y_{1:T}, x_{1:T}\right)$. However, classic state space model inference algorithms use a different parameterization of this distribution. Specifically, from the conditional independence properties of the LGSSM it follows that (see Supplement 6.3 for details)

$$\begin{aligned} p\left(x_{1:T}|y_{1:T}\right) &= p(x_1|y_{1:T}) \prod_{t=2}^{T} p\left(x_t|x_{t-1}, y_{1:T}\right) \\ &= p(x_1|y_{1:T}) \prod_{t=2}^{T} \frac{p(x_{t-1}, x_t|y_{1:T})}{p(x_{t-1}|y_{1:T})}. \end{aligned} \tag{31}$$

The inference algorithms discussed in the following yield the expectation and covariance parameters of the Gaussian distributions $p\left(x_t|y_{1:T}\right), t = 1, \ldots, T-1$ and $p\left(x_{t-1}, x_t|y_{1:T}\right), t = 2, \ldots, T$ in the numerator and denominator on the right-hand side of Eq. (31) in terms of the parameters of $p\left(y_{1:T}, x_{1:T}\right)$. More specifically, based on the results of the Kalman filter algorithm (Kalman, 1960), the so-called Kalman–Rauch–Tung–Striebel smoothing algorithm returns the parameters of $p\left(x_t|y_{1:T}\right), t = 1, \ldots, T$, and can readily be modified to also return the parameters of $p\left(x_{t-1}, x_t|y_{1:T}\right), t = 2, \ldots, T$. Together with (31) these parameters thus define $p\left(x_{1:T}|y_{1:T}\right)$. In the following, we give a high-level introduction to the algorithms for inferring $p\left(x_t|y_{1:T}\right)$ and $p\left(x_{t-1}, x_t|y_{1:T}\right)$ in the context of the univariate LGSSM studied here. For the mathematical derivations of these algorithms (i.e. answers to the question, why these algorithms actually work) and their generalization to multivariate LGSSMs we refer the reader to Supplement Section 6.3.

Before proceeding, it is helpful to differentiate the terms "filter algorithm" and "smoothing algorithm". As noted above, the aim of inference for LGSSMs is to derive probabilistic conclusions about the states of unobserved variables $x_{1:T}$ given a realization sequence of the observed variables $y_{1:T}$ and known values of $\theta$. For each $x_t, t = 1, \ldots, T$, these conclusions can in principle be derived from the joint distribution of the LGSSM as specified in Eq. (9) by appropriately conditioning on the remaining variables. Depending on whether in the conditional marginal distribution $p(x_t|y_{1:k})$ the value of $k$ is smaller than, equal to, or larger than the value of $t$, the algorithms for inferring $p(x_t|y_{1:k})$ take different forms and come under different labels (Briers et al., 2004). Specifically, for $k < t$, inferring $p(x_t|y_{1:k})$ is called "prediction," and will not be considered here. For $k = t$, inferring $p(x_t|y_{1:t})$ is referred to as "filtering". The Kalman filter thus allows to evaluate $p(x_t|y_{1:t})$ for $t = 1, \ldots, T$. As will be seen below, filtering also forms the basis for the case of $k > t$. Inferring $p(x_t|y_{1:T})$ is known as (fixed interval) "smoothing." Importantly, in addition to the information available from observing $y_{1:t}$ as in filtering, smoothing also takes into account the evolution of the observations after $x_t$ obtained a specific state. Intuitively, the smoothed conditional marginal distribution $p(x_t|y_{1:T})$ is hence more informative than the filtered conditional marginal distribution $p(x_t|y_{1:t})$.

*Filtering: Inferring* $p(x_t|y_{1:t})$. Inferring a probability distribution over the hidden variable $x_t$ given all observations up to $y_t$, i.e., $p(x_t|y_{1:t})$ is performed readily due to the fact that the LGSSM corresponds to a multivariate Gaussian distribution. As the joint distribution over $x_{1:T}$ and $y_{1:T}$ is Gaussian, all marginal and conditional marginal distributions are also Gaussian distributions. Further, as Gaussian distributions are characterized by their expectation and covariance parameters, it is only these parameters

that have to be inferred to fully characterize the filter distribution of interest

$$p\left(x_t|y_{1:t}\right) = N\left(x_t; \mu_{x_t|y_{1:t}}, \sigma^2_{x_t|y_{1:t}}\right) \tag{32}$$

for $t = 1, \ldots, T$. In (32) we have used $\mu_{x_t|y_{1:t}}$ and $\sigma^2_{x_t|y_{1:t}}$ to denote the parameters of the conditional distribution over $x_t$ given observations $y_{1:t}$. The Kalman filter corresponds to recursive expressions for the conditional expectation and covariance parameters $\mu_{x_t|y_{1:t}}$ and $\sigma^2_{x_t|y_{1:t}}$ of (32) in terms of the parameters of the "preceding distribution" $N\left(x_{t-1}; \mu_{x_{t-1}|y_{1:t-1}}, \sigma^2_{x_{t-1}|y_{1:t-1}}\right)$. For the univariate LGSSM considered here, the Kalman filter equations are given by

$$\mu_{x_t|y_{1:t}} = a\mu_{x_{t-1}|y_{1:t-1}} + k_t\left(y_t - ba\mu_{x_{t-1}|y_{1:t-1}}\right) \tag{33}$$

and

$$\sigma^2_{x_t|y_{1:t}} = (1 - k_t b)(a^2\sigma^2_{x_{t-1}|y_{1:t-1}} + \sigma^2_x) \tag{34}$$

where $k_t$ is the time-dependent Kalman gain factor (or more generally, the Kalman gain matrix, see Supplement Section 6.3.1). Intuitively, the update equation for the conditional mean $\mu_{x_t|y_{1:t}}$ in terms of $\mu_{x_{t-1}|y_{1:t-1}}$ may be interpreted from a "prediction-error-correction" perspective (Bishop, 2007; Friston, 2009), because the expectation of the latent variable $x_t$ is given by the sum of two terms: (1) the conditional expectation of the previous latent variable $x_{t-1}$ carried forward in time by the transition factor $a$ and (2) a correction term measuring the deviation between the observed value $y_t$ and its prediction based on $\mu_{x_{t-1}|y_{1:t-1}}$ given by $ba\mu_{x_{t-1}|y_{1:t-1}}$ (where the amount of correction is encoded by the Kalman gain factor $k_t$). Notably, the filtered variance $\sigma^2_{x_t|y_{1:t}}$ is independent of the observations $y_{1:t}$ and depends only on the parameters of the LGSSM. To summarize, based on the expected value and the variance of $x_1$ given $y_1$ the parameters of the distributions $p(x_t|y_{1:t})$ can successively be obtained for $t = 2, \ldots, T$ using Eqs. (33) and (34).

*Smoothing: Inferring* $p(x_t|y_{1:T})$. Like the conditional marginal distribution $p(x_t|y_{1:t})$ the conditional marginal distribution $p(x_t|y_{1:T})$ of an LGSSM is a Gaussian distribution and, hence, can be characterized by its expectation and covariance parameters. In analogy with Eq. (32), the KRTS smoothing algorithm evaluates the parameters of

$$p\left(x_t|y_{1:T}\right) = N(x_t; \mu_{x_t|y_{1:T}}, \sigma^2_{x_t|y_{1:T}}) \tag{35}$$

in terms of the parameters of the "succeeding distribution" $N(x_{t+1}; \mu_{x_{t+1}|y_{1:T}}, \sigma^2_{x_{t+1}|y_{1:T}})$ and the results of the Kalman filter. For the LGSSM under consideration, the KRTS equations are given by

$$\mu_{x_t|y_{1:T}} = \mu_{x_t|y_{1:t}} + j_t\left(\mu_{x_{t+1}|y_{1:T}} - a\mu_{x_t|y_{1:t}}\right) \tag{36}$$

and

$$\sigma^2_{x_t|y_{1:T}} := \sigma^2_{x_t|y_{1:t}} + j_t^2\left(\sigma^2_{x_{t+1}|y_{1:T}} - \sigma^2_{x_{t+1}|y_{1:t}}\right) \tag{37}$$

where $j_t$ is a correction factor analogous to the Kalman gain factor. Thus, based on the initialization of $\mu_{x_T|y_{1:T}}$ and $\sigma^2_{x_T|y_{1:T}}$, which correspond to the final results of the Kalman filter, the parameters of the distributions $p\left(x_t|y_{1:T}\right)$ for $t = T-1, T-2, \ldots, 1$ can successively be obtained using Eqs. (36) and (37). The most likely state of variables $x_{1:T}$, corresponding to the expectations $\mu_{x_t|y_{1:T}}$ of the unimodal conditional marginal distribution $p\left(x_t|y_{1:T}\right)$ can then be viewed as the "best guess" of the state of the latent variable and completes the classical inference problem for LGSSMs.

*Inferring* $p(x_{t-1}, x_t|y_{1:T})$. In addition to the most likely state of the unobserved variables $x_t$ in light of the data $y_{1:T}$, we also need to infer the joint distribution of temporally adjacent variables $x_{t-1}$ and $x_t$, $t = 2, \ldots, T$, because, together with the parameters of $p\left(x_t|y_{1:T}\right)$, the parameters of $p\left(x_{t-1}, x_t|y_{1:T}\right)$ define the distribution $p\left(x_{1:T}|y_{1:T}\right)$ (cf. Eq. (31)). From the perspective of unknown parameters $\theta$, inferring $p(x_{t-1}, x_t|y_{1:T})$ also has the benefit of obtaining an indirect estimate of the evolution parameter by means of the average conditional covariation of $x_{t-1}$ and $x_t$ for all $t = 2, \ldots, T$. This can be understood in analogy to the correlation coefficient transformation in bivariate Gaussian scenarios to the slope coefficient of simple linear regression models (see for example (Hays, 1994)), or more formally, from the viewpoint of "sufficient statistics". The distribution $p\left(x_{t-1}, x_t|y_{1:T}\right)$ is a bivariate Gaussian distribution and hence characterized uniquely by its $2 \times 1$ expectation vector $\mu_{x_{t-1,t}|y_{1:T}}$ and its $2 \times 2$ covariance matrix with off-diagonal elements $\sigma^2_{x_{t-1,t}|y_{1:T}}$. The expectation parameters can be evaluated by concatenating the expectations of $x_{t-1}$ and $x_t$. Finally, a backward recursion for the off-diagonal elements of the covariance matrix parameters is given by

$$\sigma^2_{x_{t-1,t}|y_{1:T}} := j_{t-1}\sigma^2_{x_t|y_{1:T}}. \tag{38}$$

This concludes our discussion of inferring the variational distribution $q\left(x_{1:T}\right)$ for the case that the values of the unobserved variables $\theta$ are known. As noted above, the mathematical reasons, why the recursive scheme of equations (32)–(38) works, and how the factors $k_t$ and $j_t$ are computed is explained for general, multivariate LGSSMs in Supplement Section 6.3.

#### 4.2.2. Inference of $q(x_{1:T})$ for random $\theta$

In contrast to classic applications of LGSSM inference algorithms in which the values of the unobserved variables $\theta$ are known, only the variational distribution $q(\theta)$ is known in the VB context. Intuitively, one may think that the KRTS inference machinery may still be used by simply forwarding the expectation of $\theta$ under $q(\theta)$ to the respective algorithms. However, as noted in Beal (2003) and Barber and Chiappa (2007) this is not appropriate. In fact, one can show that for the LGSSM the expectation of the log conditional joint distribution under $q(\theta)$ does not correspond to the log conditional joint distribution with expected value of $\theta$ under $q(\theta)$. This inequality is summarized in Barber and Chiappa (2007) as the "mean and fluctuation decomposition theorem". Specifically, in Barber and Chiappa (2007) the authors show that the expectation of the log joint distribution under $q(\theta)$ can be written as the sum of two components: (1) the log joint distribution conditioned on the expectation of $\theta$ under $q(\theta)$ and (2) "fluctuation" terms. Given that the additional "fluctuation" terms do not vanish in but trivial cases, standard LGSSM inference algorithms supplied with the expected value of $\theta$ under $q(\theta)$ are thus not appropriate for inference in the VB context. In order to nevertheless use standard inference algorithms in the VB framework, and hence to capitalize on the vast literature that has dealt with the numerical optimization of these algorithms in the past, (Barber & Chiappa, 2007) proposed a "unified inference theorem". The idea is to reformulate the argument of the exponential in Eq. (28) as a standard LGSSM $\tilde{p}_{\tilde{\theta}}\left(\tilde{y}_{1:T}, x_{1:T}\right)$ with known parameters $\tilde{\theta}$ by appropriately augmenting the VB–LGSSM's observed variables and unobserved parameters (here we use the tilde notation to distinguish reformulated entities from their originals). The classic inference algorithms discussed in Section 4.2.1 may then be applied to $\tilde{p}_{\tilde{\theta}}\left(\tilde{y}_{1:T}, x_{1:T}\right)$ to derive the parameters of $\tilde{p}_{\tilde{\theta}}\left(x_{1:T}|\tilde{y}_{1:T}\right)$. These parameters in turn correspond to the parameters of the argument of the exponential in Eq. (29). The details of the "mean and fluctuation decomposition" and "unified inference" theorems are discussed in Supplement Section 6.3.2.

In summary, the VB algorithm for LGSSMs can practically be applied by iterating (1) the evaluation of analytically derived variational parameter update equations, which are dependent

on the chosen form of the variational distributions, and (2) the application of a KRTS smoothing algorithm to a suitably augmented LGSSM.

### 4.3. Evaluating the variational free energy

In order to obtain an approximation to the log model evidence using the VB approach, the variational free energy integral defined in Eq. (13) has to be evaluated for the chosen generative model and variational distribution. As shown in Supplement Section 5 this term can be rewritten as the sum of two contributions (Penny, 2012): a positive average likelihood term, capturing the goodness-of-model-fit ("accuracy"), and a negative KL-divergence term between the prior and variational (approximated posterior) distribution. The latter term represents a penalty term for model complexity: it penalizes those models for which a large adjustment of model parameters is required to explain the data. To conveniently evaluate the variational free energy for the LGSSM, however, we use an alternative decomposition comprising an "energy" and an "entropy" term, see Supplement Section 6.4 for details.

## 5. A tutorial example

In this section, we return to the reaction time example discussed in the introduction (readers interested in the mathematical details of this example may consult Supplement Section 6). Based on the developments of Sections 2–4, we now assume that a researcher is interested in modeling the observed data using a latent linear stochastic time-series model of the form

$$\begin{aligned} dX(t) &= \alpha X(t)\,dt + \sigma\,dW(t), \quad t \in [0, s] \\ y_{t_i} &= bX(t_i) + c_0 + \eta_{t_i}, \quad t_i \in [0, s], i = 0, 1, \ldots, n. \end{aligned} \tag{39}$$

In the current scenario, "time" actually refers to the course of the training regime and the collection of test median reaction times. $c_0$ is a constant that is assumed to be known, and which will be discussed in some more detail below.

As mentioned in Section 2, the first step to a VB analysis of (36) in the framework discussed here is the discretization of the latent continuous time process based on the parameter transformation formulas $a = (1 + \alpha\Delta t)$ and $\sigma_x^2 = \sigma^2\Delta t$, yielding

$$\begin{aligned} x_t &= ax_{t-1} + \varepsilon_t \\ y_t &= bx_t + c_0 + \eta_t \end{aligned} \tag{40}$$

for $t = 2, \ldots, T$ and for $t = 1, \ldots, T$, respectively. To generate the data of Fig. 1, we set $T = 50$, $\Delta t = 1$, $\alpha = -0.1$ (i.e., $a = 0.9$), $\sigma = 0.1$ (i.e., $\sigma_x^2 = 0.01$), $b = -200$, $\sigma_y^2 = 0.9$, $c_0 = 100$, $x_1 = -1$, and used the LGSSM formulation to sample $x_{1:50}$ and $y_{1:50}$, of which $y_{1:50}$ are shown as the "median reaction time" data.

As discussed in Section 3 and the beginning of Section 4, the second step to a VB analysis of (39) is to embed (40) in a generative model, or, in other words, to specify prior distributions over the unobserved variables of interests, $a$, $\sigma_x^2$, $b$ and $\sigma_y^2$. Here, we opted for low informative (wide variance) prior distributions over the evolution and emission variables $a$ and $b$ and semi-informed priors over the volatility and noise parameters $\sigma_x^2$ and $\sigma_y^2$. More specifically, we set the prior distributions over $a$ and $b$ to Gaussian distributions with parameters $\mu_a := 0$, $\sigma_a^2 := 10^3$ and $\mu_b := -190$, $\sigma_b^2 := 10^3$, respectively. The prior distributions over the precisions $\lambda_x$ and $\lambda_y$ were set to Gamma distributions with parameters $a_{\lambda_x} := 10^{-1}$, $b_{\lambda_x} := 10^3$, $a_{\lambda_y} := 1$ and $b_{\lambda_x} := 10^2$. Together with the factorization property (23) and the likelihood expressed by (37) this completes the specification of the generative model.

To obtain posterior distributions over $a$, $\lambda_x$, $b$ and $\lambda_y$ and an approximation to the model log evidence, we then apply the algorithmic VB procedure as discussed in Sections 4.1–4.3. This approach is visualized in Fig. 6 for the first and eighth iteration upon initialization. In the current scenario, convergence is achieved quickly, and we can proceed to the evaluation of the results.

In Fig. 7(A), we reprise the reaction time data of Fig. 1, but now additionally plot the sampled unobserved time series $x^*_{1:50}$, corresponding to the evolution of the perceptual discrimination task over the course of the training regimes as red dots. Moreover, we also plot the inferred variational expectation of $x_{1:50}$, denoted as $\langle x_{1:50}\rangle_{q(x_{1:50})}$ upon convergence of the VB algorithm as white dots. In fact, these lie on top of the sampled data. In an experimental context, the variational expectations correspond to the "best guess" on the unobserved perceptual cognitive state. In Fig. 7(B) we are concerned with Bayesian posterior parameter inference. Specifically, using the integral transform for probability density functions, we depict the prior and posterior distributions over the latent continuous time model variables $\alpha$ and $\sigma$, and the prior and posterior distributions over the discrete time emission variables $b$ and $\sigma_y^2$. As an example for Bayesian posterior parameter inference, we may note that most of the probability mass (i.e. a 95%-credible interval) for the perceptual learning rate parameter $\alpha$ is concentrated between $-0.21$ and $-0.052$, covering the pre-specified value of $\alpha = -0.1$. Likewise, the volatility parameter $\sigma$ most supported by the data, i.e. its 95%-credible interval, lies between 0.083 and 0.12, covering the pre-specified value of $\sigma = 0.1$. By visual inspection, we note that most of the posterior probability mass for the emission parameter is centered at $b = -201$, slightly deviating from the simulated value of $b = -200$, while, for the current realization of the LGSSM, the observation noise parameter $\sigma_y^2$ is clearly underestimated. The high confidence associated with these biased estimates may be the result of the known over-compactness of mean-field approximations (e.g. Bishop, 2007; Daunizeau et al., 2011).

It is noteworthy, that credible intervals specify the posterior probability of a parameter lying in a specific interval in parameter space given the observed data. They should not be confused with confidence intervals used in classical statistics, which specify the probability of hypothetical data giving rise to an observed statistic under the assumption of a fixed and known parameter (Herzog & Ostwald, 2013). Finally, the evaluation of the variational free energy upon convergence of the VB algorithm as depicted in Fig. 6(B) allows for the specification of an approximation to the log model evidence. Of course, by itself the specific numerical value of this approximation is rather meaningless, as it depends on the form of the generative model and the data sample. However, as noted in the introduction, it represents the necessary prerequisite for formal Bayesian model comparison.

Some readers may wonder whether the assumption of a known constant $c_0$ is a limitation of the variational Bayes approach to stochastic time-series models and what it is meant to represent in the current example. Informally, the AR(1)-process form of the LGSSM (37) corresponds to a "stochastic" exponential decay towards zero. If this process is initialized at a negative value, it thus increases, and approaches zero from below. This is the behavior we posited for the latent perceptual learning process. To result in a decrease of corresponding "reaction times", the value of $b$ is thus required to be negative. Finally, to generate artificial reaction times that do not approach zero (because this is implausible from a biologically viewpoint), we introduced the constant $c_0$ to maintain final reaction times around approximately 100 ms in Fig. 1. In other words, the constant $c_0$ was merely introduced to make the example more realistic. Of course, $c_0$ could also be estimated, i.e. assume the status of an unobserved variable in the model. However, accommodating the framework for its estimation requires a somewhat more general view than the route taken in this tutorial. More specifically, in analogy to simple linear regression models in the context of the general linear model, it requires the augmentation

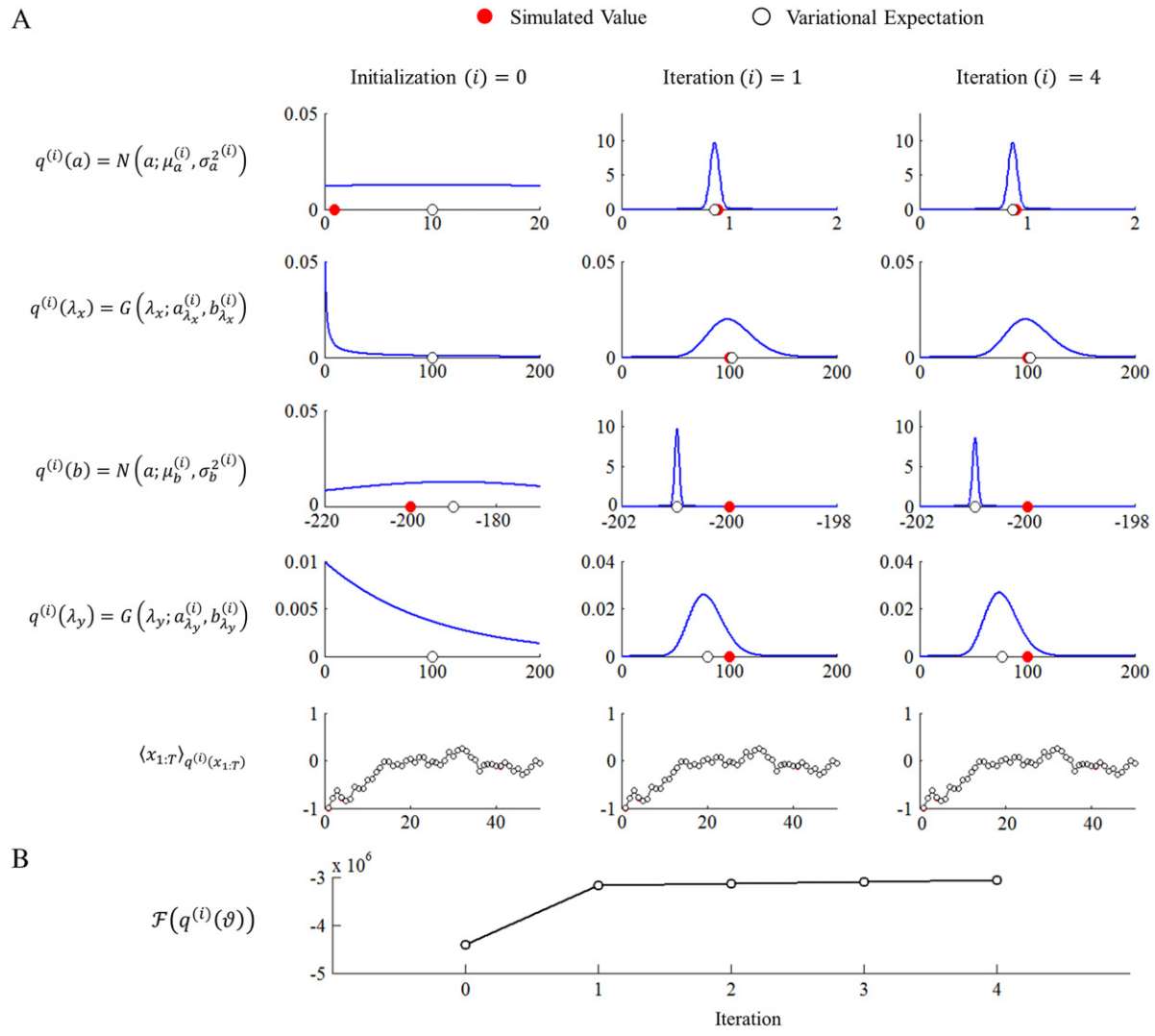


**Fig. 6. VB algorithm for LGSSMs** (A) Fig. 6(A) visualizes the variational distributions over the unobserved variables $\vartheta$ of the LGSSM upon initialization (leftmost column) and for the first and fourth iterations of the algorithm (middle and rightmost columns). Simulated expectations are shown as red dots, variational expectations are depicted as white dots. The variational probability density functions are shown as blue lines for unobserved variables $a$, $\lambda_x$, $b$ and $\lambda_y$. Low informative priors were chosen for $a$, $b$ and semi-informative priors were chosen for $\lambda_x$ and $\lambda_y$, corresponding to the variational probability density functions at initialization. (B) Fig. 6(B) tracks the variational free energy for the initialization and first four iterations. For the generative model chosen and the sampled data realization saturation of the variational free energy can be observed already at the second iteration. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

of the unobserved variable time-series to a two-dimensional column vector $\begin{pmatrix} 1 & x_t \end{pmatrix}^T$ and rendering the emission variable $b$ a two-dimensional row vector $B = (c_0 \quad b)$. This thus leads into the realm of multivariate LGSSMs, which we meant to eschew in this tutorial.

Finally, we would like to stress that although the tutorial example exploits psychological (or rather psychophysical) connotations, we do not mean to imply that VB for linear stochastic time-series offers a new method for reaction time analysis in perceptual learning scenarios. The example was merely chosen as a means to provide some intuitions for the more general concepts discussed in Sections 2–4. While it is of course not impossible to use the method discussed here in the context of learning experiments (see also Section 6.2), we happily admit that the tutorial example is somewhat contrived and in no way constitutes a novel general purpose method for psychological data analysis. It merely serves as a vehicle to demonstrate (1) how combining a latent SDE model with VB allows for inference of unobserved cognitive processes from observed data, (2) how the approach allows for the evaluation of posterior uncertainty about the unobserved variables involved, and (3), by means of the variational free energy, provides the necessary prerequisite to compare alternative formalized explanations of observed data.

## 6. Generalizations and applications

### 6.1. Generalizations and technical alternatives

The VB approach to latent stochastic time-series models is a statistical method with a wide range of applications. In fact, in its generalized form comprising an arbitrary number of hierarchical layers and allowing for nonlinear evolution and emission functions, it may be viewed as a general inference procedure for most commonly employed statistical models (including "static" models such as the general linear model and its descendants, factor analysis, and independent component analysis (Friston, 2008a,b; Roweis & Ghahramani, 1999)). We would like to emphasize that in this tutorial, we have only discussed a very specialized application of this general case.

A possible extension to nonlinear evolution and emission functions is presented in Daunizeau et al. (2009). The framework

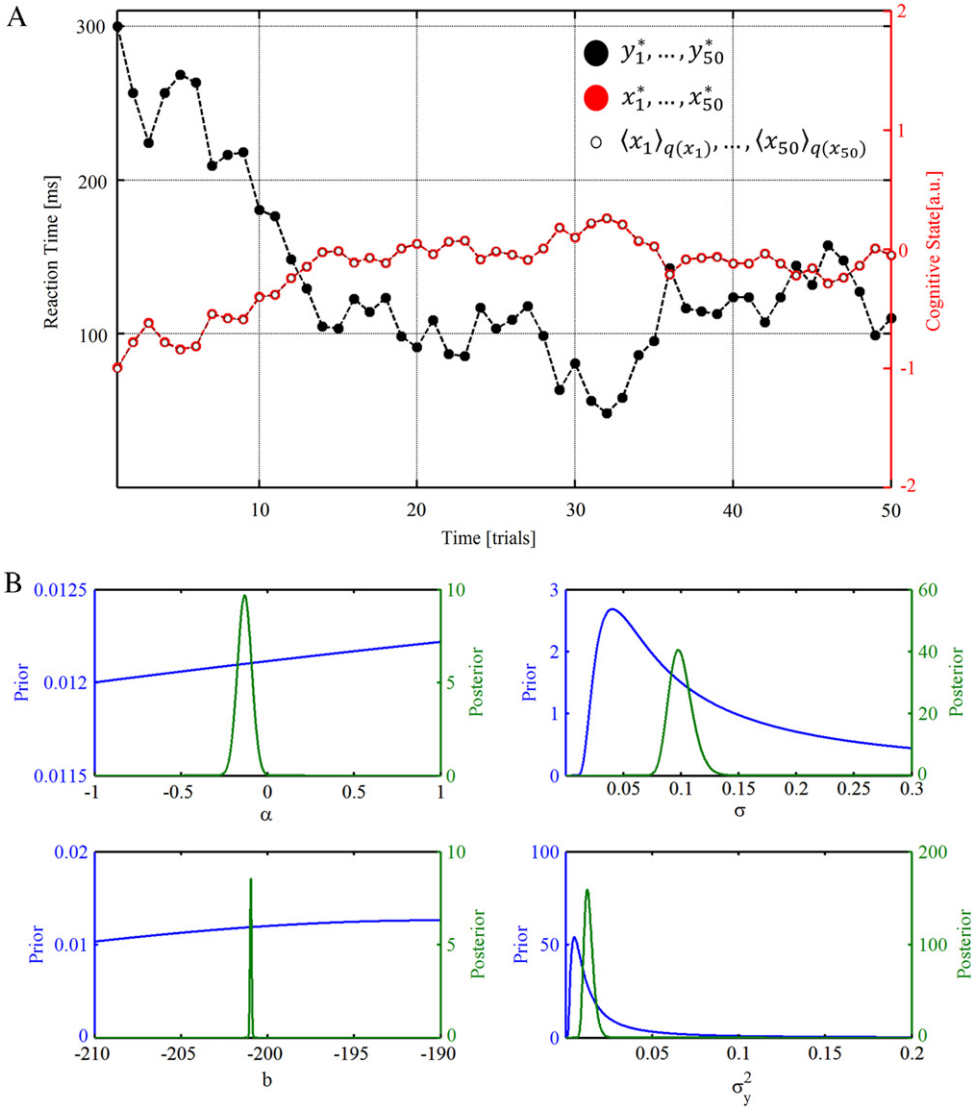


**Fig. 7. VB for the reaction time—perceptual learning example** (A) Fig. 7(A) shows the reaction time data of Fig. 1 $y^*_{1:50}$ as black dots and the sampled underlying evolution of the perceptual discrimination process $x^*_{1:50}$ as red dots. On top of this, the inferred variational expectations $\langle x_t \rangle_{q(x_t)}$ $(t = 1, \ldots, 50)$ upon convergence of the VB algorithm are depicted as white dots. (B) Fig. 7(B) depicts the prior and variational posterior distributions over the unobserved variables $\alpha$, $\sigma$, $b$ and $\sigma_y^2$ that parameterize the learning curve data model. Based on the variational posteriors shown, Bayesian credible intervals can be derived as discussed in the main text. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

discussed therein is based on a latent stochastic differential equation, as in the current tutorial. However, in contrast to the framework discussed here, the framework of Daunizeau et al. (2009) is applicable to arbitrary nonlinear SDEs and emission functions and thus a much wider range of statistical models. To allow for VB inference Daunizeau et al. (2009) use a local linearization approach in terms of the system's Jacobian and a Laplace (Gaussian) approximation to the posterior distributions over unobserved variables. Archambeau and colleagues (Archambeau, Cornford, Opper, & Shawe-Taylor, 2007; Archambeau, Shen, Cornford, & Shawe-Taylor, 2008) offer a different perspective on VB for SDE models by computing posterior distributions over infinite dimensional sample paths. In their case, a Euler–Maruyama discretization step is eschewed and the approach is rooted in the approximation of Gaussian processes (Rasmussen & Williams, 2006).

Finally Friston et al. (2008) have developed a VB inference procedure for random dynamical systems in generalized coordinates under the label "Dynamic Expectation–Maximization (DEM)". This framework differs from the one considered in this tutorial as no backward smoothing pass is required and thus operates in an online fashion. Additionally, in contrast to the latent Wiener process of the current tutorial and the nonlinear framework of Daunizeau et al. (2009), the additive noise terms are assumed to be differentiable. As discussed in Daunizeau et al. (2012), this assumption renders the model assumptions more constrained with respect to SDEs and thus potentially more powerful for inference in realistic dynamic systems. Like the framework by Daunizeau et al. (2009), it relies on a Laplace approximation to the posterior distribution. DEM has later been generalized to "Variational Filtering" (Friston, 2008a,b) by disposal of the Laplace approximation, but under retention of the mean-field approximation, and "Generalized Filtering" (Friston et al., 2010; Li et al., 2011) by disposal of the mean-field approximation, but under retention of the Laplace approximation.

With respect to algorithmic approaches for posterior distribution inference, approximate Bayesian techniques can roughly be categorized into sampling-based and deterministic approaches. A principle alternative to the VB approach for probabilistic time-series models are thus Markov chain Monte Carlo (MCMC) approaches. A simple and general answer to the question of whether an analytic or sampling-based approach for posterior

distribution estimation is preferable is unlikely to exist, as each specific instantiation of an algorithmic approach in a specific modeling context will have advantages and disadvantages (Barber, 2012). Generally speaking, it may be the case that MCMC methods afford exact solutions for the posterior distribution in low-dimensional parameter space scenarios, while VB methods allow for tackling high-dimensional parameter space scenarios under minimal computational cost with reasonable results (Shen et al., 2010). Currently, new VB approaches are usually validated against MCMC methods in low-dimensional scenarios and have repeatedly been shown to produce comparable, but computationally more efficient, results. Like for MCMC approaches (Lunn, 2013), there exists a general purpose, user-friendly implementation of VB methods in the form of variational message passing and the associated VIBES software package (Winn, 2005). Based on a graphical model framework, VIBES allows for the derivation of VB algorithms for a wide range of models with a significant reduction in the necessary algebraic derivations compared to the full analytical approach taken in the Supplement.

With respect to the question of Bayesian model selection Penny (2012) has recently compared the Bayesian information criterion (BIC) (Schwarz, 1978), Akaike's information criterion (AIC) (Akaike, 1974) and the "corrected" AIC (AICc) (Hurvich & Tsai, 1989) with the variational free energy. This comparison was carried out in the context of the general linear model and non-stochastic time-series models of the "Dynamic Causal Model"-type (Friston et al., 2003) for functional magnetic resonance imaging. As discussed in Penny (2012) it can be shown that BIC corresponds to a special variational free energy approximation to the model evidence in the infinite data limit under uninformative priors and exact variational inference. Because AIC and AICc do not correspond to approximations to the log model evidence, but favor models of minimal expected KL-divergence with respect to the true model (Burnham & Anderson, 2004), their formal comparison to the variational free energy is averted. Using numerical simulations, Penny (2012) showed that the relative ability of AIC, AICc, BIC, and the variational free energy to recover the true underlying models depends on the specific model type and the signal-to-noise ratio chosen. More specifically, in the general linear model context, AIC and BIC were more error-prone than the variational free energy at low signal-to-noise ratio, an effect that decreased in higher signal-to-noise ratio scenarios. For the specific time-series model chosen, the variational free energy was found to provide the best model selection ability. Naturally, these findings are only suggestive, as they result from numerical simulation rather than analytical considerations, and are constrained to the particular generative model class studied. A formal and numerical investigation of which model comparison criterion best to employ may thus be a valuable endeavor in concrete applications of the framework discussed here (see also Wipf and Nagarajan (2009) for the influence of Laplace approximations to the lower-bound property of variational free energy with respect to the log model evidence).

### 6.2. Applications in mathematical psychology

The general framework sketched in this tutorial can be applied to a wide range of modeling problems in mathematical psychology, provided the researcher is interested in the following: from the model formulation perspective, the researcher should be interested in an unobserved (cognitive) process that unfolds over time and exhibits some degree of variance or volatility. As demonstrated by the reaction time example, VB for stochastic time-series models then allows for the derivation of probabilistic statements about the temporal evolution of the latent process and the parameters governing its deterministic and probabilistic dynamics. Whether the latent process is considered in continuous time (i.e. SDE-like) or discrete time (i.e. LGSSM-like) is not crucial, the framework considered here accommodates both cases. It should be noted, however, that one fundamental aspect of the modeling framework of this tutorial is that the unobserved process allows for its indirect observation at each time-point. From a model estimation and evaluation perspective, the use of VB for stochastic time-series models is conditional on the researcher's interest in a computationally-efficient, analytical Bayesian evaluation of the formulated model and an overall aim to quantify the uncertainty over time-varying and time-invariant unobserved variables in light of data. Finally, from a model comparison perspective, the use of the variational free energy is conditional on an interest in comparing different models in their plausibility to explain a given data set with the fundamental aim of Bayesian model selection. In the following, we highlight some concrete cases of relevance to mathematical psychology. For more comprehensive reviews of the broad applicability of (stochastic) dynamical system theory in psychological and cognitive sciences, we refer the interested reader to the books by Port and Van Gelder (1995) and Ward (2002).

The most direct application to phenomena of interest in mathematical psychology is a characterization of the temporal evolution of cognitive processes in trial-by-trial modeling and longitudinal data (see e.g. Pernet, Sajda, and Rousselet (2011), Ward (2002)). Here, we highlight the examples of Daunizeau, den Ouden, Pessiglione, Kiebel, Friston et al. (2010), Daunizeau, den Ouden, Pessiglione, Kiebel, Stephan et al. (2010) and Mathys, Daunizeau, Friston, and Stephan (2011). Daunizeau, den Ouden, Pessiglione, Kiebel, Friston et al. (2010) and Daunizeau, den Ouden, Pessiglione, Kiebel, Stephan et al. (2010) established a meta-Bayesian approach (experimental Bayesian inference about psychological processes modeled as Bayesian inference) in the context of an audio–visual associative learning task. Specifically, the technical approach employed by Daunizeau, den Ouden, Pessiglione, Kiebel, Friston et al. (2010) and Daunizeau, den Ouden, Pessiglione, Kiebel, Stephan et al. (2010) rests on a combination of the VB approach to stochastic time-series models as discussed here under (inverse) Bayesian decision theory optimality criteria (see for example DeGroot (2004) and Robert (2007) for a discussion of Bayesian decision theory). Based on the trial-by-trial fluctuations in reaction times, the authors were able to show that the participants' behavior was more plausibly explained by a model that exploited the dynamic probabilistic nature of the experimental task than by a model that ignores changes in the statistical regularities of the environment. More generally, the framework developed by Daunizeau, den Ouden, Pessiglione, Kiebel, Friston et al. (2010) and Daunizeau, den Ouden, Pessiglione, Kiebel, Stephan et al. (2010) allows for probabilistic inference about the dynamic observer representations of temporally evolving environmental states and, by means of Bayesian model selection, to discern the preference structure (loss/utility function) participants may employ under constrained optimality criteria. Mathys and coworkers (Mathys et al., 2011) extended this work with a focus on (1) connecting classic reinforcement learning models (e.g. Sutton & Barto, 1998) with a Bayesian perspective and (2) overcoming limitations of ideal Bayesian learning models, such as the implicit computational complexity, questionable biological implementation and failure to account for individual differences. Specifically, in comparison to the example considered here, (Mathys et al., 2011) focused on inference in a generative model of deeper hierarchical structure: the generative model in Mathys et al. (2011) comprises two latent levels evolving according to Gaussian random walks and a Bernoulli distribution emission function. The specific application pursued here is to create learning models capable of describing subjectively optimal behavior, which are objectively maladaptive, in the context of understanding psychiatric diseases.

Another area of application of hierarchical time-series models in psychology is the analysis of longitudinal single- and multi-subject data obtained in the study of learning and/or developmental questions. As an example, Oravecz, Tuerlinckx, and Vandekerckhove (2011) introduced a hierarchical Ornstein–Uhlenbeck process model for continuous repeated measurement data with the aim of capturing inter-individual differences in the context of affect core dynamics. This study may be regarded as a prime example of the vast applicability of stochastic time-series models to the analysis of the structural and random factors contributing what may best be characterized as "individual pathways of change" (Molenaar & Newell, 2010).

The relationship between SDE approaches to reaction-time modeling, usually referred to as drift–diffusion models (DDMs) (Ratcliff, 1978; Ratcliff & Van Dongen, 2011; Smith & Ratcliff, 2004) and value-based decision making (Busemeyer et al., 2006; Busemeyer & Townsend, 1993; Huang et al., 2012) and the approach discussed here is less direct: in the application of DDM-like models to reaction time data, the temporal evolution of the process of interest is only observable by means of its distribution of endpoints, i.e. reaction time distributions for correct or false responses. This fact necessitated the development of sophisticated temporal integration schemes for the experimental evaluation of DDM models, see e.g. Ratcliff and Tuerlinckx (2002), Vandekerckhove and Tuerlinckx (2007) and Wagenmakers, van der Maas, and Grasman (2007). As noted above, one necessary condition for the applicability of the current framework is the (indirect) observation of the underlying process at discrete time-points while it evolves. Nevertheless, the DDM approach has recently been noticed as a fruitful cross-section between formal mathematical psychology and neuroimaging research, see for example Busemeyer et al. (2006), Hunt et al. (2012), O'Connell, Dockree, and Kelly (2012), Philiastides, Auksztulewicz, Heekeren, and Blankenburg (2011), Philiastides, Ratcliff, and Sajda (2006) and Ratcliff, Philiastides, and Sajda (2009). This is because neuroimaging methodology such as magneto/encephalography (M/EEG) allows, at least in principle, to track the evolution of cognitive processes while they evolve. The VB approach for stochastic time-series models thus has high potential as a method of choice in the future formal integration of mathematical psychology and brain imaging methodology.

For mathematical psychologists with an interest in biology it may further be noteworthy that the framework discussed here corresponds closely to the standard problem of neuroimaging time-series analysis. In the M/EEG domain for example, the aim is to characterize the cortical dynamics of neuronal population activity, but the data at hand only correspond to an indirect observation of this process from electrical fluctuations measured at the scalp. Unsurprisingly, VB methods for time-series data have been crucial in the development of biologically meaningful approaches to M/EEG time-series analysis, such as Dynamic Causal Modeling for event-related potentials (David, Harrison, & Friston, 2005; David et al., 2006), time–frequency domain signals (Chen, Kiebel, & Friston, 2008; Chen et al., 2012) and neural field activity (Moran, Pinotsis, & Friston, 2013; Pinotsis, Moran, & Friston, 2012). Likewise, it has long been noticed that fMRI data corresponds to the indirect observation of neural activity through the filter of the hemodynamic response (Logothetis, 2008). Further, fMRI data exhibits time-series data characteristics such as serial correlations of noise and thus "static" models such as the general linear model are not necessarily the most meaningful approach to fMRI time series analysis, see for example (Woolrich, Ripley, Brady, & Smith, 2001; Worsley & Friston, 1995). This insight led to the development of Dynamic Causal Modeling for fMRI (Friston et al., 2003; Stephan & Friston, 2010). Recently, the significance of "neural noise" has received increasing attention in the literature (Daunizeau et al., 2012; Li et al., 2011), specifically with respect to the characterization of endogenous brain dynamics at rest (Raichle, 2009).

Finally, the VB approach to general stochastic time-series models corresponds closely to the mathematical formulation of one of the most comprehensive global brain theories to date, a body of work usually referred to as the "Free Energy Principle (FEP)" (for reviews, see Friston (2010, 2009), Friston, Thornton, and Clark (2012)). The FEP offers a unifying account of a plethora of psychological phenomena, such as perceptual categorization, unconscious perceptual inference and perceptual learning (Friston & Kiebel, 2009; Kiebel, Daunizeau, & Friston, 2008; Kiebel, von Kriegstein, Daunizeau, & Friston, 2009), attentional control (Feldman & Friston, 2010), consciousness (Hohwy, 2012), action observation (Friston, Mattout, & Kilner, 2011; Kilner, Friston, & Frith, 2007), motor planning (Adams, Shipp, & Friston, 2012; Friston, 2011; Friston, Samothrakis, & Montague, 2012), value-based decision making (Friston, Adams, & Montague, 2012), emotional valence (Joffily & Coricelli, 2013)—and, amongst others, Freudian ideas (Carhart-Harris & Friston, 2010) and life itself (Friston, 2013). In verbose terms (Bastos et al., 2012), the FEP postulates that biological systems are allostatic, i.e. they act to maintain homeostasis. Formally speaking, this means that they minimize the entropy (dispersion) of their interoceptive and exteroceptive states. Under the FEP, entropy is considered as the average surprise over time, which means that biological systems minimize the surprise associated with their sensory states at each point in time. In the context of the FEP, "surprise" is defined as the negative log model evidence (log marginal likelihood), from which it follows that biological systems continually maximize the Bayesian evidence for their generative model of sensory inputs. Depending on the generative model, minimization of surprise can be computationally intractable and is hence achieved indirectly via the variational free energy bound on the log model evidence as discussed in Section 3. With respect to perception, maximizing the variational free energy then corresponds to Bayesian filtering of sensory inputs, also known as "Predictive Coding" (Rao & Ballard, 1999). With respect to behavior, action must minimize the occurrence of surprising internal and external states. A formal scheme to achieve this under a generative model is "Active Inference" (Brown, Adams, Parees, Edwards, & Friston, 2013). While theoretically appealing, the formal experimental validation of deductions from the mathematical formulation of the FEP remains sparse (for a first example in this direction, see Lieder, Daunizeau, Garrido, Friston, and Stephan (2013)). One of the core motivations for the composition of this tutorial has been to ease the access to the mathematical underpinnings of the FEP, with the aim of deriving experimentally testable predictions based on its formalization and clarify the explanatory requirements that alternative hypotheses to the FEP must achieve.

## 7. Conclusion

In this tutorial, we have sketched a variational framework for Bayesian inference in latent linear stochastic time-series models. We have (1) viewed the discrete time LGSSM, in other words, a latent AR(1)-process, as an approximation of continuous time latent stochastic differential equations, (2) have sketched the VB approach, and (3) have shown how it is practically applied to a numerical example to achieve the two principle aims of the Bayesian paradigm: the quantification of posterior uncertainty about unobserved model variable values and the approximation of the log model evidence. The development and validation of Bayesian techniques for latent random dynamical systems is an active area of research (see e.g. Barber, Cemgil, and Chiappa (2011)). We hope that this tutorial provides an accessible introduction to this flourishing field and demonstrates its wide applicability in the domain of mathematical psychology.

## Acknowledgments

This work was supported by the BMBF Bernstein II initiative (Förderkennzeichen: 01GQ1001C) and the Max Planck Society. We thank three anonymous reviewers for their constructive criticism of our manuscript.

## Appendix A. Non-negativity of the KL-divergence

The non-negativity of the KL-divergence $\mathrm{KL}\left(q\left(x\right)||p(x)\right)$ is central to the VB approach. Here, we follow (Bishop, 2007, p. 55–56) to show why this property holds. Two aspects are central: the negative logarithm is a convex function, and for convex functions Jensen's inequality applies. Convex functions are defined by the property that every straight line connecting two points on the function's graph lies above it, or formally: for $f : [x_1, x_2] \subset \mathbb{R} \rightarrow \mathbb{R}$ and $q \in [0, 1]$

$$f\left(qx_1 + (1-q)\,x_2\right) \leqslant qf\left(x_1\right) + (1-q)f(x_2) \tag{A.1}$$

does hold. Intuitively, (A.1) can be extended to more than two points $x_i$, $i = 1, \ldots, n$ with $q_i \geqslant 0$ and $\sum_{i=1}^{n} q_i = 1$ in the form

$$f\left(\sum_{i=1}^{n} q_i x_i\right) \leqslant \sum_{i=1}^{n} q_i f\left(x_i\right) \tag{A.2}$$

(A.2) can in turn, intuitively, be extended to a continuum of points $x$ and associated values $q(x)$, where $q\left(x\right) \geqslant 0$ and $\int q\left(x\right) dx = 1$ as

$$f\left(\int q\left(x\right) x dx\right) \leqslant \int q(x) f\left(x\right) dx. \tag{A.3}$$

From a probabilistic viewpoint, $\int q\left(x\right) x\, dx$ corresponds to the expectation of $x$ under $q(x)$ and $\int q\left(x\right) f\left(x\right) dx$ corresponds to the expectation of $f\left(x\right)$ under $q(x)$, i.e., for convex $f$ we have

$$\mathbb{E}_{q(x)}\left(f\left(x\right)\right) \geqslant f\left(\mathbb{E}_{q(x)}\left(x\right)\right). \tag{A.4}$$

The results (A.2)–(A.4) are known as Jensen's inequality (e.g. Ash & Doléans-Dade, 2000). Noting from real analysis that the logarithm is a concave function, and $f := -\ln$ hence a convex function, we thus have for the KL-divergence as defined in Eq. (14):

$$\begin{aligned}\mathrm{KL}\left(q\left(x\right)||p(x)\right) &:= \int q\left(x\right) \ln\left(\frac{q\left(x\right)}{p\left(x\right)}\right) dx \\ &= -\int q\left(x\right) \ln\left(\frac{p\left(x\right)}{q\left(x\right)}\right) dx \\ &\geqslant -\ln \int q\left(x\right) \frac{p\left(x\right)}{q\left(x\right)} dx = -\ln 1 = 0.\end{aligned} \tag{A.5}$$

Also note that the KL-divergence $\mathrm{KL}(q\left(x\right)||p(x))$ vanishes for $p\left(x\right) := q(x)$, as in this case the logarithmic term in the integral evaluates to zero for all $x$. For further discussion of properties of the KL-divergence, see for example Cover and Thomas (1991).

## Appendix B. Marginal likelihood decomposition

By definition of the variational free energy in (13), we have

$$\mathrm{F}\left(q\left(\vartheta\right)\right) = \int q\left(\vartheta\right) \ln\left(\frac{p\left(y, \vartheta\right)}{q\left(\vartheta\right)}\right) d\vartheta. \tag{B.1}$$

Using the definition of conditional probability, we have

$$\mathrm{F}\left(q\left(\vartheta\right)\right) = \int q\left(\vartheta\right) \ln\left(p(y)\frac{p\left(\vartheta|y\right)}{q\left(\vartheta\right)}\right) d\vartheta. \tag{B.2}$$

Using the properties of the logarithm and the linearity of integrals, from (B.2) it follows that

$$\mathrm{F}\left(q\left(\vartheta\right)\right) = \int q\left(\vartheta\right) \ln p(y) d\vartheta + \int q\left(\vartheta\right) \ln\left(\frac{p\left(\vartheta|y\right)}{q\left(\vartheta\right)}\right) d\vartheta. \tag{B.3}$$

With the linearity of integrals, we then also have

$$\mathrm{F}\left(q\left(\vartheta\right)\right) = \ln p(y) \int q\left(\vartheta\right) d\vartheta + \int q\left(\vartheta\right) \ln\left(\frac{p\left(\vartheta|y\right)}{q\left(\vartheta\right)}\right) d\vartheta \tag{B.4}$$

and because $q\left(\vartheta\right)$ is a probability distribution (and thus integrates to 1) and again with the properties of the logarithm, we obtain

$$\mathrm{F}\left(q\left(\vartheta\right)\right) = \ln p(y) - \int q\left(\vartheta\right) \ln\left(\frac{q\left(\vartheta\right)}{p\left(\vartheta|y\right)}\right) d\vartheta. \tag{B.5}$$

The definition of the KL-divergence in (14) then allows to write (B.5) as

$$\mathrm{F}\left(q\left(\vartheta\right)\right) = \ln p(y) - KL\left(q\left(\vartheta\right)||p\left(\vartheta|y\right)\right) \tag{B.6}$$

which is equivalent to (12).

## References

Adams, R. A., Shipp, S., & Friston, K. J. (2012). Predictions not commands: active inference in the motor system. *Brain Structure & Function*, http://dx.doi.org/10.1007/s00429-012-0475-5.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. http://dx.doi.org/10.1109/TAC.1974.1100705.

Archambeau, C., Cornford, D., Opper, M., & Shawe-Taylor, J. (2007). Gaussian process approximations of stochastic differential equations. Available from: http://core.kmi.open.ac.uk/display/21067.

Archambeau, C., Shen, Y., Cornford, D., & Shawe-Taylor, J. (2008). Variational inference for diffusion processes. In C. Platt, D. Koller, Y. Singer, & S. Roweis (Eds.), *Neural information processing systems (NIPS), Vol. 20* (pp. 17–24). Cambridge, US: The MIT Press.

Arnold, L. (1998). *Random dynamical systems*. Berlin; New York: Springer.

Ash, R. B., & Doléans-Dade, C. (2000). *Probability and measure theory*. San Diego: Harcourt/Academic Press.

Barber, David (2012). *Bayesian reasoning and machine learning*. Cambridge University Press.

Barber, David, Cemgil, A. T., & Chiappa, S. (Eds.) (2011). *Bayesian time series models*. Cambridge: University Press.

Barber, D., & Chiappa, S. (2007). Unified inference for variational Bayesian linear Gaussian state-space model. In B. Schölkopf, P. Platt, & T. Hofmann (Eds.), *Advances in neural information processing systems 19: Proceedings of the 2006 conference* (pp. 81–88). Cambridge, US: The MIT Press.

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., & Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron*, *76*(4), 695–711. http://dx.doi.org/10.1016/j.neuron.2012.10.038.

Beal, Matthew J. (2003). *Variational algorithms for approximate Bayesian inference*. (Ph.D. thesis), London: Gatsby Computational Neuroscience Unit, University College London, Available from http://www.cse.buffalo.edu/faculty/mbeal/thesis/.

Bishop, C. M. (2007). *Pattern recognition and machine learning*. New York: Springer, 1st ed. 2006. Corr. 2nd printing.

Briers, M., Doucet, A., & Maskell, S. (2004). Smoothing algorithms for state-space models. Available from http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.109.5006.

Brown, H., Adams, R. A., Parees, I., Edwards, M., & Friston, K. (2013). Active inference, sensory attenuation and illusions. *Cognitive Processing*, http://dx.doi.org/10.1007/s10339-013-0571-3.

Brown, S. D., Ratcliff, R., & Smith, P. L. (2006). Evaluating methods for approximating stochastic differential equations. *Journal of Mathematical Psychology*, *50*(4), 402–410. http://dx.doi.org/10.1016/j.jmp.2006.03.004.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference understanding AIC and BIC in model selection. *Sociological Methods & Research*, *33*(2), 261–304. http://dx.doi.org/10.1177/0049124104268644.

Busemeyer, Jerome R., Jessup, R. K., Johnson, J. G., & Townsend, J. T. (2006). Building bridges between neural models and complex decision making behaviour. *Neural Networks: The Official Journal of the International Neural Network Society*, *19*(8), 1047–1058. http://dx.doi.org/10.1016/j.neunet.2006.05.043.

Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: a dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review*, *100*(3), 432–459.

Carhart-Harris, R. L., & Friston, K. J. (2010). The default-mode, ego-functions and free-energy: a neurobiological account of Freudian ideas. *Brain*, *133*(4), 1265–1283. http://dx.doi.org/10.1093/brain/awq010.

Casella, G., & Berger, R. L. (2002). *Statistical inference*. Australia: Pacific Grove, CA: Thomson Learning.

Chappell, Michael, Groves, Adrian, & Woolrich, Mark (2008). TR08MC1: the FMRIB variational bayes tutorial, Oxford. Available from http://users.fmrib.ox.ac.uk/~chappell/papers/TR07MC1.pdf.

Chen, C. C., Kiebel, S. J., & Friston, K. J. (2008). Dynamic causal modelling of induced responses. *NeuroImage*, *41*(4), 1293–1312. http://dx.doi.org/10.1016/j.neuroimage.2008.03.026.

Chen, C.-C., Kiebel, S. J., Kilner, J. M., Ward, N. S., Stephan, K. E., Wang, W.-J., et al. (2012). A dynamic causal model for evoked and induced responses. *NeuroImage*, *59*(1), 340–348. http://dx.doi.org/10.1016/j.neuroimage.2011.07.066.

Cover, T. M., & Thomas, (1991). *Elements of information theory*. New York: Wiley.

Daunizeau, J., David, O., & Stephan, K. E. (2011). Dynamic causal modelling: a critical review of the biophysical and statistical foundations. *NeuroImage*, *58*(2), 312–322. http://dx.doi.org/10.1016/j.neuroimage.2009.11.062.

Daunizeau, Jean, den Ouden, H. E. M., Pessiglione, M., Kiebel, S. J., Friston, K. J., & Stephan, K. E. (2010). Observing the observer (II): deciding when to decide. *PloS One*, *5*(12), e15555. http://dx.doi.org/10.1371/journal.pone.0015555.

Daunizeau, Jean, den Ouden, H. E. M., Pessiglione, M., Kiebel, S. J., Stephan, K. E., & Friston, K. J. (2010). Observing the observer (I): meta-bayesian models of learning and decision-making. *PloS One*, *5*(12), e15554. http://dx.doi.org/10.1371/journal.pone.0015554.

Daunizeau, J., Friston, K. J., & Kiebel, S. J. (2009). Variational Bayesian identification and prediction of stochastic nonlinear dynamic causal models. *Physica D. Nonlinear Phenomena*, *238*(21), 2089–2118. http://dx.doi.org/10.1016/j.physd.2009.08.002.

Daunizeau, J., Stephan, K. E., & Friston, K. J. (2012). Stochastic dynamic causal modelling of fMRI data: should we care about neural noise?. *NeuroImage*, *62*(1), 464–481. http://dx.doi.org/10.1016/j.neuroimage.2012.04.061.

David, O., Harrison, L., & Friston, K. J. (2005). Modelling event-related responses in the brain. *NeuroImage*, *25*(3), 756–770. http://dx.doi.org/10.1016/j.neuroimage.2004.12.030.

David, O., Kiebel, S. J., Harrison, L. M., Mattout, J., Kilner, J. M., & Friston, K. J. (2006). Dynamic causal modeling of evoked responses in EEG and MEG. *NeuroImage*, *30*(4), 1255–1272. http://dx.doi.org/10.1016/j.neuroimage.2005.10.045.

Dayan, P., & Abbott, L. F. (2005). *Theoretical neuroscience: computational and mathematical modeling of neural systems* (1). Cambridge, US: The MIT Press.

DeGroot, M. H. (2004). *Optimal statistical decisions*. Hoboken, N.J.: Wiley-Interscience.

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *39*(1), 1–38. http://dx.doi.org/10.2307/2984875.

Domingos, P. (1999). The role of Occam's Razor in knowledge discovery. *Data Mining and Knowledge Discovery*, *3*, 409–425.

Efron, B. (2013). Bayes' theorem in the 21st century. *Science*, *340*(6137), 1177–1178. http://dx.doi.org/10.1126/science.1236536.

Feldman, H., & Friston, K. J. (2010). Attention, uncertainty, and free-energy. *Frontiers in Human Neuroscience*, *4*, 215. http://dx.doi.org/10.3389/fnhum.2010.00215.

Friston, K. J. (2007). *Statistical parametric mapping - The analysis of funtional brain images*. Amsterdam: Academic Press, Elsevier.

Friston, K. (2008a). Hierarchical models in the brain. *PLoS Computational Biology*, *4*(11), e1000211. http://dx.doi.org/10.1371/journal.pcbi.1000211.

Friston, K. J. (2008b). Variational filtering. *NeuroImage*, *41*(3), 747–766. http://dx.doi.org/10.1016/j.neuroimage.2008.03.017.

Friston, K. (2010). The free-energy principle: a unified brain theory?. *Nature Reviews Neuroscience*, *11*(2), 127–138. http://dx.doi.org/10.1038/nrn2787.

Friston, K. (2009). The free-energy principle: a rough guide to the brain?. *Trends in Cognitive Sciences*, *13*(7), 293–301. http://dx.doi.org/10.1016/j.tics.2009.04.005.

Friston, K. (2011). What is optimal about motor control? *Neuron*, *72*(3), 488–498. http://dx.doi.org/10.1016/j.neuron.2011.10.018.

Friston, K. (2013). Life as we know it. *Journal of The Royal Society Interface*, *10*(86), http://dx.doi.org/10.1098/rsif.2013.0475.

Friston, K., Adams, R., & Montague, R. (2012). What is value-accumulated reward or evidence?. *Frontiers in Neurorobotics*, *6*, 11. http://dx.doi.org/10.3389/fnbot.2012.00011.

Friston, K. J., Harrison, L., & Penny, W. (2003). Dynamic causal modelling. *NeuroImage*, *19*(4), 1273–1302.

Friston, K., & Kiebel, S. (2009). Cortical circuits for perceptual inference. *Neural Networks: The Official Journal of the International Neural Network Society*, *22*(8), 1093–1104. http://dx.doi.org/10.1016/j.neunet.2009.07.023.

Friston, K., Mattout, J., & Kilner, J. (2011). Action understanding and active inference. *Biological Cybernetics*, *104*(1–2), 137–160. http://dx.doi.org/10.1007/s00422-011-0424-z.

Friston, K., Mattout, J., Trujillo-Barreto, N., Ashburner, J., & Penny, W. (2007). Variational free energy and the Laplace approximation. *NeuroImage*, *34*(1), 220–234. http://dx.doi.org/10.1016/j.neuroimage.2006.08.035.

Friston, K. J., Penny, W., Phillips, C., Kiebel, S., Hinton, G., & Ashburner, J. (2002). Classical and Bayesian inference in neuroimaging: theory. *NeuroImage*, *16*(2), 465–483. http://dx.doi.org/10.1006/nimg.2002.1090.

Friston, K., Samothrakis, S., & Montague, R. (2012). Active inference and agency: optimal control without cost functions. *Biological Cybernetics*, *106*(8–9), 523–541. http://dx.doi.org/10.1007/s00422-012-0512-8.

Friston, K., Stephan, K., Li, B., & Daunizeau, J. (2010). Generalised filtering. *Mathematical Problems in Engineering*, *2010*, http://dx.doi.org/10.1155/2010/621670.

Friston, K., Thornton, C., & Clark, A. (2012). Free-energy minimization and the dark-room problem. *Frontiers in Psychology*, *3*, 130. http://dx.doi.org/10.3389/fpsyg.2012.00130.

Friston, K. J., Trujillo-Barreto, N., & Daunizeau, J. (2008). DEM: a variational treatment of dynamic systems. *NeuroImage*, *41*(3), 849–885. http://dx.doi.org/10.1016/j.neuroimage.2008.02.054.

Fujimoto, K., Satoh, A., & Fukunaga, S. (2011). System identification based on variational Bayes method and the invariance under coordinate transformations. CDC-ECE (S. 3882–3888). Available from http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6160563.

Havlicek, M., Friston, K. J., Jan, J., Brazdil, M., & Calhoun, V. D. (2011). Dynamic modeling of neuronal responses in fMRI using cubature Kalman filtering. *NeuroImage*, *56*(4), 2109–2128. http://dx.doi.org/10.1016/j.neuroimage.2011.03.005.

Hays, W. L. (1994). *Statistics*. Fort Worth: Harcourt Brace College Publishers.

Herzog, S., & Ostwald, D. (2013). Experimental biology: sometimes Bayesian statistics are better. *Nature*, *494*(7435), 35. http://dx.doi.org/10.1038/494035b.

Hinton, G. E., & van Camp, D. (1993). Keeping the neural networks simple by minimizing the description length of the weights. In *COLT'93, Proceedings of the sixth annual conference on Computational learning theory*. New York, NY, USA: ACM, http://dx.doi.org/10.1145/168304.168306. (S. 5–13).

Hohwy, J. (2012). Attention and conscious perception in the hypothesis testing brain. *Frontiers in Consciousness Research*, *3*, 96. http://dx.doi.org/10.3389/fpsyg.2012.00096.

Honerkamp, J. (1993). *Stochastic dynamical systems: concepts, numerical methods, data analysis*. New York, NY: Wiley.

Huang, K., Sen, S., & Szidarovszky, F. (2012). Connections among decision field theory models of cognition. *Journal of Mathematical Psychology*, *56*(5), 287–296. http://dx.doi.org/10.1016/j.jmp.2012.07.005.

Hunt, L. T., Kolling, N., Soltani, A., Woolrich, M. W., Rushworth, M. F. S., & Behrens, T. E. J. (2012). Mechanisms underlying cortical activity during value-guided choice. *Nature Neuroscience*, *15*(3), 470–476. http://dx.doi.org/10.1038/nn.3017. S1–3.

Hurvich, C. M., & Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, *76*(2), 297–307. http://dx.doi.org/10.1093/biomet/76.2.297.

Joffily, M., & Coricelli, G. (2013). Emotional valence and the free-energy principle. *PLoS Computational Biology*, *9*(6), e1003094. http://dx.doi.org/10.1371/journal.pcbi.1003094.

Jordan, M. I. (1999). *Learning in graphical models*. Cambridge, Mass: MIT Press.

Kalman, R. (1960). A new approach to linear filtering and prediction problems. *Transactions of the ASME – Journal of Basic Engineering*, *82*(Series D), 35–45.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795. http://dx.doi.org/10.2307/2291091.

Kiebel, S. J., Daunizeau, J., & Friston, K. J. (2008). A hierarchy of time-scales and the brain. *PLoS Computational Biology*, *4*(11), e1000209. http://dx.doi.org/10.1371/journal.pcbi.1000209.

Kiebel, S. J., von Kriegstein, K., Daunizeau, J., & Friston, K. J. (2009). Recognizing sequences of sequences. *PLoS Computational Biology*, *5*(8), e1000464. http://dx.doi.org/10.1371/journal.pcbi.1000464.

Kilner, J. M., Friston, K. J., & Frith, C. D. (2007). Predictive coding: an account of the mirror neuron system. *Cognitive Processing*, *8*(3), 159–166. http://dx.doi.org/10.1007/s10339-007-0170-2.

Kloeden, P. E., & Platen, E. (1999). *Numerical solution of stochastic differential equations*. New York: Springer, 1st ed. 1992. Corr. 4th printing.

Øksendal, B. K. (2003). *Stochastic differential equations: an introduction with applications*. Berlin; New York: Springer.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86. http://dx.doi.org/10.2307/2236703.

Lee, M. D. (2008). Three case studies in the Bayesian analysis of cognitive models. *Psychonomic Bulletin & Review*, *15*(1), 1–15.

Lehmann, E., & Casella, G. (1998). *Theory of point estimation (Springer texts in statistics)*. New York: Springer.

Li, B., Daunizeau, J., Stephan, K. E., Penny, W., Hu, D., & Friston, K. (2011). Generalised filtering and stochastic DCM for fMRI. *NeuroImage*, *58*(2), 442–457. http://dx.doi.org/10.1016/j.neuroimage.2011.01.085.

Lieder, F., Daunizeau, J., Garrido, M. I., Friston, K. J., & Stephan, K. E. (2013). Modelling trial-by-trial changes in the mismatch negativity. *PLoS Computational Biology*, *9*(2), e1002911. http://dx.doi.org/10.1371/journal.pcbi.1002911.

Lodewyckx, T., Kim, W., Lee, M., Tuerlinckx, F., Kuppens, P., & Wagenmakers, E.-J. (2011). A tutorial on Bayes factor estimation with the product space method. *Journal of Mathematical Psychology*, *55*(5), 331–347. http://dx.doi.org/10.1016/j.jmp.2011.06.001.

Logothetis, N. K. (2008). What we can do and what we cannot do with fMRI. *Nature*, *453*(7197), 869–878. http://dx.doi.org/10.1038/nature06976.

Lunn, D. (2013). *The BUGS book: a practical introduction to Bayesian analysis*. Boca-Raton, FL: Chapman & Hall Texts in Statistical Science Series.

MacKay, D. J. C. (1995). Free energy minimisation algorithm for decoding and cryptanalysis. *Electronics Letters*, *31*(6), 446–447. http://dx.doi.org/10.1049/el:19950331.

MacKay, David J. C. (2003). *Information theory, inference and learning algorithms*. Cambridge University Press.

Mathys, C., Daunizeau, J., Friston, K. J., & Stephan, K. E. (2011). A bayesian foundation for individual learning under uncertainty. *Frontiers in Human Neuroscience*, *5*, 39. http://dx.doi.org/10.3389/fnhum.2011.00039.

McLachlan, G. J., & Krishnan, (2008). *The EM algorithm and extensions*. Hoboken, N.J.: Wiley-Interscience.

Mil'shtein, G. N. (2011). *Numerical integration of stochastic differential equations*. Dordrecht; London: Springer.

Molenaar, P. C. M., & Newell, K. M. (2010). *Individual pathways of change—statistical models for analyzing learning and development*. American Psychological Association.

Moran, R., Pinotsis, D. A., & Friston, K. (2013). Neural masses and fields in dynamic causal modeling. *Frontiers in Computational Neuroscience*, *7*, 57. http://dx.doi.org/10.3389/fncom.2013.00057.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. Cambridge, MA: MIT Press.

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, *47*(1), 90–100. http://dx.doi.org/10.1016/S0022-2496(02)00028-7.

Neal, R., & Hinton, G. E. (1998). A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models* (pp. 355–368). Kluwer: Academic Publishers.

O'Connell, R. G., Dockree, P. M., & Kelly, S. P. (2012). A supramodal accumulation-to-bound signal that determines perceptual decisions in humans. *Nature Neuroscience*, *15*(12), 1729–1735. http://dx.doi.org/10.1038/nn.3248.

Oravecz, Z., Tuerlinckx, F., & Vandekerckhove, J. (2011). A hierarchical latent stochastic differential equation model for affective dynamics. *Psychological Methods*, *16*(4), 468–490. http://dx.doi.org/10.1037/a0024375.

Ozaki, T. (1992). A bridge between nonlinear time series models and nonlinear stochastic dynamical systems: a local linearization approach. *Statistica Sinica*, *2*(1), 113–135.

Penny, W. D. (2012). Comparing dynamic causal models using AIC, BIC and free energy. *NeuroImage*, *59*(1), 319–330. http://dx.doi.org/10.1016/j.neuroimage.2011.07.039.

Penny, S.J.R. (2000). Variational bayes for 1-dimensional mixture models. Available from www.robots.ox.ac.uk/~sjrob/Pubs/vbmog.ps.gz.

Penny, W., Kiebel, S., & Friston, K. (2003). Variational Bayesian inference for fMRI time series. *NeuroImage*, *19*(3), 727–741.

Penny, W. D., Stephan, K. E., Mechelli, A., & Friston, K. J. (2004). Comparing dynamic causal models. *NeuroImage*, *22*(3), 1157–1172. http://dx.doi.org/10.1016/j.neuroimage.2004.03.026.

Pernet, C. R., Sajda, P., & Rousselet, G. A. (2011). Single-trial analyses: why bother?. *Frontiers in Perception Science*, *322*, http://dx.doi.org/10.3389/fpsyg.2011.00322.

Philiastides, M. G., Auksztulewicz, R., Heekeren, H. R., & Blankenburg, F. (2011). Causal role of dorsolateral prefrontal cortex in human perceptual decision making. *Current Biology: CB*, *21*(11), 980–983. http://dx.doi.org/10.1016/j.cub.2011.04.034.

Philiastides, M. G., Ratcliff, R., & Sajda, P. (2006). Neural representation of task difficulty and decision making during perceptual categorization: a timing diagram. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *26*(35), 8965–8975. http://dx.doi.org/10.1523/JNEUROSCI.1655-06.2006.

Pinotsis, D. A., Moran, R. J., & Friston, K. J. (2012). Dynamic causal modeling with neural fields. *NeuroImage*, *59*(2), 1261–1274. http://dx.doi.org/10.1016/j.neuroimage.2011.08.020.

Port, R. F., & Van Gelder, T. (1995). *Mind as motion explorations in the dynamics of cognition*. Cambridge (Mass.); London: Bradford Book.

Press, W. H. (2007). *Numerical recipes: the art of scientific computing*. Cambridge, UK; New York: Cambridge University Press.

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257–286.

Raichle, M. E. (2009). A paradigm shift in functional brain imaging. *The Journal of Neuroscience*, *29*(41), 12729–12734. http://dx.doi.org/10.1523/JNEUROSCI.4366-09.2009.

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87. http://dx.doi.org/10.1038/4580.

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. Cambridge, Mass.: MIT Press.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108. http://dx.doi.org/10.1037/0033-295X.85.2.59.

Ratcliff, R., Philiastides, M. G., & Sajda, P. (2009). Quality of evidence for perceptual decision making is indexed by trial-to-trial variability of the EEG. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(16), 6539–6544. http://dx.doi.org/10.1073/pnas.0812589106.

Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin & Review*, *9*(3), 438–481.

Ratcliff, R., & Van Dongen, H. P. A. (2011). Diffusion model for one-choice reaction-time tasks and the cognitive effects of sleep deprivation. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(27), 11285–11290. http://dx.doi.org/10.1073/pnas.1100483108.

Rauch, H., Striebel, C., & Tung, F. (1965). Maximum likelihood estimates of linear dynamic systems. *Journal of the American Institute of Aeronautics and Astronautics*, *3*(8), 1445–1450.

Robert, C. P. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation* (2nd ed.). New York: Springer.

Roberts, S. J., & Penny, W. D. (2002). Variational Bayes for generalized autoregressive models. *IEEE Transactions on Signal Processing*, *50*(9), 2245–2257. http://dx.doi.org/10.1109/TSP.2002.801921.

Roweis, S., & Ghahramani, Z. (1999). A unifying review of linear Gaussian models. Available from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.30.5555.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464. http://dx.doi.org/10.2307/2958889.

Shannon, C. E. (1948). *The Bell System Technical Journal*, *27*, 379–423, 623–656.

Shen, Y., Archambeau, C., Cornford, D., Opper, M., Shawe-Taylor, J., & Barillec, R. (2010). A comparison of variational and Markov chain monte carlo methods for inference in partially observed stochastic dynamic systems. *Journal of Signal Processing Systems*, *61*(1), 51–59. http://dx.doi.org/10.1007/s11265-008-0299-y.

Shiffrin, R. M., Lee, M. D., Kim, W., & Wagenmakers, E.-J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical bayesian methods. *Cognitive Science*, *32*(8), 1248–1284. http://dx.doi.org/10.1080/03640210802414826.

Shumway, D. R. H., & Stoffer, P. D. S. (2011). Time series analysis and its applications. In *Springer texts in statistics*. New York: Springer.

Smith, P. L. (2000). Stochastic dynamic models of response time and accuracy: a foundational primer. *Journal of Mathematical Psychology*, *44*(3), 408–463. http://dx.doi.org/10.1006/jmps.1999.1260.

Smith, P. L., & Ratcliff, R. (2004). Psychology and neurobiology of simple decisions. *Trends in Neurosciences*, *27*(3), 161–168. http://dx.doi.org/10.1016/j.tins.2004.01.006.

Stephan, K. E., & Friston, K. J. (2010). Analyzing effective connectivity with fMRI. *Wiley Interdisciplinary Reviews. Cognitive Science*, *1*(3), 446–459. http://dx.doi.org/10.1002/wcs.58.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: an introduction*. The MIT Press.

Tuckerman, M. E. (2010). *Statistical mechanics: theory and molecular simulation*. Oxford University Press.

Tzikas, D. G., Likas, C. L., & Galatsanos, N. P. (2008). The variational approximation for Bayesian inference. *IEEE Signal Processing Magazine*, *25*(6), 131–146. http://dx.doi.org/10.1109/MSP.2008.929620.

Van Brunt, B. (2004). *The calculus of variations*. New York: Springer.

Vandekerckhove, J., & Tuerlinckx, F. (2007). Fitting the Ratcliff diffusion model to experimental data. *Psychonomic Bulletin & Review*, *14*(6), 1011–1026.

Visser, I. (2011). Seven things to remember about hidden Markov models: a tutorial on Markovian models for time series. *Journal of Mathematical Psychology*, *55*(6), 403–415. http://dx.doi.org/10.1016/j.jmp.2011.08.002.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*(5), 779–804.

Wagenmakers, E.-J., van der Maas, H. L. J., & Grasman, R. P. P. P. (2007). An EZ-diffusion model for response time and accuracy. *Psychonomic Bulletin & Review*, *14*(1), 3–22.

Ward, L. M. (2002). *Dynamical cognitive science*. Cambridge, Mass: The MIT Press.

Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, *44*(1), 92–107. http://dx.doi.org/10.1006/jmps.1999.1278.

Winn, J., & Bishop, C. (2005). Variational message passing. *Journal of Machine Learning Research*, *6*, 661–694.

Wipf, D., & Nagarajan, S. (2009). A unified Bayesian framework for MEG/EEG source imaging. *NeuroImage*, *44*(3), 947–966. http://dx.doi.org/10.1016/j.neuroimage.2008.02.059.

Woolrich, M. W., Ripley, B. D., Brady, M., & Smith, S. M. (2001). Temporal autocorrelation in univariate linear modeling of FMRI data. *NeuroImage*, *14*(6), 1370–1386. http://dx.doi.org/10.1006/nimg.2001.0931.

Worsley, K. J., & Friston, K. J. (1995). Analysis of fMRI time-series revisited–again. *NeuroImage*, *2*(3), 173–181. http://dx.doi.org/10.1006/nimg.1995.1023.

Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *The Annals of Statistics*, *11*(1), 95–103. http://dx.doi.org/10.2307/2240463.

Zhang, J., Bogacz, R., & Holmes, P. (2009). A comparison of bounded diffusion models for choice in time controlled tasks. *Journal of Mathematical Psychology*, *53*(4), 231–241. http://dx.doi.org/10.1016/j.jmp.2009.03.001.