Entropy **2014**, *16*, 3832-3847; doi:10.3390/e16073832

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.com/journal/entropy

Article

Variational Bayes for Regime-Switching Log-Normal Models

Hui Zhao and Paul Marriott *

University of Waterloo, 200 University Avenue West, Waterloo, ON N2L 3G1, Canada;
E-Mail: h6zhao@uwaterloo.ca

* Author to whom correspondence should be addressed; E-Mail: pmarriot@uwaterloo.ca;
Tel.: +1-519-888-4567.

*Received: 14 April 2014; in revised form: 12 June 2014 / Accepted: 7 July 2014 /
Published: 14 July 2014*

Abstract: The power of projection using divergence functions is a major theme in information geometry. One version of this is the variational Bayes (VB) method. This paper looks at VB in the context of other projection-based methods in information geometry. It also describes how to apply VB to the regime-switching log-normal model and how it provides a computationally fast solution to quantify the uncertainty in the model specification. The results show that the method can recover exactly the model structure, gives the reasonable point estimates and is very computationally efficient. The potential problems of the method in quantifying the parameter uncertainty are discussed.

Keywords: information geometry; variational Bayes; regime-switching log-normal model; model selection; covariance estimation

1. Introduction

While, in principle, the calculation of the posterior distribution is mathematically straightforward, in practice, the computation of many of its features, such as posterior densities, normalizing constants and posterior moments, is a major challenge in Bayesian analysis. Such computations typically involve high dimensional integrals, which often have no analytical or tractable forms. The variational Bayes (VB) method was developed to generate tractable approximations to these quantities. This method provides analytic approximations to the posterior distribution by minimizing the Kullback–Leibler (KL) divergence from the approximations to the actual posterior and has been demonstrated to be computationally very fast.

VB gains its computational advantages by making simplifying assumptions about the posterior dependence structure. For example, in the simplest form, it assumes posterior independence between selected sets of parameters. Under these assumptions, the resultant approximate posterior is either known analytically or can be computed by a simple iteration algorithm similar to the Expectation-Maximization (EM) algorithm. In this paper, we show that, as well as having advantages of computational speed, the VB algorithm does an excellent job of model selection, in particular in finding the appropriate number of regimes.

While the simplification in the dependence gives computational advantages, it also comes at a cost. For example, we also found that the posterior variance may be underestimated. In [1], we propose a novel method to compute the true posterior covariance matrix by only using the information obtained from VB approximations.

The use of projections to particular families is, of course, not new to information geometry (IG). In [2], we find the famous Pythagorean results concerning projection using α-divergences to α-families, and other important results on projections based on divergences can be found in [3] and [4] (Chapter 7).

1.1. Variational Bayes

Suppose, in a Bayesian inference problem that we use $q(\boldsymbol{\tau})$ to approximate the posterior $p(\boldsymbol{\tau}|y)$, where y is the data and $\boldsymbol{\tau} = \{\tau_1, \cdots, \tau_p\}$ the model parameter vector. The KL divergence between them is defined as,

$$\text{KL}\left[q(\boldsymbol{\tau})||p(\boldsymbol{\tau}|\mathbf{y})\right] = \int q(\boldsymbol{\tau}) \log \frac{q(\boldsymbol{\tau})}{p(\boldsymbol{\tau}|\mathbf{y})} d\boldsymbol{\tau}, \tag{1}$$

provided the integral exists. We want to balance two things, having the discrepancy between p and q small, while keeping q tractable. Hence, we want to seek $q(\boldsymbol{\tau})$, which minimizes Equation (1), while keeping $q(\tau)$ in an analytically tractable form. First, note that the evaluation of Equation (1) requires $p(\boldsymbol{\tau}|\mathbf{y})$, which may be unavailable, since in the general Bayesian problem, its normalizing constant is one of the main intractable integrals. However, we note that:

$$\begin{aligned}\text{KL}\left[q(\boldsymbol{\tau})||p(\boldsymbol{\tau}|\mathbf{y})\right] &= \int q(\boldsymbol{\tau}) \log \frac{q(\boldsymbol{\tau})}{p(\boldsymbol{\tau}|\mathbf{y})p(\mathbf{y})} d\boldsymbol{\tau} + \log p(\mathbf{y}) \\ &= -\int q(\boldsymbol{\tau}) \log \frac{p(\boldsymbol{\tau},\mathbf{y})}{q(\boldsymbol{\tau})} d\boldsymbol{\tau} + \log p(\mathbf{y}).\end{aligned} \tag{2}$$

Thus, minimizing Equation (1) is equivalent to maximizing the first term of the right-hand side of Equation (2). The key computational point is that, often, the term $p(\boldsymbol{\tau}, \mathbf{y})$ is available even when the full posterior $\frac{p(\tau,y)}{\int p(\tau,y)d\tau}$ is not.

Definition 1. *Let* $F(q) = \int q(\boldsymbol{\tau}) \log \frac{p(\boldsymbol{\tau},\mathbf{y})}{q(\boldsymbol{\tau})} d\boldsymbol{\tau}$ *and:*

$$\hat{q} = \underset{q \in Q}{\arg\max}\, F(q), \tag{3}$$

where Q is a predetermined set of probability density functions over the parameter space. Then $\hat{q}$ is called the variational approximation or variational posterior distribution, and functions of $\hat{q}$ (such as mean, variance, etc.), are called variational parameters.

Some of the power of Definition 1 comes when we assume that all elements of Q have tractable posteriors. In that case, all variational parameters will then also be tractable when the optimization can be achieved. A prime example of a choice for Q is the set of all densities that factorize as

$$q(\boldsymbol{\tau}) = \prod_{i=1}^{d} q_i(\tau_i).$$

This reduces the computational problem from computing a high dimensional integral to one of computing a number of one-dimensional ones. Furthermore, as we see in the example of this paper, it is often the case that the variational families are standard exponential families (since they are often 'maximum entropy models' in some sense), and the optimisation problem (3) can be solved by simple iterative methods with very fast convergence.

The core of the method builds on the basis of the principle of the variational free energy minimization in physics, which is concerned with finding the maxima and minima of a functional over a class of functions, and the method gains its name from this root. Early developments of the method can be found in machine learning, especially in applications on neural networks [5,6]. The method has been successfully applied in many different disciplines and domains, for example, in independent component analysis [7,8], graphical models [9,10], information retrieval [11] and factor analysis [12].

In the statistical literature, an early application of the variational principle can be found in the work of [13] to construct Bayes estimators. In recent years, the method has obtained more attention from both the application and theoretical perspective, for example [14–18].

1.2. Regime-Switching Models

In this paper, we illustrate the strengths and weaknesses of VB through a detailed case study. In particular, we look at a model that is used in finance, risk management and actuarial science, the so-called regime-switching log-normal model (RSLN) proposed, in this context, by [19].

Switching between different states, or regimes, is a common phenomenon in many time series, and regime-switching models, originally proposed by [20], have been used to model these switching processes. As demonstrated in [21], the maximum likelihood estimate (MLE) does not give a simple method to deal with parameter uncertainty; for details of this method, see [21]. The asymptotic normality of maximum likelihood estimators may not apply for sample sizes commonly found in practice. Hence, to understand parameter uncertainty, [21] considered the RSLN model in a Bayesian framework using the Metropolis–Hastings algorithm. Furthermore, model uncertainty, in particular selecting the correct number of regimes, is a major issue. Hence, model selection criteria have to be used to choose the "best" model. Hardy [19] found that a two-regime RSLN model maximized the Bayes information criterion (BIC) [22] for both monthly TSE 300 total return data and S&P 500 total return data; however, according to the Akaike information criterion (AIC) [23], a three-regime model was the optimal on S&P data. To account for the model uncertainty associated with the number of regimes, [24] offered a trans-dimensional model using reversible jump MCMC [25]. We note that BIC is not necessarily ideal for model selection with state space models [26], while it is still commonly used in the literature.

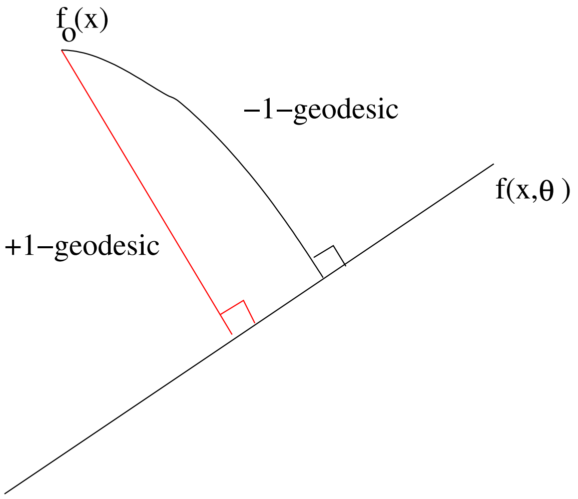
MCMC methods make possible the computation of all posterior quantities; however there are a number of practical issues associated with their implementation. A primary concern is determining that the generated chain has, in fact, "converged". In practice, MCMC practitioners have to resort to convergence diagnostic techniques. Furthermore, the computational cost can be a concern. Other implementational issues include the difficulty of making good initalisation choices, implementing the MCMC algorithm in one long chain or several shorter chains in parallel, *etc*. Detailed discussions can be found in [27].

One of the main contributions of this paper is to apply the variational Bayes (VB) method to the RSLN model and present a solution to quantify the uncertainty in model specification. The VB method is a technique that provides analytical approximations to the posterior quantities, and in practice, it is demonstrated to be a very much faster alternative to MCMC methods.

2. Variational Bayes and Informational Geometry

In this section, we explore the relationship between VB and IG, in particular the statistical properties of divergence-based projections onto exponential families. Here, we used the IG of [2], in particular the ± 1 dual affine parameters for exponential families. One of the most striking results from [2] is the Pythagorean property of these dual affine coordinate systems. This is illustrated in Figure 1, which shows a schematic representing a model space containing the distribution $f_0(x)$ and an exponential family $f(x;\theta)$.

Figure 1. Projections onto an exponential family.



The Pythagorean result comes from using the KL divergence to project onto the exponential family $f(x;\theta) = \nu(x)\exp\{s(x)\theta - \psi(\theta)\}$, *i.e.*,

$$\min_{\theta} \int -\log \frac{f(x;\theta)}{f_0(x)} f_0(x)dx.$$

All distributions that project to the same point form a -1-flat space defined by all distributions $f(x)$ with the same mean, *i.e.*,

$$E_{\widehat{\theta}}(s(x)) = E_{f(x)}(s(x)),$$

and further, it is Fisher orthogonal to the $+1$-flat family $f(x;\theta)$. The statistical interpretation of this concerns the behaviour of a model $f(x,\theta)$ when the data generation process does not lie in the model.

In contrast to this, we have the VB method, which uses the reverse KL divergence for the projection, *i.e.*,

$$\min_{\theta} \int \log \frac{f(x;\theta)}{f_0(x)} f(x;\theta) dx.$$

This results in a Fisher orthogonal projection, shown in Figure 1, but now using a $+1$-flat family. This does not have the property that the mean of $s(x)$ is constant, but as we shall see, it does have nice computational properties when used in the context of Bayesian analysis.

In order to investigate the information geometry of VB, we consider two examples. The first, in Section 3.1, is selected to maximally illustrate the underlying geometric issues and to get some understanding of the quality of the VB approximation. The second, in Section 3.2, shows an important real-world application from actuarial science and is illustrated with simulated and real data.

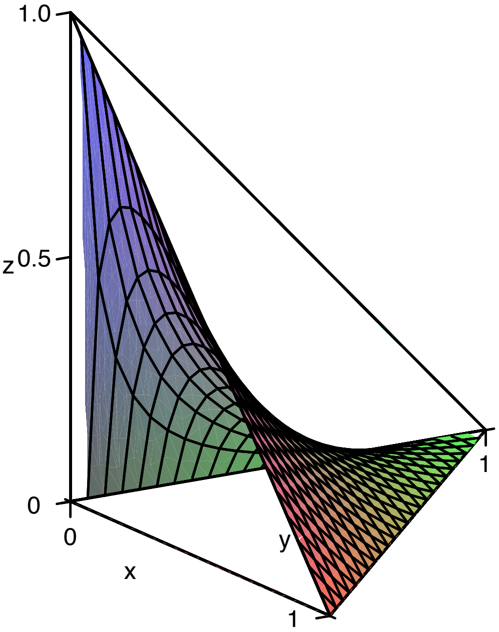
3. Applications of Variational Bayes

3.1. Geometric Foundation

We consider the simplest model that shows dependence. Let X_1, X_2 be two binary random variables, with distribution $\pi := (\pi_{00}, \pi_{10}, \pi_{01}, \pi_{11})$, where $P(X_1 = i, X_2 = j) = \pi_{ij}$, $i, j \in \{0, 1\}$. Further, let the marginal distributions be denoted by $\pi_1 = P(X_1 = 1), \pi_2 = P(X_2 = 1)$. We want to consider the geometry of the VB projection from a general distribution to the family of independent distributions. This represents the way that VB gains its computational advantages by simplifying the posterior dependence structure.

The model space is illustrated in Figure 2, where π is represented by a point in the three simplex, and the independence surface, where $\pi_{00}\pi_{11} = \pi_{10}\pi_{01}$, is also shown.

Figure 2. Space of distributions with independence surface: marginal probability and dependence.



Both the interior of the simplex and independence surface are exponential families, and it is convenient to use the natural parameters for the interior of the simplex:

$$\xi_1 = \log \frac{\pi_{10}}{\pi_{00}}, \xi_2 = \log \frac{\pi_{01}}{\pi_{00}}, \xi_3 = \log \frac{\pi_{11}\pi_{00}}{\pi_{10}\pi_{01}}$$

where the independence surface is given by $\xi_3 = 0$. The independence surface can also be parameterised by the marginal distributions π_1, π_2 or the corresponding natural parameters $\xi_i^{ind} := \log(\pi_i/(1-\pi_i))$. For any distribution, π, represented in natural parameters by (ξ_1, ξ_2, ξ_3), has its VB approximation defined implicitly by the simultaneous equations:

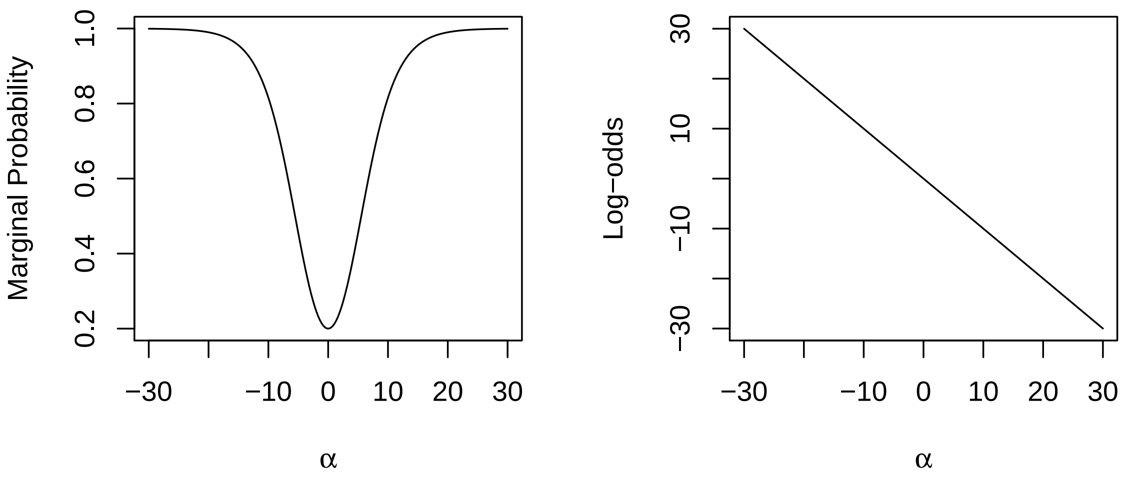
$$\xi_1^{ind}(\pi_1) = \xi_1 + \xi_3\pi_2, \quad (4)$$

$$\xi_2^{ind}(\pi_2) = \xi_2 + \xi_3\pi_1. \quad (5)$$

These can be solved, as is typical with VB methods, by iterating updated estimates of π_1 and π_2 across the two equations. We show this in a realistic example in the following section.

Having seen the VB solution in this simple model, we can investigate the quality of the approximation. If we were using the forward KL project, as proposed by [2], then the mean will be preserved by the approximation, while, of course, the variance structure is distorted. In the case of using the reverse KL projection, as used by VB, the mean will not be preserved, but in this example, we can investigate the distortion explicitly. Let $(\xi_1(\alpha), \xi_2(\alpha), \xi_3(\alpha))$ be a $+1$-geodesic, which cuts the independence surface orthogonally and is parameterised by α, where $\alpha = 0$ corresponds to the independence surface. In this example, all such geodesics can be computed explicitly. Figure 3 shows the distortion associated with the VB approximation. In the left-hand panel, we show the mean, which is the marginal probability, $P(X_1 = 1)$, for all points on the orthogonal geodesic. We see, as expected, that this is not constant, but it is locally constant at $\alpha = 0$, showing that the distortion of the mean can be small near the independence surface. The right-hand panel shows the dependence, as measured by the log-odds, for points on the geodesic. As expected, the VB does not preserve the dependence structure; indeed, it is designed to exploit the simplification of the dependence structure.

Figure 3. Distortion implied by variational Bayes (VB) approximation.



3.2. Variational Bayes for the RSLN Model

The regime-switching log-normal model [19] with a fixed finite number, K, of regimes can be described as a bivariate discrete time process with the observed data sequence $w_{1:T} = \{w_t\}_{t=1}^T$ and the unobserved regime sequence $S_{1:T} = \{S_t\}_{t=1}^T$, where $S_t \in \{1, \cdots, K\}$ and T is the number of observations. The logarithm of w_t, denoted by $y_t = \log w_t$, is assumed normally distributed, having mean μ_i and variance σ_i^2 both dependent on the hidden regime S_t. The sequence of $S_{1:T}$ is assumed to follow a first order Markov chain having transition probabilities $A = (a_{ij})$ with the probabilities $\pi = (\pi_i)_{i=1}^K$ to start the first regime.

The RSLN model is a special case of more general state-space models, which were studied in detail by [28]. In this paper, we use this model and simulated and real data to illustrate the VB method in practice. We also calibrate its performance by referring to [24], which used MCMC methods to fit the same model to the same data. Here, we are regarding the MCMC analysis as a form of “gold-standard", but with the cost of being orders-of-magnitude slower than VB in computational time.

In the Bayesian framework, we use a symmetric Dirichlet prior for π, that is $p(\pi) = \text{Dir}(\pi; \frac{C^\pi}{K}, \cdots, \frac{C^\pi}{K})$, for $C^\pi > 0$. Let a_i denote the $i - th$ row vector of A. The prior for A is chosen as $p(A) = \prod_{i=1}^K p(a_i) = \prod_{i=1}^K \text{Dir}(a_i; \frac{C^A}{K}, \cdots, \frac{C^A}{K})$, for $C^A > 0$, and the prior distribution for $\{(\mu_i, \sigma_i^2)\}_{i=1}^K$ is chosen to be normal-inverse gamma, $p(\{\mu_i, \sigma_i^2\}_{i=1}^K) = \prod_{i=1}^K \text{N}(\mu_i|\sigma_i^2; \gamma, \frac{\sigma_i^2}{\eta^2})\text{IG}(\sigma_i^2; \alpha, \beta)$. In the above setting, C^π, C^A, γ, η^2, α and β are hyper-parameters. Thus, the joint posterior distribution of $\pi, A, \{\mu_i, \sigma_i^2\}_{i=1}^K$, and $S_{1:T}$ is $P(\pi, A, \{\mu_i, \sigma_i^2\}_{i=1}^K, S_{1:T}|y_{1:T})$ and is proportional to:

$$p(S_1|\pi) \prod_{t=1}^{T-1} p(S_{t+1}|S_t; A) \prod_{t=1}^{T} p(y_t|S_t; \{\mu_i, \sigma_i^2\}_{i=1}^K) p(\pi) p(A) p(\{\mu_i, \sigma_i^2\}_{i=1}^K). \tag{6}$$

This posterior distribution and its corresponding marginal posterior distributions are analytically intractable. In VB, we seek an approximation of Equation (6), denoted by $q(\pi, A, \{\mu_i, \sigma_i^2\}_{i=1}^K, S_{1:T})$, to which we want to balance two things: having the discrepancy between Equation (6) and q small, while keeping q tractable. In general, there are two ways to choose q. The first is to specify a particular distributional family for q, for example the multivariate normal distribution. The other is to choose q with a simpler dependency structure than that of Equation (6); for example, we choose q, which factorizes as:

$$q(\pi, A, \{\mu_i, \sigma_i^2\}_{i=1}^K, S_{1:T}) = q(\pi) \prod_{i=1}^K q(a_i) \prod_{i=1}^K q(\mu_i|\sigma_i^2) q(\sigma_i^2) q(S_{1:T}). \tag{7}$$

The Kullback–Leibler (KL) divergence [29] can be used as the measure of dissimilarity between Equations (6) and (7). For succinctness, we denote $\tau = (\pi, A, \{\mu_i, \sigma_i^2\}_{i=1}^K, S_{1:T})$; thus the KL divergence is defined as:

$$\text{KL}(q(\tau) \mid\mid p(\tau|y)) = \int q(\tau) \log \frac{q(\tau)}{p(\tau|y)} d\tau. \tag{8}$$

Note that the evaluation of Equation (8) requires $p(\tau|y)$, which is unavailable. However, we note that:

$$\text{KL}(q(\tau) \mid\mid p(\tau|y)) = \log p(y) - \int q(\tau) \log \frac{p(\tau, y)}{q(\tau)} d\tau$$

Given the factorization Equation (7), this can be written as:

$$\mathrm{KL}(q(\tau)\,||\,p(\tau|y)) =$$

$$\log p(y) - \int \sum_{S_{1:T}} q(\pi)q(A)\prod_{i=1}^{K} q(\mu_i|\sigma_i^2)q(\sigma_i^2)q(S_{1:T}) \log \frac{p(\pi, A, \{\mu_i, \sigma_i^2\}_{i=1}^{K}, S_{1:T}, y_{1:T})}{q(\pi)q(A)\prod_{i=1}^{K} q(\mu_i|\sigma_i^2)q(\sigma_i^2)q(S_{1:T})} d\pi dA d\{\mu_i, \sigma_i^2\}_{i=1}^{K}$$

Consider first the $q(\pi)$ term. The right-hand side can be rearranged as:

$$\mathrm{KL}\left(q(\pi) \middle|\middle| \frac{\exp\left[\int \sum_{S_{1:T}} q(S_{1:T})q(A)\prod_{i=1}^{K} q(\mu_i|\sigma_i^2)q(\sigma_i^2) \log p(\pi, A, \{\mu_i, \sigma_i^2\}_{i=1}^{K}, s_{1:T}, y_{1:T}) dA d\{\mu_i, \sigma_i^2\}_{i=1}^{K} \right]}{Z_\pi} \right) + K_\pi, \tag{9}$$

where:

$$K_\pi = \int \sum_{S_{1:T}} q(S_{1:T})q(A)\prod_{i=1}^{K} q(\mu_i|\sigma_i^2)q(\sigma_i^2)q(S_{1:T}) \log q(A)\prod_{i=1}^{K} q(\mu_i|\sigma_i^2)q(\sigma_i^2) dA d\{\mu_i, \sigma_i^2\}_{i=1}^{K} - \log Z_\pi + \log p(y),$$

and Z_π is a normalizing term. The first term of Equation (9) is the only term that depends on $q(\pi)$. Thus, the minimum value of $\mathrm{KL}(q(\tau)\,||\,p(\tau|y))$ is achieved when this term equals zero. Hence, we obtained:

$$q(\pi) = \frac{\exp\left[\int \sum_{S_{1:T}} q(S_{1:T})q(A)\prod_{i=1}^{K} q(\mu_i|\sigma_i^2)q(\sigma_i^2) \log p(\pi, A, \{\mu_i, \sigma_i^2\}_{i=1}^{K}, s_{1:T}, y_{1:T}) dA d\{\mu_i, \sigma_i^2\}_{i=1}^{K} \right]}{Z_\pi} \tag{10}$$

Given the joint distribution of $p(\pi, A, \{\mu_i, \sigma_i^2\}_{i=1}^{K}, s_{1:T}, y_{1:T})$ in the form of Equation (6), the straightforward evaluation of Equation (10) results in:

$$q(\pi) \;\propto\; \prod_{i=1}^{K} \pi_i^{\frac{C_\pi^K}{K} + w_i^s - 1} = \mathrm{Dir}(\pi, w_1^\pi, \cdots, w_K^\pi);\; w_i^\pi = \frac{C_\pi^K}{K} + w_i^s, w_i^s = \mathrm{E}_{q(S_{1:T})}[S_{1,i}] \tag{11}$$

where $S_{1,i} = 1$, if the process is in state i at time 1, and zero otherwise.

Similarly, we can rearrange Equation (9) with respect to $\{q(a_i)\}_{i=1}^{K}$, $\{q(\mu_i|\sigma_i^2)\}_{i=1}^{K}$, $\{q(\sigma_i^2)\}_{i=1}^{K}$ and $q(S_{1:T})$, respectively, and using the same arguments, then we can obtain:

$$q(A) = \prod_{i}^{k} \mathrm{Dir}(a_i; w_{i1}^A, ..., w_{ik}^A);\; w_{ij}^A = \frac{C^A}{K} + v_{ij}^s, \tag{12}$$

$$q(\mu_i|\sigma_i^2) = \mathrm{N}\left(\gamma_i', \frac{\sigma_i^2}{\kappa_i}\right), \gamma_i' = \frac{\eta^2\gamma + p_i^s}{\eta^2 + q_i^s}, \kappa_i = \eta^2 + q_i^s \tag{13}$$

$$q(\sigma_i^2) = \mathrm{IG}\left(\alpha_i', \beta_i'\right), \alpha_i' = \alpha + \frac{q_i^s}{2}, \beta_i' = \beta + \frac{r_i^s}{2} + \frac{\eta^2}{2}(\gamma_i' - \gamma)^2 \tag{14}$$

$$q(S_{1:T}) = \frac{\prod_{i=1}^{k} \pi_i^{*S_{1,i}} \prod_{t=1}^{T-1}\prod_{i=1}^{k}\prod_{j=1}^{k} a_{ij}^{*S_{t,i}S_{t+1,j}} \prod_{t=1}^{T}\prod_{i=1}^{k} \theta^{*S_{t,i}}}{\tilde{Z}}, \tag{15}$$

where $S_{t,i} = 1$, if the process in state i at time t, and zero otherwise, and with $\pi_i^* = e^{\mathrm{E}_{q(\pi)}[\log \pi_i]}$, $a_{ij}^* = e^{\mathrm{E}_{q(A)}[\log(a_{ij})]}$, $\theta_{i,t}^* = e^{\mathrm{E}_{q(\mu_i|\sigma_i^2)q(\sigma_i^2)}[\log \phi_i(y_t)]}$, $v_{ij}^s = \sum_{t=1}^{T-1} \mathrm{E}_{q(S_{1:T})}\left[S_{t,i}S_{t+1,j}\right]$, $p_i^s = \sum_{t=1}^{T} \mathrm{E}_{q(S_{1:T})}[s_{t,i}]y_t$, $q_i^s = \sum_{t=1}^{T} \mathrm{E}_{q(S_{1:T})}[s_{t,i}]$, $r_i^s = \sum_{t=1}^{T}(\gamma_i' - y_t)^2 E_{q(S)}[s_{t,i}]$. Here, ψ is the digamma function, ϕ is the normal density function and the exact functional forms used in the updates are shown in Algorithm 1.

Algorithm 1 Variational Bayes algorithm for the regime-switching log-normal model (RSLN) model.

```
Initialize $w_i^{s(0)}$, $v_{ij}^{s\,(0)}$, $p_i^{s(0)}$, $q_i^{s(0)}$, and $r_i^{s(0)}$ at step 0
while $w_i^{\pi(t-1)}$, $w_{ij}^{A(t-1)}$, $\gamma_i'^{(t-1)}$, $\alpha_i'^{(t-1)}$, $\beta_i'^{(t-1)}$, $\pi_i^{*(t-1)}$, $a_{ij}^{*\,(t-1)}$, and $\theta_{i,t}^{*\,(t-1)}$ do not converge do
    1. Compute $w_i^{\pi(t)}$, $w_{ij}^{A(t)}$, $\gamma_i'^{(t)}$, $\kappa_i^{(t)}$, $\alpha_i'^{(t)}$, and $\beta_i'^{(t)}$at step $t$ by
        $w_i^{\pi(t)} = \frac{C_\pi^K}{K} + w_i^{s(t-1)}, \quad w_{ij}^{A(t)} = \frac{C_\pi^A}{K} + v_{ij}^{s\,(t-1)}, \quad \gamma_i'^{(t)} = \frac{\eta^2\gamma + p_i^{s(t-1)}}{\eta^2 + q_i^{s(t-1)}},$
        ${\kappa_i}^{(t)} = \eta^2 + q_i^{s(t-1)}, \quad \alpha_i'^{(t)} = \alpha + \frac{q_i^{s(t-1)}}{2}, \quad \beta_i'^{(t)} = \beta + \frac{r_i^{s(t-1)}}{2} + \frac{\eta^2}{2}(\gamma_i'^{(t)} - \gamma)^2$
    2. Compute $\pi_i^{*(t)}$, $\theta_{i,t}^{*\,(t)}$ and $a_{ij}^{*\,(t)}$ at step $t$ by:
        $\pi_i^{*(t)} = \exp\big(\psi(w_i^{\pi(t)}) - \psi(\sum_i w_i^{\pi(t)})\big), \quad a_{ij}^{*\,(t)} = \exp\big(\psi(w_{ij}^{A(t)}) - \psi(\sum_{j=1} w_{ij}^{A(t)})\big)$
        $\theta_{i,t}^{*\,(t)} = \exp\big( - \frac{1}{2}\log 2\pi - \frac{1}{2}(\log \beta_i'^{(t)} - \psi(\alpha_i'^{(t)})) - \frac{1}{2}\left((y_t - \gamma_i'^{(t)})^2 \frac{\alpha_i'^{(t)}}{\beta_i'^{(t)}} + \frac{1}{{\kappa_i}^{(t)}}\right)\big)$
    3. Compute $w_i^{s(t)}$, $v_{ij}^{s\,(t)}$, $p_i^{s(t)}$, $q_i^{s(t)}$, and $r_i^{s(t)}$ at step $t$ by:
        $w_i^{s(t)} = \mathrm{E}_{q^{(t)}(S_{1:T})}[S_{1,i}], \; v_{ij}^{s\,(t)} = \sum_{t=1}^{T-1} \mathrm{E}_{q^{(t)}(S_{1:T})}\left[S_{t,i}S_{t+1,j}\right], \; p_i^{s(t)} = \sum_{t=1}^{T-1} \mathrm{E}_{q^{(t)}(S_{1:T})}[s_{t,i}]y_t,$
        $q_i^{s(t)} = \sum_{t=1}^{T-1} \mathrm{E}_{q^{(t)}(S_{1:T})}[s_{t,i}], \; r_i^{s(t)} = \sum_{t=1}^{T-1}(\gamma_i'^{(t)} - y_t)^2 E_{q^{(t)}(S)}[s_{t,i}]$
  $t \Leftarrow t + 1$
end while
```

The VB method proceeds, as was shown with the simple Equations (4) and (5), by iterative updating the variational parameters to solve a set of simultaneous equations. In this example, the update equations for the variables $\pi, A, \{\mu_i, \sigma_i^2\}_{i=1}^K, S_{1:T}$ are given explicitly by Algorithm 1. For the initialisation, we choose symmetric values for most of the parameters and choose random values for others, as appropriate. For this example, this worked very satisfactory, although we note that for more general state space models [28], states that find good initial values can be non-trivial.

3.3. Interpretation of Results

First, all approximating distributions above turn out to lie in well-known parametric families. The only unknown quantities are the parameters of these distributions, which are often called the variational parameters.

The evaluation of parameters of $q(\pi)$, $q(A)$, $q(\mu_i|\sigma_i^2)$, and $q(\sigma_i^2)$ requires knowledge of $q(S_{1:T})$, and also, the evaluation of π_i^*, a_{ij}^* and $\theta_{i,t}^*$ requires knowledge of $q(\pi)$, $q(A)$, $q(\mu_i|\sigma_i^2)$ and $q(\sigma_i^2)$. This structure leads to an iterative updating scheme, described in Algorithm 1.

The main computational effort in Algorithm 1 is computing $\mathrm{E}_{q(S_{1:T})}[S_{t,i}]$ and $\mathrm{E}_{q(S_{1:T})}\left[S_{t,i}S_{t+1,j}\right]$, which have no simple tractable forms. We note that the distributional form of $q(S_{1:T})$ has a very similar structure as the conditional distribution of $p(S_{1:T}|Y_{1:T},\tau)$ for which the forward-backward algorithm [30] is commonly used to compute $\mathrm{E}_{p(S_{1:T}|Y_{1:T},\tau)}[S_{t,i}|Y_{1:T},\boldsymbol{\tau}]$ and $\mathrm{E}_{p(S_{1:T}|Y_{1:T},\tau)}\left[S_{t,i}S_{t+1,j}|Y_{1:T},\tau\right]$. Therefore, we also use the forward-backward algorithm to compute $\mathrm{E}_{q(S_{1:T})}[S_{t,i}]$ and $\mathrm{E}_{q(S_{1:T})}\left[S_{t,i}S_{t+1,j}\right]$.

The conditional distribution of $q(\mu_i|\sigma_i^2)$ is $\mathrm{N}\left(\mu_i|\sigma_i^2;\gamma_i',\frac{\sigma_i^2}{\kappa_i}\right)$, then the marginal distribution of μ_i is the location-scale t distribution, denoted as $t_{2\alpha_i'}(\mu_i;\gamma_i',\frac{\kappa_i}{\beta_i'/\alpha_i'})$, where the density function of $t_\nu(x;\mu,\lambda)$ is defined as $p(x|\nu,\mu,\lambda)=\frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})}\left(\frac{\lambda}{\pi\nu}\right)^{\frac{1}{2}}\left[1+\frac{\lambda(x-\mu)^2}{\nu}\right]^{-\frac{\nu+1}{2}}$, for $x,\mu\in(-\infty,+\infty)$ and $\nu,\lambda>0$.

4. Numerical Studies

4.1. Simulated Data

In this section, we applied the VB solutions to four sets of simulated data, which are used in [24]. Through these simulated studies, we will test the performance of VB on detecting the number of regimes and compare it with those of the BIC and the MCMC methods [24]. For this paper, we present only an initial study with a relatively small number of datasets. The results are highly promising, but more extensive studies are needed to draw comprehensive conclusions. Furthermore, see [28] for general results on VB in hidden state space models.

To estimate the number of regimes, we construct a matrix, called the relative magnitude matrix (RMM), defined as $A'=\left(\hat{a}_{ij}'\right)$, where $\hat{a}_{ij}'=\frac{w_{ij}^A}{w_0^A}$, $w_0^A=\sum_{i=1}^K\sum_{j=1}^K w_{ij}^A$ and w_{ij}^A is the parameter of $q(A)$. Our model selection procedure is to fit a VB with a large number of regimes and to examine the rows and columns in the RMM. If the values of the entries in the $i-th$ row and the $i-th$ column of A' are all equal to $\frac{C^A/K}{T-1+C^A\times K}$, then we will declare the regime i nonexistent. This method is validated by the following observations. It can be shown that the parameter of v_{ij}^s in w_{ij}^A is equal to the number of times the process leaves regime i and enters regime j. Therefore, for the $i-th$ regime, the values of zero for all of v_{ji}^s and v_{ij}^s with $j=1,\cdots,K$ indicate that there is no transition process entering or leaving regime i.

Table 1 specifies the parameters for the four cases, and we generate 671 observations for each case (equal to the number of months from January 1956 to September 2011). The parameters used in Case 1 are identical to the maximum likelihood estimates for TSX monthly return data from 1956 to 1999 [19]. Case 2 only has one regime present. Case 3 is similar to Case 1, but the two regimes have the same mean. Case 4 adds a third regime. For each case, we use MLE to fit a one-regime, two-regime, three-regime and four-regime RSLN model and report the corresponding BIC and log-likelihood scores. We then misspecify the number of regimes and run a four-regime VB algorithm.

Table 1. Parameters of the simulated data.

Case	**Regime 1**	**Regime 2**	**Regime 3**	**Transition Probability**
	(μ_i, σ_i)	(μ_i, σ_i)	(μ_i, σ_i)	
1	(0.012, 0.035)	(−0.016, 0.078)	-	$\begin{pmatrix} 0.963 & 0.037 \\ 0.210 & 0.790 \end{pmatrix}$
2	(0.014, 0.050)	-	-	-
3	(0.000, 0.035)	(0.000, 0.078)	-	$\begin{pmatrix} 0.963 & 0.037 \\ 0.210 & 0.790 \end{pmatrix}$
4	(0.012, 0.035)	(−0.016, 0.078)	(0.04, 0.01)	$\begin{pmatrix} 0.953 & 0.037 & 0.01 \\ 0.210 & 0.780 & 0.01 \\ 0.80 & 0.190 & 0.01 \end{pmatrix}$

Table 2 shows the number of iterations that VB takes to converge in each case and the corresponding computational time (on a MacBook, 2 GHz processor). On average, VB converges after a hundred iterations and takes about one minute. On the same computer, a 10^4-iteration Reverse Jump MCMC (RJMCMC) will take about 10 h to finish. Using diagnostics, this seemed to be enough for convergence, while not being an "unfair" comparison in terms of time with VB. We can see that the computational efficiency will be a very attractive feature of the VB method. The results of the BIC with the log-likelihood (in parentheses), the relative magnitude matrices and the posterior probabilities for the models with the different number of regimes estimated by MCMC (cited from Hartman and Heaton [24]) are given in Table 3. In Case 1, the BIC favors the two-regime model. The posterior probability estimated by MCMC for the one-regime model is the largest, but there is still a large probability for the two regime model. Note that the prior specification for the number of regimes can effect these numbers and is always an issue with these forms of multidimensional MCMC. The relative magnitude matrix clearly shows that there are only two regimes whose $\hat{a}'_{ij}$ are not negligible. This implies VB removes excess transition and emission processes and discovers the exact number of hidden regimes. In Case 2 and Case 3, both VB and the BIC can select the correct number of regimes, and the posterior probability for the one-regime model estimated by MCMC is still the largest. In Case 4, VB does not detect the third regime. The transition probability to this regime is only 0.01, and the means and standard deviations of Regime 1 make the rare data from Regime 3 easily merged within the data from Regime 1. From Table 3, it is clear that for all of the cases, the log-likelihood always increases as the number of regimes increase.

Table 2. Computational efficiency of VB.

	Case 1	**Case 2**	**Case 3**	**Case 4**
Iterations to converge	62	182	132	94
Computational time [s]	27.161	80.842	58.510	45.044

Table 3. The estimated number of regimes by VB, BIC and MCMC.

Case	No. of Regimes	MLE BIC (Log Likelihood)	RJMCMC Posterior Probability	VB Relative Magnitude Matrix
1	1	$1,108.875(1,115.384)$	0.647	$\begin{pmatrix} 0.14357 & 0.00004 & 0.00004 & 0.03153 \\ 0.00004 & 0.00004 & 0.00004 & 0.00004 \\ 0.00004 & 0.00004 & 0.00004 & 0.00004 \\ 0.03018 & 0.00004 & 0.00004 & 0.79428 \end{pmatrix}$
	2	$1,158.227(1,174.499)$	0.214	
	3	$1,156.370(1,182.405)$	0.088	
	4	$1,153.150(1,188.948)$	<0.052	
2	1	$1,045.448(1,051.957)$	0.864	$\begin{pmatrix} 0.99944 & 0.00004 & 0.00004 & 0.00004 \\ 0.00004 & 0.00004 & 0.00004 & 0.00004 \\ 0.00004 & 0.00004 & 0.00004 & 0.00004 \\ 0.00004 & 0.00004 & 0.00004 & 0.00004 \end{pmatrix}$
	2	$1,038.360(1,054.632)$	0.109	
	3	$1,030.733(1,056.768)$	0.020	
	4	$1,026.882(1,062.680)$	<0.006	
3	1	$1,110.903(1,117.411)$	0.629	$\begin{pmatrix} 0.11322 & 0.00004 & 0.00004 & 0.02647 \\ 0.00004 & 0.00004 & 0.00004 & 0.00004 \\ 0.00004 & 0.00004 & 0.00004 & 0.00004 \\ 0.02659 & 0.00004 & 0.00004 & 0.83327 \end{pmatrix}$
	2	$1,139.214(1,155.486)$	0.221	
	3	$1,131.904(1,157.719)$	0.098	
	4	$1,121.921(1,157.940)$	<0.052	
4	1	$1,044.819(1,051.328)$	0.641	$\begin{pmatrix} 0.22643 & 0.00004 & 0.00004 & 0.05518 \\ 0.00004 & 0.00004 & 0.00004 & 0.00004 \\ 0.00004 & 0.00004 & 0.00004 & 0.00004 \\ 0.05377 & 0.00004 & 0.00004 & 0.66417 \end{pmatrix}$
	2	$1,092.610(1,108.881)$	0.203	
	3	$1,087.435(1,113.470)$	0.094	
	4	$1,080.240(1,116.038)$	<0.06	

4.2. Real Data

In this section, we apply the VB solution to the TSX monthly total return index in the period from January, 1956, to December, 1999 (528 observations in total and studied in [19,21]).

A four-regime VB is implemented first. VB converges after 100 iterations about 34.284 s (on a MacBook, 2 GHz processor). The relative magnitude matrix, given in Table 4, clearly shows that VB identifies two regimes. This matches both of the BIC and AIC-based results [19]. Based on these results, we then fit a two-regime VB, which converges after 83 iterations in about 14.241 s. Table 5 gives the marginal distributions for all of the parameters. Figure 4 presents the corresponding density functions, where we can see that all of the plots show a symmetric and bell-shaped pattern.

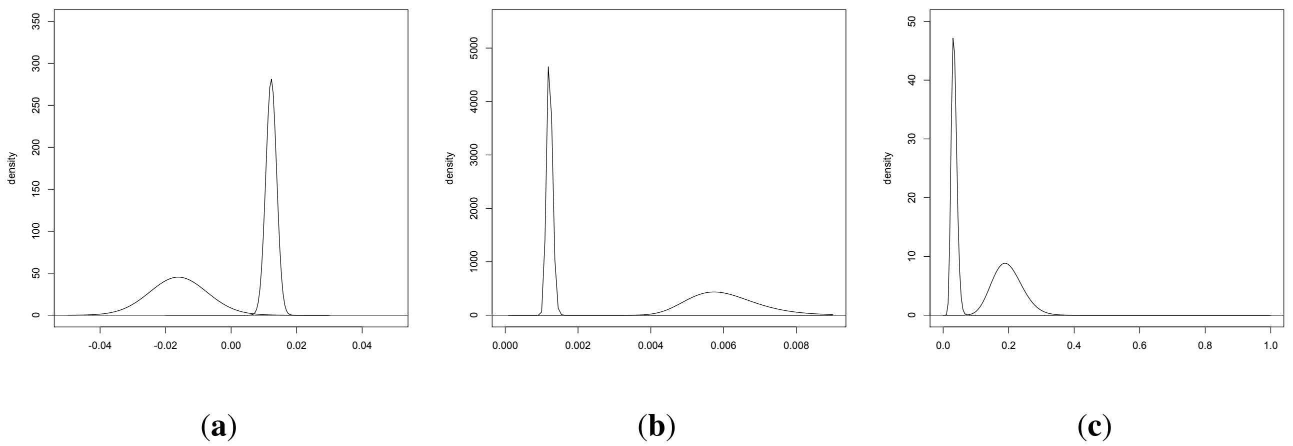
Table 4. Estimations of the number of regimes for TSXdata.

	January 1956–December 1999
R. M. M.	$\begin{pmatrix} 0.11496 & 0.00005 & 0.00005 & 0.02803 \\ 0.00005 & 0.00005 & 0.00005 & 0.00005 \\ 0.00005 & 0.00005 & 0.00005 & 0.00005 \\ 0.02853 & 0.00005 & 0.00005 & 0.82791 \end{pmatrix}$

Table 5. The marginal distributions of the parameters estimated by VB.

Parameter	**Distribution**	**Mean**	***s.d.***	**Transition Probability**
μ_1	$t_{454.61}(0.0123, 370778.19)$	0.0123	0.00165	-
σ_1^2	$\mathrm{IG}(227.30, 0.28)$	0.00122(0.0349)	0.00008	-
μ_2	$t_{80.39}(-0.0161, 12987.55)$	-0.0161	0.00889	-
σ_2^2	$\mathrm{IG}(40.20, 0.24)$	0.00603(0.0777)	0.00098	-
$p_{1,2}$	$\mathrm{Beta}(15.21, 434.78)$	0.0338	0.00851	$\begin{pmatrix} 0.9662 & 0.0338 \\ 0.1969 & 0.8031 \end{pmatrix}$
$p_{2,1}$	$\mathrm{Beta}(15.00, 61.21)$	0.1969	0.04525	

Figure 4. The VB marginal distributions of the parameters. (**a**) μ_2 (left) and μ_1 (right); (**b**) σ_1^2 (left) and σ_2^2 (right) ; (**c**) $p_{1,2}$ (left) and $p_{2,1}$ (right) .



(**a**) (**b**) (**c**)

Table 6 (the upper part) gives the maximum likelihood estimates (cited from [19]), mean parameters computed by the MCMC method (cited from [21]) and mean parameters computed by VB. It clearly shows that the point estimates by VB are very close to those by MLE and MCMC. The numbers in parenthesis in Table 6 are the standard deviations computed by the three methods, respectively. It is worth noting that all of the variance estimated by VB are smaller than those by the MLE or MCMC methods. In fact, some other researchers also report the underestimation of posterior variance in other VB applications, for example [31,32]. In the paper [1], we look at some diagnostics methods that can assess how well the VB approximates the true posterior, particularly with regards to its covariance structure. The methods proposed also allow us to generate simple corrections when the approximation error is large.

Table 6. Estimates and standard deviations by VB, MLE and MCMC.

	μ_1	σ_1	$p_{1,2}$	μ_2	σ_2	$p_{2,1}$
VB	0.0123(0.00165)	0.0349(0.00008)	0.0338(0.00851)	−0.0161(0.00889)	0.0777(0.00098)	0.1969(0.04525)
MLE	0.0123(0.002)	0.0347(0.001)	0.0371(0.012)	−0.0157(0.010)	0.0778(0.009)	0.2101(0.086)
MCMC	0.0122(0.002)	0.0351(0.002)	0.0334(0.012)	−0.0164(0.010)	0.0804(0.009)	0.2058(0.065)

5. Conclusions

Variational Bayes can be thought of in terms of information geometry as a projection-based approximation technique; it provides a framework to approximate posteriors. We applied this method to the regime-switching log-normal model and provide solutions to account for both model uncertainty and parameter uncertainty. The numerical results show that our method can recover exactly the number of regimes and gives reasonable point estimates. The VB method is also demonstrated to be very computationally efficient.

The application on the TSX monthly total return index data in the period from January 1956 to December 1999, confirms the similar results in the literature in finding the number of regimes.

Author Contributions

The article was written by Hui Zhao under the guidance of Paul Marriott. All authors have read and approved the final manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Zhao, H.; Marriott, P. Diagnostics for variational bayes approximations. **2013**, arXiv:1309.5117.
2. Amari, S.-I. *Differential-Geometrical Methods in Statistics*; Springer: New York, NY, USA, 1990.
3. Eguchi, S. Second order efficiency of minimum contrast estimators in a curved exponential family. *Ann. Stat.* **1983**, *11*, 793–803.
4. Kass, R.; Vos, P. *Geometrical Foundations of Asymptotic Inference*; Wiley: New York, NY, USA, 1997.

5. Hinton, G.E.; van Camp, D. Keeping neural networks simple by minimizing the description length of the weights. In Proceedings of the 6th ACM Conference on Computational Learning Theory, Santa Cruz, CA, USA, 26–28 July 1993; ACM: New York, NY, USA, 1993.
6. MacKay, D. Developments in Probabilistic Modelling with Neural Networks—Ensemble Learning. In *Neural Networks: Artifical Intelligence and Industrial Applications*; Springer: London, UK, 1995; pp. 191–198.
7. Attias, H. Independent Factor Analysis. *Neur. Comput.* **1999**, *11*, 803–851.
8. Lappalainen, H. Ensemble Learning For Independent Component Analysis. In Proceedings of the First International Workshop on Independent Component Analysis, Aussois, France, 11–15 January 1999; pp. 7–12.
9. Beal, M.; Ghahramani, Z. The variational Bayesian EM algorithm for incomplete data: With application to scoring graphical model structures. *Bayesian Stat.* **2003**, *7*, 453–463.
10. Winn, J. *Variational Message Passing and its Applications*. Ph.D. Thesis, Department of Physics, University of Cambridge, Cambridge, UK, 2003.
11. Blei, D.M.; Ng, A.Y.; Jordan, M.I.; Lafferty, J. Latent Dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
12. Ghahramani, Z.; Beal, M.J. A Variational Inference for Bayesian Mixtures of Factor Analysers. *Adv. Neur. Inf. Process. Syst.* **2000**, *12*, 449–455.
13. Haff, L.R. The Variational Form of Certain Bayes Estimators. *Ann. Stat.* **1991**, *19*, 1163–1190.
14. Faes, C.; Ormerod, J.T.; Wand, M.P. Variational Bayesian Inference for Parametric and Nonparametric Regression With Missing Data. *J. Am. Stat. Assoc.* **2011**, *106*, 959–971.
15. McGrory, C.; Titterington, D.; Reeves, R.; Pettitt, A.N. Variational Bayes for estimating the parameters of a hidden Potts model. *Stat. Comput.* **2009**, *19*, 329–340.
16. Ormerod, J.T.; Wand, M.P. Gaussian Variational Approximate Inference for Generalized Linear Mixed Models. *J. Comput. Graph. Stat.* **2011**, *21*, 1–16.
17. Hall, P.; Humphreys, K.; Titterington, D.M. On the Adequacy of Variational Lower Bound Functions for Likelihood-Based Inference in Markovian Models with Missing Values. *J. R. Stat. Soc. Ser. B* **2002**, *64*, 549–564.
18. Wang, B.; Titterington, M. Convergence Properties of a general algorithm for calculating variational Bayesian estimates for a normal mixture model. *Bayesian Anal.* **2006**, *1*, 625–650.
19. Hardy, M.R. A Regime-Switching Model of Long-Term Stock Returns. *N. Am. Actuar. J.* **2001**, *5*, 41–53.
20. Hamilton, J.D. A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle. *Econometrica* **1989**, *57*, 357–384.
21. Hardy, M.R. Bayesian Risk Management for Equity-Linked Insurance. *Scand. Actuar. J.* **2002**, *2002*, 185–211.
22. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **1978**, *6*, 461–464.
23. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Autom. Control* **1974**, *19*, 716–723.
24. Hartman, B.M.; Heaton, M.J. Accounting for regime and parameter uncertainty in regime-switching models. *Insur. Math. Econ.* **2011**, *49*, 429–437.

25. Green, P.J. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **1995**, *82*, 711–732.
26. Watanabe, S. *Algebraic Geometry and Statistical Learning Theory*; Cambridge University Press: Cambridge, UK, 2009.
27. Brooks, S.P. Markov Chain Monte Carlo Method and Its Application. *J. R. Stat. Soc. Ser. D* **1998**, *47*, 69–100.
28. Ghahramani, Z.; Hinton, G.E. Variational learning for switching state-space models. *Neur. Comput.* **1998**, *12*, 831–864.
29. Kullback, S.; Leibler, R.A. On information and sufficiency. *Ann. Math. Stat* **1951**, *22*, 79–86.
30. Baum, L.E.; Petrie, T.; Soules, G.; Weiss, N. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *Ann. Math. Stat.* **1970**, *41*, 164–171.
31. Rue, H.; Martino, S.; Chopin, N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. Ser. B* **2009**, *71*, 319–392.
32. Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer: New York, NY, USA, 2006.