

Predicting time series: A Pólya-Gamma data augmentation scheme within the Non-Homogeneous hidden Markov model

Constandina Koki^{1,*}, Loukia Meligkotsidou², and Ioannis Vrontos¹

¹Athens Univeristy of Economics and Bussiness, 76 Patission Str,
Athens, Greece

²National and Kapodistrian University of Athens,
Panepistimioupolis, Athens, Greece

*Corresponding author: Constandina Koki, kokiconst@aueb.gr

Abstract

We consider Non-Homogeneous Hidden Markov Models (NHHMMs) for forecasting univariate time series. We introduce two state NHHMMs where the time series are modeled via different predictive regression models for each state. Also, the time-varying transition probabilities depend on exogenous variables through a logistic function. In a hidden Markov setting, inference for logistic regression coefficients becomes complicated and in some cases impossible due to convergence issues. In this paper we try to address this problem. We use a new latent variable scheme, that utilizes the Pólya-Gamma class of distributions. Given an available set of predictors, we allow for model uncertainty regarding the predictors that affect the series both linearly – in the mean – and non-linearly – in the transition matrix. Predictor selection and inference on the model parameters are based on a MCMC scheme with reversible jump steps. Single-step and multiple-steps-ahead predictions are obtained based on the most probable model, median probability model or a Bayesian Model Averaging (BMA) approach. Using simulation experiments, as well as an empirical study on realized volatility data set, illustrate the performance of our algorithm in various setups, in terms of mixing properties, model selection and predictive ability.

Keywords: Non Homogeneous Hidden Markov Models; Model selection; Forecasting; Pólya-Gamma Data Augmentation, Realized Volatility

JEL classification: C11;C15;C51;C52;C53

1 Introduction

Discrete-time finite state-space Homogeneous Hidden Markov models (HHMMs) have been extensively studied and used to model stochastic processes that consist of an observed process and a latent (hidden) sequence of states which is assumed to affect the observation sequence, see for example Cappé et al. [2005]

and Billio et al. [1999]. Bayesian inference, using Markov chain Monte Carlo (MCMC) techniques, has enhanced the applicability of HHMMs and has led to the construction of more complex model specifications including NHHMMs. Initially, Diebold et al. [1994] studied the two-state Gaussian Non-Homogeneous Hidden Markov Models (NHHMMs) where the time varying transition probabilities were modeled via logistic functions. Their approach was based on the Expectation-Maximization algorithm (EM). Filardo and Gordon [1998] adopted a Bayesian perspective to overcome technical and calculation issues of classical approaches. Since then, various Bayesian methods have been proposed in the literature. For example, Spezia [2006] modeled the time-varying transition probabilities via a logistic function depending on exogenous variables and performed model selection based on the Bayes factor. In the same spirit, Meligkotsidou and Dellaportas [2011] considered an m -state NHHMM and assumed that the elements of the transition matrix are linked through exogenous variables with a multinomial logistic link, whereas the observed process conditional on the unobserved process follows an autoregressive model of order p . They accommodate and exploit model uncertainty within their Bayesian modeling –they allow variable selection to covariates affecting only the transition probability matrix– to improve the predictive ability of NHHMMs on economic data series.

Motivated by the fact that, when there is model uncertainty the data augmentation scheme of Holmes and Held [2006] for estimating the multinomial/logistic regression coefficients in Meligkotsidou and Dellaportas [2011] is unstable, this paper extends the work of Meligkotsidou and Dellaportas [2011] by employing recent methodological advances on Pólya-Gamma data augmentation for predictive discrete-time, finite state-space, NHHMMs that can be used for modeling and predicting univariate time-series. We consider two-state NHHMMs (easily extended to m -state NHHMMs) in which the time series are modeled via different predictive regression models for each state, whereas the transition probabilities are modeled via logistic regressions. Given an available set of predictors, we allow for model uncertainty regarding the predictors that affect the series both linearly–directly in the mean regressions– and non-linearly –in the transition probability matrix.

We perform Bayesian inference for the proposed Non-Homogeneous Pólya-Gamma Hidden Markov model (NHPG) based on MCMC. Our MCMC scheme aims at overcoming difficulties and solving convergence issues arising with existing MCMC algorithms, thus offering an improved inferential procedure for estimation and variable selection in NHHMMs, as well as more accurate forecasts. To this end we exploit the missing data representation of hidden Markov models and construct an MCMC algorithm based on data augmentation, consisting of several steps. First, the latent sequence of states is sampled via the Scaled Forward-Backward algorithm of Scott [2002], which is a modification of the Forward-Backward algorithm of Baum et al. [1970] who used it to implement the classical EM algorithm. Then, following Meligkotsidou and Dellaportas [2011], we use a logistic regression representation of the transition probabilities and simulate the parameters of the mean predictive regression model for each state, via Gibbs sampling steps.

Many data-augmentation or Metropolis-Hastings algorithms have been proposed to infer the logistic regression model, see for example O’Brien and Dunson [2004], Holmes and Held [2006], Frühwirth-Schnatter and Frühwirth [2010], Fussl et al. [2013], Polson et al. [2013]. Using the data augmentation scheme

of the Holmes and Held [2006], as in Meligkotsidou and Dellaportas [2011], may result in convergence issues, especially in cases that there exists model uncertainty. To deal with this issue, we use the Pólya-Gamma data augmentation scheme of Polson et al. [2013]. The authors in Polson et al. [2013] proved that this data augmentation scheme has a significantly improved performance. The recent work of Holsclaw et al. [2017] confirms that using Pólya-Gamma data augmentation to parametrize the transition probabilities of NHHMMs results in an algorithm that mixes well and provides adequate estimates of the model parameters. Finally, we incorporate variable selection within our MCMC scheme by using the well-known model selection method of Green [1995], the reversible jump algorithm. Within our algorithm we perform a couple of reversible jump steps to allow for a different set of covariates to affect the mean equation and the transition probabilities.

Different approaches have been used in the literature to cope with the model selection problem. The use of information criteria, such as Akaike’s Information Criterion (AIC, Akaike et al. [1973]), the Bayesian Information Criterion (BIC) of Schwarz [1978], the Deviance Information Criterion (DIC, Spiegelhalter et al. [2002]) or the Widely applicable Bayesian Information Criterion (WBIC, Watanabe [2013]), is another approach to variable selection. Holsclaw et al. [2017] consider a NHHMM similar to ours for modeling multivariate meteorological time series data. In that paper, the transition probabilities are modeled via multinomial logistic regressions affected by a specific set of exogenous variables. The authors use the BIC criterion for choosing the best model among a specified class of models. We extend this work by considering the problems of statistical inference and variable selection jointly, in a purely Bayesian setting, and we describe the proposed methodology in the context of 2-state NHHMMs, noting though that it can be easily extended to the case of m -state models. The proposed model is flexible, since we do not decide a priori which covariates affect the observed or the unobserved process. Instead, we have a common pool of covariates $\{X\}$ and within the MCMC algorithm we gauge which covariates are included in subset $\{X^{(1)}\}$, affecting the mean predictive equation of the observed process, and which covariates are included in subset $\{X^{(2)}\}$, affecting the time-varying transition probabilities.

The probabilistic approach we use is based on the calculation of the posterior distribution of different NHPGs or equivalently on computing the posterior probabilities of different NHPGs. Posterior probabilities can be used either for selecting the most probable model (i.e., making inference using the model with the highest posterior probability), or for Bayesian model averaging (i.e., producing inferences averaged over different NHHMMs). Barbieri and Berger [2004] argue that the optimal predictive model is not necessarily the model with the highest posterior probability but the median probability model, which is defined as the model consisting of those covariates which have overall posterior probability of being included in the model –inclusion probability– greater or equal to 0.5. Our method allows us to calculate the posterior probability of the model, as well as the probabilities of inclusion. To this end, we develop a Reversible Jump Markov Chain Monte Carlo (RJMCMC) algorithm (Green [1995], Hastie and Green [2011]) which explores the model space by jumping between different hidden Markov models. A more applied tutorial can be found in Waagepetersen and Sorensen [2001]. A study for comparing variable selection methods is well

presented in O'Hara and Sillanpää [2009] whilst Dellaportas et al. [2002] study the variable selection methods in the context of model choice.

We use our NHPG model for predicting realized volatility. Accurate forecasting of future volatility is important for asset allocation, portfolio construction and risk management see Gospodinov et al. [2006]. A review on the realized volatility literature can be found in McAleer and Medeiros [2008]. A vast amount of literature has investigated the relationship between volatility and macroeconomic and/or financial variables (see for example, Paye [2012], Christiansen et al. [2012], Meligkotsidou et al. [2018] among others). The proposed NHPG is able to capture not only the linear relationship between the logarithm of realized volatility and a set of predictors, as in Christiansen et al. [2012], but the nonlinear relationship, as well as other special characteristics of the analyzed series, such as heteroscedasticity and autocorrelation. NHPG outperforms the models of Meligkotsidou and Dellaportas [2011] and Christiansen et al. [2012] and the homogeneous HMM with autoregressive terms, in terms of forecasting ability, see Table 5.

Finally, we use the Continuous Rank Probability Score (CRPS) as our main forecasting scoring rule. In our setting, the predictive density of the MCMC output is multimodal and thus scoring rules that are sensible to distance should be preferred (Gneiting and Raftery [2007]). The logarithmic scoring rule for evaluating different models and their predictive accuracy is not appropriate, since the estimated predictive distribution is multimodal. It gives harsh penalty for low probability events (Boero et al. [2011], Gneiting and Raftery [2007]) prefers the forecast density that is less informative (Machete [2013]). We propose the use of the Continuous Ranked Probability Score which is defined using the cumulative predictive distribution function. This proper scoring rule has gained a lot of interest in the meteorological community (Grimmett et al. [2006]), as a better alternative for assessing the quality of forecasts as well as for validating the model performance.

In summary, our main contributions of our paper are the following

1. We present experimental evidence in support of the claim that NHPG has an improved performance in terms of variable selection and forecasting ability when comparing it with Meligkotsidou and Dellaportas [2011], at no cost of computational burden and running time.
2. We provide a stable algorithm without the need of tuning, that can be used as a black box for predicting time series.
3. We propose a flexible model that can detect the linear and a non-linear relationship between the predictors and the studied time series.
4. We improve the forecasts on the realized volatility data set of Christiansen et al. [2012].

The paper proceeds as follows. In Section 2 we briefly describe the proposed model and in Section 3 we present analytically our Bayesian computational strategy both for the model with a fixed number of predictors and for that with an unknown number of predictors. Then, in Section 4, we briefly present the forecasting criteria we used to assess the predictive ability of our method. A case study with model uncertainty is presented in Section 5. Next, we apply our methodology to monthly realized volatility data and present a comparative

analysis for this particular dataset in Section 6. Finally, Section 7 concludes the paper.

2 The Non-Homogeneous Pólya-Gamma Hidden Markov Model

In this section, we present the proposed Non-Homogeneous Pólya-Gamma hidden Markov model (NHPG) for modeling univariate time series. Consider an observed random process $\{Y_t\}$ and a hidden underlying process $\{Z_t\}$ which is a two-state non-homogeneous discrete-time Markov chain that determines the states of the observed process. Let y_t and z_t be the realizations of the observed random process $\{Y_t\}$ and of the hidden process $\{Z_t\}$, respectively. We assume that at time t , $t = 1, \dots, T$, y_t depends on the current state z_t and not on the previous states. Consider also a set of $r - 1$ available predictors $\{X_t\}$ with realization $x_t = (1, x_{1t}, \dots, x_{r-1t})$ at time t . A subset of the predictors $X_t^{(1)} \subseteq \{X_t\}$ of length $r_1 - 1$ is used in the regression model for the observed process and a subset $X_t^{(2)} \subseteq \{X_t\}$ of length $r_2 - 1$ is used to model/describe the dynamics of the time-varying transition probabilities. Thus, we allow the covariates to affect the observed process $\{Y_t\}$ in a non-linearly.

The observed random process $\{Y_t\}$ can be written in the form

$$Y_t = g(Z_t) + \epsilon_t,$$

where $g(Z_t) = X_{t-1}^{(1)} B_{Z_t}$ is a linear function, $B_{Z_t} = (b_{0Z_t}, b_{1Z_t}, \dots, b_{r_1-1Z_t})'$ are the regression coefficients and $\epsilon_t \sim \mathcal{N}(0, \sigma_{Z_t}^2)$. We use $\mathcal{N}(\mu, \sigma^2)$ to denote the normal distribution with mean μ and variance σ^2 . In a less formal way, if s represents the hidden states, the observed series given the unobserved process has the form

$$Y_t | Z_t = s \sim \mathcal{N}(X_{t-1}^{(1)} B_s, \sigma_s^2), \quad s = 1, 2.$$

The dynamics of the unobserved process $\{Z_t\}$ can be described by the time-varying transition probabilities, which depend on the predictors $X_t^{(2)}$ and are given by the following relationship

$$P(Z_{t+1} = j | Z_t = i) = p_{ij}^{(t)} = \frac{\exp(x_t^{(2)} \beta_{ij})}{\sum_{j=1}^2 \exp(x_t^{(2)} \beta_{ij})}, \quad i, j = 1, 2,$$

where $\beta_{ij} = (\beta_{0,ij}, \beta_{1,ij}, \dots, \beta_{r_2-1,ij})'$ is the vector of the logistic regression coefficients to be estimated. Note that for identifiability reasons, we adopt the convention of setting, for each row of the transition matrix, one of the β_{ij} to be a vector of zeros. Without loss of generality, we set $\beta_{ij} = \beta_{ji} = \mathbf{0}$ for $i, j = 1, 2, i \neq j$. Hence, the for $\beta_i = \beta_{ii}$, $i = 1, 2$ probabilities can be written in a simpler form

$$p_{ii}^{(t)} = \frac{\exp(x_t^{(2)} \beta_i)}{1 + \exp(x_t^{(2)} \beta_i)} \text{ and } p_{ij}^{(t)} = 1 - p_{ii}^{(t)}, \quad i, j = 1, 2, \quad i \neq j.$$

The unknown quantities of the NHPG are $\{\theta_s = (B_s, \sigma_s^2), \beta_s, s = 1, 2\}$, i.e., the parameters in the mean predictive regression equation and the parameters in

the logistic regression equation for the transition probabilities of the unobserved process $\{Z_t\}$, $t = 1, \dots, T$. Our model and the methods developed in this paper can be easily generalized into an m -state NHHMM, where the rows of the transition matrix are modeled by multinomial logistic regressions.

3 Bayesian Inference and Computational Strategy

This section presents the Bayesian approach to inference for the NHPG model. The key steps in our proposed framework are the following. First, for a given Hidden Markov model with time-varying probabilities, we construct a Markov chain which has as stationary distribution the posterior distribution of the model parameters. Simulation of this Markov chain provides, after some burn in period and adequately many iterations, samples from the posterior distribution of interest; see, for details, Besag et al. [1995]. Second, for a given set of competing models each including a different set of predictors in the mean regression and/or in the transition probabilities equation, we base our inference about the models on their posterior probabilities. This improves over approach which considers the models separately and chooses the best model via significance tests or via model selection criteria.

3.1 The MCMC Sampling Scheme

We present a summary of the MCMC algorithm that we have constructed for joint inference on model specification and model parameters, before we give the detailed explanation of these steps.

1. Start with initial values of $\beta, \theta = (B, \sigma^2)$.
2. Calculate the probabilities of the time-varying transition matrix.
3. Run a Scaled Forward-Backward (Scott [2002]) algorithm to simulate the hidden states given the parameters of the model.
4. Simulate the parameters of the regression model for the mean via a Gibbs sampler method.
5. Simulate the coefficients β using the Pólya-Gamma representation by Polson et al. [2013].
6. Use a double reversible jump algorithm to update the set of covariates that affect the transition matrix and those that affect the mean regression model.
7. Make one-step-ahead predictions conditional on the simulated unknown quantities.

Repeat steps 3-6 until convergence and then repeat steps 3-7.

3.2 Inference for fixed sets of predictors

Below we provide detailed guidelines on how to estimate the parameters of a given NHHMM, i.e., for fixed sets of predictors used in the mean equation and the transition probabilities $X^{(1)}$ and $X^{(2)}$, respectively. In the MCMC algorithm we update in turn (i) the latent variables z^T given the current values of the model parameters by using the scaled Forward-Backward algorithm (Scott [2002]) (ii) the logistic regression coefficients by adopting the auxiliary variables method of Polson et al. [2013] given the sequence of states z^T , and (iii) the mean regression coefficients conditional on z^T by using the Gibbs sampling algorithm.

Let $y^T = (y_1, \dots, y_T)$ be the history of the observed process, $z^T = (z_1, \dots, z_T)$ the sequence of states up to time T , and let $f_s(\cdot)$ denote the normal probability density function of $Y_t | Z_t = s$, $s = 1, 2$ and $\pi_1(z_1)$ the initial distribution of Z_1 . The joint likelihood function of the data, y^T , and the sequence of states, z^T , is given by

$$\begin{aligned} \mathcal{L}(\theta, \beta) &= \pi(y^T, z^T | X, \theta, \beta) = \pi(y^T | z^T, X, \theta, \beta) \pi(z^T | X, \theta, \beta) \\ &= \pi_1(z_1) f_{z_1}(y_1) \prod_{t=2}^T p_{z_{t-1} z_t}^{(t-1)} f_{z_t}(y_t) \\ &= \prod_{i=1}^2 \prod_{j=1}^2 \left[\prod_{t: z_t=j} p_{ij}^{(t-1)} \right] \left(\frac{1}{2\pi\sigma_j^2} \right)^{N_j/2} \exp \left\{ -\frac{1}{2\sigma_j^2} (Y_j - X_j^{(1)'} B_j)' (Y_j - X_j^{(1)'} B_j) \right\}. \end{aligned}$$

We use the notation N_s , $s = 1, 2$ for the number of times the chain was in state s , that is $N_s = \sum_{t=1}^T I(Z_t = s)$, with I the indicator function. If a prior distribution $\pi(\theta, \beta)$ is specified for the model parameters, then inference on all the unknown quantities in the model is based on their joint posterior distribution $\pi(\theta, \beta, z^T | y^T) \propto \pi(\theta, \beta) \pi(y^T, z^T | \theta, \beta)$.

For the parameters in the mean predictive regression equation, we use conjugate prior distributions, i.e., $\sigma_s^2 \sim \mathcal{IG}(p, q)$, $B_s | \sigma_s^2 \sim \mathcal{N}(L_0, \sigma_s^2 V_0)$, $s = 1, 2$, where \mathcal{IG} denotes the Inverted-Gamma distribution. After some straightforward algebra we derive the marginal posterior distribution for the state specific parameters σ_s and conditional posterior distribution for B_s ,

$$\begin{aligned} \sigma_s^2 | y^T, z^T &\sim \mathcal{IG} \left(p + \frac{n_s}{2}, q + \frac{1}{2} (L_0' V_0^{-1} L_0 + Y_s' Y_s - L_s' V_s^{-1} L_s) \right), \\ B_s | \sigma_s^2, z^T, y^T &\sim \mathcal{N}(L_s, \sigma_s^2 V_s), \end{aligned}$$

with $V_s = (V_0^{-1} + X_s^{(1)'} X_s^{(1)})^{-1}$ and $L_s = V_s (V_0^{-1} L_0 + X_s^{(1)'} Y_s)$.

To make inference about the logistic regression coefficients we use the auxiliary variables method of Polson et al. [2013] as described in Subsection 3.2.1. Given an auxiliary variable ω_s , a conjugate prior for the logistic regression coefficients β_s , $s = 1, 2$ is multivariate normal distribution $\mathcal{N}(m_{\beta_s}, V_{\beta_s})$.

3.2.1 Simulation of the logistic regression coefficients

In a two-state NHHMM we can model the two diagonal elements of probability transition matrix by linking them to the set of covariates using a logistic link. We use the data augmentation scheme of Polson et al. [2013] since, as shown in

their work, the estimation of logistic regression coefficients using this scheme is superior, in terms of convergence and mixing, among all the competing methods.

Given the unobserved (latent) data $z^T = (z_1, \dots, z_T)$ we define, for $t = 1 \dots, T - 1$, the quantity $\tilde{Z}_{t+1}^s = I[Z_{t+1} = Z_t = s]$. In words, the quantity $\sum_t \tilde{Z}_{t+1}^s$ is the number of times that the chain was at the same state for two consecutive time periods. Then,

$$p\left(\tilde{Z}_{t+1}^s = 1 \mid x_t^{(2)}\right) = p_{ss}^t = \frac{\exp\left(x_t^{(2)}\beta_s\right)}{1 + \exp\left(x_t^{(2)}\beta_s\right)} \Leftrightarrow \text{logit}(p_{ss}^t) = x_t^{(2)}\beta_s, \quad s = 1, 2.$$

Polson et al. [2013] proved that binomial likelihoods –thus Bernoulli likelihoods in our simpler case– parametrized by log odds can be represented as mixtures of Gaussian distributions with respect to the Pólya-Gamma distribution. The main result of Polson et al. [2013] is that letting $p(\omega)$ be the density of a latent variable ω with $\omega \sim \mathcal{PG}(b, 0)$, for $b > 0$, the following integral identity holds for all $a \in \mathbb{R}$:

$$\frac{\exp(\psi)^a}{(1 + \exp(\psi))^b} = 2^{-b} \exp(k\psi) \int_0^\infty \exp(-\omega\psi^2/2) p(\omega) d\omega,$$

where $k = a - b/2$. Furthermore, the conditional distribution of $\omega \mid \psi$ is also Pólya-Gamma, $\mathcal{PG}(b, \psi)$. Using the previous result and setting $\Omega_s = \text{diag}\{\omega_{1,s}, \dots, \omega_{N_s,s}\}$ as a set of latent variables, the likelihood for each state $s = 1, 2$ is

$$\begin{aligned} \mathcal{L}(\beta_s, \omega_s) &= \prod_{t=1}^{N_s} \left\{ \frac{\exp\left(x_t^{(2)}\beta_s\right)}{1 + \exp\left(x_t^{(2)}\beta_s\right)} \right\}^{\tilde{z}_t} \left\{ \frac{1}{1 + \exp\left(x_t^{(2)}\beta_s\right)} \right\}^{1-\tilde{z}_t} \\ &\propto \prod_{t=1}^{N_s} \exp\left(k_t x_t^{(2)}\beta_s\right) \int_0^\infty \exp\left\{-\omega_{t,s} \left(x_t^{(2)}\beta_s\right)^2 / 2\right\} p(\omega_{t,s}) d\omega_{t,s}. \end{aligned}$$

Conditioning on Ω_s , one can derive the expression

$$\pi(\beta \mid z^t, \omega_s) \propto \pi(\beta) \prod_{t=1}^{N_s} \exp\left\{-\frac{\omega_{t,s}}{2} \left(\left(x_t^{(2)}\beta_s\right)^2 - \frac{2k_t x_t^{(2)}\beta_s}{\omega_{t,s}} + \frac{k_t^2}{\omega_{t,s}^2} \right)\right\}.$$

Assuming as prior distributions $\omega \sim \mathcal{PG}(b, 0)$ and $\beta \sim \mathcal{N}(m_{\beta_0}, V_{\beta_0})$, simulation from the posterior distribution can be done iteratively in two steps:

$$\omega_{t,s} \mid \tilde{z}_t \sim \mathcal{PG}\left(1, x_t^{(2)}\beta_s\right), \quad t = 1 : N_s, \quad s = 1, 2,$$

$$\beta_s \mid \tilde{Z}, \Omega_s \sim \mathcal{N}(m_{\omega_s}, V_{\omega_s}),$$

$$V_{\omega_s} = \left(X^{(2)'}\Omega_s X^{(2)} + V_{\beta_0}^{-1}\right)^{-1} \quad \text{and} \quad m_{\omega_s} = V_{\omega_s} \left(X^{(2)'}k + V_{\beta_0}^{-1}m_{\beta_0}\right),$$

where \mathcal{PG} denotes the Pólya-Gamma distribution and $k = (\tilde{z}_1 - 1/2, \dots, \tilde{z}_{N_s} - 1/2)$.

3.3 Inference under model uncertainty

Here, we consider the full model comparison problem where the uncertainty about which predictors should be included in the mean regression model is taken

into account together with the uncertainty about the predictors that should be included in the transition probability equation. In this setting, the RJMCMC does not need any tuning and hence it can be used as a black box.

Suppose that a prior $\pi(k)$ is specified over k models (M_1, M_2, \dots, M_k) in a countable set \mathcal{K} and for each k we are given a prior distribution $\pi(\theta_k | k)$ along with a likelihood $\mathcal{L}(y | \theta_k, k)$ for data y . The joint prior for θ and k is $\pi(k, \theta_k) = \pi(\theta_k | k)\pi(k)$. When a move of type m from $\tilde{x} = (k, \theta_k)$ to $\tilde{x}^* = (k^*, \theta_{k^*}^*)$ is proposed from the proposal distribution g and if $j_m(\tilde{x})$ denotes the probability that move m is attempted at state \tilde{x} and $j_{m^*}(\tilde{x}^*)$ the probability of the reverse move, we accept the proposed move with probability $\alpha_m(\tilde{x}, \tilde{x}^*) = \min\{1, A_m(\tilde{x}, \tilde{x}^*)\}$ where

$$A_m(\tilde{x}, \tilde{x}^*) = \frac{\mathcal{L}(y^T | \tilde{x}^*) \pi(\theta^* | k^*) \pi(k^*) j_{m^*}(\tilde{x}^*) g'_m(u^* | \tilde{x}^*, k^*) \left| \frac{\partial(\theta_{k^*}^*, u^*)}{\partial(\theta_k, u)} \right|}{\mathcal{L}(y^T | \tilde{x}) \pi(\theta | k) \pi(k) j_m(\tilde{x}) g_m(u | \tilde{x}, k^*)},$$

and $\left| \frac{\partial(\theta_{k^*}^*, u^*)}{\partial(\theta_k, u)} \right|$ is the Jacobian of the transformation.

In each step, we choose to add or remove one covariate with probability 0.5 and then we randomly choose which covariate we will add/remove. We propose a new value for the mean equation coefficients B^* or for the regression equation coefficients β^* from the full conditional posterior density, conditionally on the other coefficients. Thus, the Jacobian of the transformation will be equal to unity. To be more specific, if we want to update the covariates in the mean equation, the proposal distribution g' is just the product of the two conditional posterior distributions. With some straightforward matrix algebra, the acceptance probability for the mean equation is $\alpha_B = \min\{1, A_B\}$ and the acceptance probability for the transition matrix is $\alpha_\beta = \min\{1, A_\beta\}$ where

$$A_B = \frac{j_{m^*}(k^*)}{j_m(k)} \prod_{s=1}^2 \frac{|V_s^*|^{1/2} |V_{0s}|^{1/2}}{|V_s|^{1/2} |V_{0s}^*|^{1/2}} \times \exp \left\{ -\frac{1}{2\sigma_s^2} \left(L_{0s}' V_{0s}^{*-1} L_{0s}^* - L_s^{*'} V_s^{*-1} L_s^* - L_{0s}' V_{0s}^{-1} L_{0s} + L_s' V_s^{-1} L_s \right) \right\}$$

and

$$A_\beta = \frac{j_{m^*}(k^*)}{j_m(k)} \prod_{s=1}^2 \frac{|V_{\omega s}^*|^{1/2} |V_{\beta_{0s}}|^{1/2}}{|V_{\omega s}|^{1/2} |V_{\beta_{0s}^*}|^{1/2}} \times \exp \left\{ -\frac{1}{2\sigma_s^2} \left(L_{\beta_{0s}}^{*'} V_{\beta_{0s}}^{*-1} L_{\beta_{0s}}^* - L_{\omega s}^{*'} V_{\omega s}^{*-1} L_{\omega s}^* - L_{\beta_{0s}}' V_{\beta_{0s}}^{-1} L_{\beta_{0s}} + L_{\omega s}' V_{\omega s}^{-1} L_{\omega s} \right) \right\}.$$

4 Bayesian Forecasting and Scoring rules

4.1 One-step-ahead predictions

The proposed modeling and inferential approach is used for forecasting. The posterior predictive density can not be found in closed form, but can be evaluated numerically. Given model M , the predictive distribution of y_{T+1} is

$$f_p(y_{T+1} | y^T) = \int f(y_{T+1} | y^T, z^T, M, \beta_M, \theta_M) \pi(\beta_M, \theta_M | y^T) d\beta_M d\theta_M,$$

where $f(y_{T+1} | y^T, z^T, \beta_M, \theta_M) = \sum_{s=1}^2 P(Z_{T+1} = s | Z_T = z_T) f_s(y_{T+1})$. In practice, we follow an iterative procedure within our MCMC algorithm to draw a sample from the posterior predictive distribution. At the r -th iteration of our algorithm, the algorithm chooses model M_r . Furthermore, the hidden states and the unknown parameters $\beta_{M_r}, \theta_{M_r}$ are simulated as described in Subsection 3.2. To make an one-step-ahead prediction (i.e., simulate y_{T+1}), we first simulate the hidden state for time $T+1$ from the discrete distribution based on the transition probabilities $P(Z_{T+1}^{(r)} = s | Z_T = z_T^{(r)})$, $s = 1, 2$, and then, conditional on the hidden state, we draw a value y_{T+1}^r from $\mathcal{N}(X_T^{(1)} B_{s, M_r}, \sigma_{s, M_r}^2)$, $s = 1, 2$. Given y_{T+l} , Z_{T+l} and the covariates X_{T+l-1} , for $l=1, \dots, L$, we may also update the transition matrix P^{T+l} , simulate Z_{T+l+1} and finally simulate the prediction y_{T+l+1} from its respective predictive distribution. In this way, in each iteration we obtain sequentially a sample of L one-step-subsequent predictions.

4.2 Forecasting criteria

In forecasting experiments one needs to evaluate the quality of the obtained forecasts. Moreover, predictive accuracy is valued not only for its own sake but also for comparing different models (Gelman et al. [2014]). The accuracy of the model's forecasts can be used as a metric to evaluate the model's performance (Geweke and Whiteman [2006]).

Advances in numerical integration via MCMC algorithms made probabilistic forecasts possible, which are in most cases preferable. Besides, having the posterior predictive distribution, one can obtain point forecasts using suitable scoring functions (Gneiting [2011]). Scoring rules provide summary measures for the evaluation of probabilistic forecasts by assigning a numerical score based on the forecast and on the event or value that it materializes. We refer to Gneiting and Raftery [2007], Machete [2013] for a review on the theory and properties of scoring rules. A widely used, extensively studied and quite powerful criterion is the Logarithmic Score (LS), see Gelman et al. [2014], Gschlößl and Czado [2007] and references therein. It is based on the logarithm of the posterior predictive density evaluated at the observed value. However, LS lacks robustness as it involves harsh penalty for low probability events and thus is sensitive to extreme cases (Boero et al. [2011]). Besides, comparing the entropies of the forecasts, Machete [2013] showed that LS prefers the forecast density that is less informative. In the same spirit Gneiting and Raftery [2007] noticed that measures which are not sensitive to distance give no credit for assigning high probabilities to values near but not identical to the one materializing. Sensitivity to distance seems desirable when predictive distributions tend to be multimodal, which is the case of our model. To deal with this, one could calculate the Continuous Ranked Probability Score (CRPS) which is based on the cumulative predictive distribution, see Appendix 7.3 for the definition. Recently, Boero et al. [2011] argued that when density forecasts are collected in histogram format, then the ranked probability score has advantages over the other studied scoring rules.

To compute the CRPS for the forecast y_l we use the identity of Székely and Rizzo [2005],

$$CRPS(F_p, y_l) = \frac{1}{2} E_F |Y - Y'| - E_F |Y - y_l|,$$

were Y, Y' are independent copies of a random variable with the posterior predictive distribution function F_p (see also Gschlößl and Czado [2007]).

Finally, for the sake of completeness we give the definition of the Mean Square Forecast Error, $MSFE = \frac{1}{L} \sum_{i=T+1}^{T+L} (y_i - \hat{y}_i)^2$ and the Mean Absolute Forecast Error, $MAFE = \frac{1}{L} \sum_{i=T+1}^{T+L} |y_i - \hat{y}_i|$. Based on the MCMC sample, the computed point forecasting criteria is just the average score of the aforementioned criteria for all the sample values.

5 Simulation Study

We have conducted a series of simulation experiments to assess the performance of the proposed approach in terms of inference, model selection and predictive ability. We have scrutinized our algorithms, using different sample sizes and assigning various values to the parameters. Our experiments are carried out using MATLAB 2017b on a Windows 10 system with 32GB of RAM and Intel Core i7 8-core processor.

To assess the inferential ability of our model we benchmarked our model with the model proposed by Meligkotsidou and Dellaportas [2011](M&D) (without model uncertainty) and with a Homogeneous Hidden Markov Model (HHMM), see Appendix 7.2. The NHPG is at least as good as the M&D model –in forecasting ability, efficiency and sample quality–, but has less computational complexity and hence the algorithm is faster. In Table 6 we present a summary of the case study of the fixed model. We report the sample quality, convergence and mixing criteria that we use in Appendix 7.4.

The superiority of our model among all the benchmark models is shown in the case study with model uncertainty, Section 5.1. We compare NHPG with existing variable selection schemes, such as the model of Meligkotsidou and Dellaportas [2011], a HHMM with a variable selection using reversible jump and a model using the spike and slab prior for variable selection as studied in Narisetty and He [2014], referred to as BAyesian Shrinking And Diffusing priors (BASAD), see Appendix 7.1.

The data were generated either from a HHMM or from a NHHMM with covariates simulated from independent normal distributions. We found that the mean equation parameters converge rapidly whereas the logistic regression coefficients needed some burn in period to converge. The hidden chain Z^T was well estimated. For each iteration we kept a replication of the hidden chain and compared it with the real simulated hidden chain, using a 1-0 loss function (see Figure 1).

Furthermore, to test the predictive ability of our model we kept L out-of-sample observations. We calculated, for all the competing models, the CRPS, the MSFE and the MAFE. However, we note that due to the large out-of-sample period we only report the averages (for all the draws) of the aforementioned forecasting criteria. In all our experiments we found that our model outperforms the all competing models in forecasting the observed process.

5.1 Case study: The NHPG with model uncertainty

The main applications in which our algorithm considerably improves over the benchmark models –M&D, BASAD– is when there exists model uncertainty.

We compared results from the three models (NHPG, M&D, BASAD) subject to model uncertainty. We simulated data from a NHHMM of size $T = 1200$. From a common pool of independently normally distributed covariates $X = \{1, X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9\}$ with means $\mu_x = [4, 3, -2, -5, 2.5, -4, -6, 7, 1]$ and variances $\sigma_x^2 = [1, 1, 0.5, 1, 1, 1, 0.5, 2, 1.5]$ we used 3 covariates $X^{(1)} = \{1, X_1, X_2, X_3\}$ affecting the mean equation and $X^{(2)} = \{1, X_1, X_2, X_4\}$ the transition matrix. The mean equation parameters were $B_1 = [2, -0.3, 2, 2]'$, $\sigma_1^2 = 1.5$ and $B_2 = [1, 3, 4, 3]'$, $\sigma_2^2 = 0.8$ whereas the logistic regression coefficients where $\beta_1 = [1.5, 1, 2, 3]'$ and $\beta_2 = [3, -2.5, 4, 1]'$, for the two states respectively. We kept $L = 96$ out-of-sample observations and we computed a sequence of one-step-ahead forecasts of the real observed process. Our results are based on a sample of 15000 predictions after discarding an initial burn in period of 10000 iterations. In this forecasting analysis we also included the HHMM with variable selection, in the mean regression model. We used non-informative priors for the unknown parameters σ_s^2, B_s, β_s , $s = 1, 2$, that is $\sigma_s^2 \sim \mathcal{IG}(0.1, 0.1)$, $B_s | \sigma_s^2 \sim \mathcal{N}(0, 100\sigma_s^2 \times I)$ and finally $\beta_s \sim \mathcal{N}(0, 100 \times I)$. Also, as suggested by Narisetty and He [2014] and Narisetty et al. [2018] we used as hyperparameters values $\tau_{0B,n}^2 = \frac{\hat{\sigma}^2}{10T}$, $\tau_{1B,n}^2 = \hat{\sigma}^2 \max\left(\frac{r_1^{2.1}}{100T}, \log(T)\right)$ and $\tau_{0\beta,n}^2 = \frac{1}{T}$, $\tau_{1\beta,n}^2 = \max\left(\frac{r_2^{2.1}}{100T}, 1\right)$ with $\hat{\sigma}^2$ be the estimated variance of the data Y . In Table 1 we report here the runtimes for every methodology. In terms of time efficiency, the NHPG is slower than BASAD but more than two times faster than M&D.

Runtimes per 1000 iterations			
	NHPG	M&D	BASAD
sec	310	693	160

Table 1: Summary of runtimes (in seconds per 1000 iterations). NHPG denotes the proposed methodology. The NHPG is slower than BASAD but more than two times faster than M&D.

We found that our approach was able to identify—as the most probable or the median probability model—the correct data generating process. In contrast with the proposed methodology, in Table 2, we show that for this specific case study the median probability model of the competing methodologies was not the correct—true—one. As far as forecasting is concerned, as shown in Table 3, the NHHMM had the best performance according to all forecasting criteria. Supplementary to Table 3 are the plots in Figure 2. The empirical posterior predictive distributions of the four competing models, for three randomly selected out-of-sample periods, are plotted in the same plot along with the actual value. Figure 3 gives a graphical indication of the superiority of our approach in terms of forecasting.

Finally, for each MCMC iteration we kept a replicated chain of the hidden process and we compared it with the true simulated chain. Using the 0-1 Loss function we computed the average number of misestimated states in each chain. All three approaches had similar performance according to this criterion. Specifically, from the chain with 1104 hidden states, NHHMM failed to recognize 2 states per iteration, M&D 3 states per iteration and BASAD methodology one state per iteration. A virtualization of the estimation of the hidden process

against the true hidden process is shown in Figure 1. This figure presents the thinned version (1:2 observations) of the simulated time series along with the true hidden process and an estimate of the hidden process using the proposed methodology.

Note that we present this specific example as a case where the M&D and BASAD models do not perform well in terms of variable selection. However, there existed some cases in our extensive simulation study that the aforementioned methodologies were able to identify the true generating process, whereas our method does identify the true model in all cases (results not shown for reasons of space limitation).

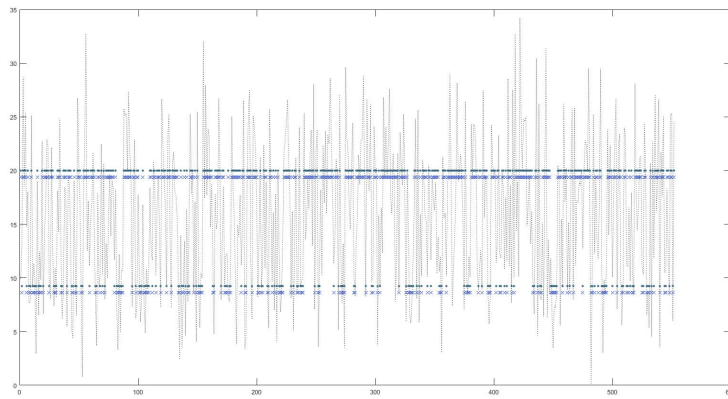


Figure 1: Observed process (black dotted line) and hidden process. The true hidden states are marked with blue x and the realized simulated states are marked with blue dots. The true hidden process is well estimated by the estimated hidden process.

	True Model	Median probability model		
		NHPG	M&D	BASAD
M E	X_1, X_2, X_3	X_1, X_2, X_3	X_1, X_2, X_3	X_2, X_4, X_9
T M	X_1, X_2, X_4	X_1, X_2, X_4	$X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_9$	X_1, X_2, X_4

Table 2: Median probability models using the three aforementioned methodologies. The first row (M E) points to the covariates used in the Mean Equation and the second row (T M) points to the covariates of the Transition Matrix. NHPG, M&D, BASAD denote the proposed methodology, the methodology proposed by Meligkotsidou and Dellaportas [2011] and the methodology proposed by Narisetty and He [2014], respectively. The proposed methodology identifies the true data generating process.

	Forecasting Criteria			
	NHPG	M&D	HHMM	BASAD
E(CRPS)	-1.9526	-3.6829	-2.6597	-2.4952
MAFE	3.9271	4.3911	5.4101	5.0611
MSFE	32.8856	39.4958	53.8280	49.1432

Table 3: Summary of the results of the variable selection methods on simulated data. NHPG denotes the proposed methodology, M&D the methodology proposed by Meligkotsidou and Dellaportas [2011], HHMM the Homogeneous model and BASAD the methodology proposed by Narisetty and He [2014]. The best performance for each criterion is denoted by bold values.

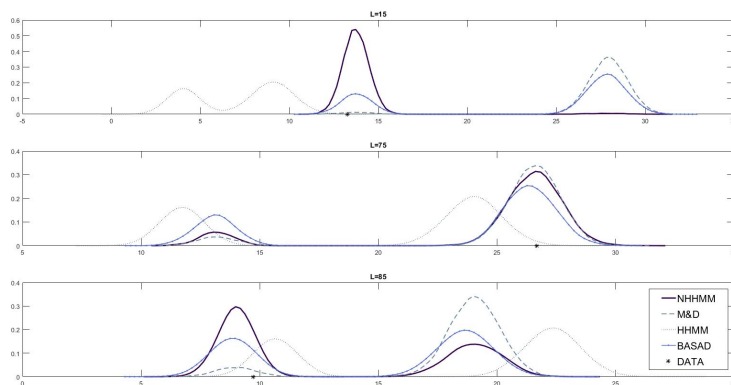


Figure 2: The plots show the empirical posterior predictive distributions based on a normal kernel function for three randomly selected out-of-sample forecasts, $L = 15, 75, 85$, using the NHHMM (black continuous line), M&D (gray dashed line), the HHMM (dotted line) and the BASAD (dotted-dashed blue line). Actual out-of-sample values are marked with asterisks. These plots visualize the advantage of using the proposed methodology.

6 Empirical Application: Realized volatility data

Financial volatility has been extensively studied in the literature due to its crucial role in various financial fields, such as asset pricing, risk management, investment and asset allocation among others, see Gospodinov et al. [2006]. Several studies have considered predicting realized stock volatility using various financial and/or economic predictors (see for example, Mitnik et al. [2015], Meligkotsidou et al. [2018], Christiansen et al. [2012], Paye [2012]). We use the NHPG to assess the predictive ability of 13 financial variables in forecasting future volatility.

6.1 The data

We applied our method to realized stock market volatility data, as described in detail by Christiansen et al. [2012]. Specifically, we use the 'long' sample of the U.S. equity market, S&P500. The realized volatility is the squared root of the realized variance for asset class i in month t expressed as the sum of squared intra-period (daily) returns

$$RV_{i,t} = \sqrt{\sum_{\tau=1}^{u_t} r_{i,t,\tau}^2}, \quad t = 1 \dots, T,$$

where $r_{i,t,\tau}$ is the τ th daily continuously compounded return of month t for asset i with u_t the trading days. Thus $\sum_{\tau=1}^{u_t} r_{i,t,\tau}^2$ is the realized variance for asset class i in month t . The distribution of the realized daily variances are highly non-normal and skewed to the right, but the logarithms of the realized variances are approximately normal and thus they have better behavior (see Andersen et al. [2003]). Hence, in the following analysis we study the natural logarithm of the realized volatility series, $\ln(RV_{i,t})$ $t = 1 \dots, T$.

The data are observed in a monthly basis, from December 1926 to December 2015—note that we use a five-years extended dataset compared to the dataset of Christiansen et al. [2012]. The out-of-sample forecast evaluation period is set to eight years, i.e., 96 observations from December 2007 until December 2015. We had a burn in period of 60000 iterations and we generated 40000 MCMC draws. We used non-informative priors for the unknown parameters $\sigma_s^2, B_s, \beta_s, s = 1, 2$, that is $\sigma_s^2 \sim \mathcal{IG}(0.15, 0.15)$, $B_s | \sigma_s^2 \sim \mathcal{N}(0, 100\sigma_s^2 \times I)$ and finally $\beta_s \sim \mathcal{N}(0, 100 \times I)$.

Following Christiansen et al. [2012] and Meligkotsidou et al. [2018] we took into account 13 macroeconomic and financial standardized predictive covariates. Particularly, from a list of equity market variables and risk factors, we considered dividend price ratio (DP) and earnings price ratio (EP) (Welch and Goyal [2008]), lagged equity market returns (MKT), in order to capture the leverage effect, that is the asymmetric response of volatility to positive and negative returns (Nelson [1991]). We also used the risk factors of Fama and French [1993], that is, the size factor (SMB), value factor (HML) and a short-term reversal factor (STR). From the set of interest rates, spreads and bond market factors we included the treasury bill rate (TBL), i.e., the interest rate on a three-month Treasury bill, the long-term return (LTR) on long-term government bonds, the term spread (TMS), i.e., the difference between the long-term yield and Treasury bill rate, the relative T-bill rate (RTB) as the difference between, T-bill rate and its 12-month moving average and the relative bond rate (RBR), as the difference between LTR and its 12 month moving average (Welch and Goyal [2008]). To proxy for weighted credit risk we also used the default spread (DEF) defined as the yield spread between BAA and AAA rated bonds. Lastly, we considered the macroeconomic variable, inflation rate (INF), the monthly growth rate of CPI. The strong contemporaneous relation between the volatility and the business conditions implies that lagged volatility plays important role in forecasting (see Paye [2012], Baillie et al. [2019]). Besides, quoting Christiansen et al. [2012], we include at least one autoregressive term, “since volatility is fairly persistent it is important to include autoregressive terms in the predictive regression to investigate whether there is additional predictive content of the macroeconomic

and financial variables that goes beyond the information contained in lagged volatility”. We ran a series of experiments for this data. Specifically we performed our analysis using the predictors described and then we repeated the analysis using the predictors plus autoregressive terms (AR) of lag 1, 2 and lag 3.

6.2 Results

Our in-sample analysis revealed some interesting findings. Based on the posterior probabilities of inclusion we see that if we do not include any AR terms in the pool of the predictors, then the median NHPG model has three predictors affecting only the mean equation of the series. Thus, based on the median probability model, the realized volatility series is considered to be a homogeneous hidden Markov model. The probabilities of staying at the same state are in this case high, concluding that the states are highly persistent. When we add the AR(1) term, the included predictors in the median probability model are also three but they affect the series both linearly and not linearly. We observe that an autoregressive term explains a big fraction of the variance of the realized volatility. Adding more AR terms (of lag 2 and lag 3), the median probability model remained almost the same as in the case of the model with one autoregressive term. Furthermore, in our out-of-sample analysis, we did not encounter any significant improvement in the forecasting ability of the models with AR(2) and AR(3) terms. We note that this result confirms the findings of Christiansen et al. [2012], who also used only one AR term of lag 1 in their analysis. Hereafter, even though –based on the CRPS– the model with the best performance was the one with the AR(1) term, we present the results of both the model with no autoregressive terms (NHPG₀) and the model with one autoregressive term (NHPG₁), for the sake of completeness. Also, we compare our results with those of the model of Christiansen et al. [2012], CSS hereafter, who consider a Bayesian Model Averaging (BMA) approach using a linear model with one autoregressive term and a stochastic model search algorithm (MC³). We note that for this model we allowed for much longer burn in period, as suggested by the authors, of 500000 draws. Moreover we included in our comparative analysis the model of Meligkotsidou and Dellaportas [2011] (M&D) and the Homogeneous Hidden Markov Model (HHMM) with one autoregressive term.

Figure 3 shows a plot of the realized volatility data (blue line) together with the probabilities of staying at the same state (e.g. if at time t we are at state 1 then the red dot at time t shows the probability of staying at state 1 at time $t + 1$, that is the transition probability p_{11}^t). The shaded bars represent the time period that the chain was in state 1, based on the smoothed probabilities of being above 0.5 for NHPG₁. Furthermore, in Figure 4 we present a thinned (1:5) in-sample realization of the observed process inferred by our algorithm, i.e., using the in-sample estimations of the parameters and the states to reproduce the realized volatility series, along with the real data. The in-sample evaluation of the observed process gives an indication of the good performance of the estimation procedure.

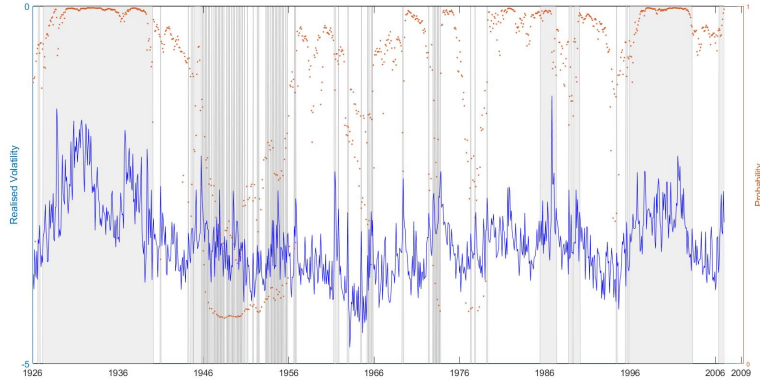


Figure 3: Time series of the monthly realized volatility of the Standard & Poor (*S&P*) 500 index (in logarithmic scale) for the period 1926-2007, using the $NHPG_1$. Red dots are the posterior mean probabilities of staying at the same state. We calculated the smoothed probability of staying in state 1. If the probability is above 0.5 we mark this event with gray-shaded bar.

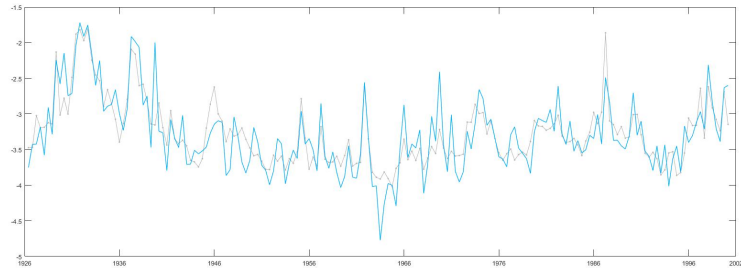


Figure 4: This figure presents a thinned realization of the observed process (1:5 observations) as calculated using our $NHPG_1$ (gray solid line). For comparison we also present the thinned time series of the studied variable.

6.2.1 Model Selection

Our model selection algorithm did not assign high probability to any specific model indicating that there exists model uncertainty. In Table 4, we summarize the posterior probabilities of inclusion for each predictor, both for the mean equation and for the transition matrix for the $NHPG_0$, $NHPG_1$, M&D and the posterior probabilities of inclusions for CSS and HHMM. Our methodology—when the autoregressive term of lag 1 was included—was not only able to identify which covariates affect the realized volatility series but also to decide how the covariates affect the series, i.e., linearly or non-linearly. The number of the predictors defining the median probability $NHPG_1$ has diminished to three (instead of thirteen). Specifically, we found that the MKT affects the series linearly, the

SMB affects the series non-linearly and the DEF both linearly and non linearly. The predictors that were included in NHPG_1 are in common with the predictors included in the HHMM. The linear autoregressive model of Christiansen et al. [2012] identifies four predictors with probability at least 0.5, three of them being the same with NHPG_1 and HHMM, that is the MKT, the DEF and the EP plus the predictor STR. However, the algorithm M&D includes all the predictors in the mean equation model, while it includes the predictors DP, MKT, TBL and DEF in the logistic regression for the transition probabilities.

Posterior probabilities of inclusion									
	NHPG_0		NHPG_1		M&D		HHM	CSS	
Covariates	M E	T M	M E	T M	M E	T M			
DP	0.01	0.04	0.08	0	1	0.77	0.3		0.38
EP	0.98	0.12	0.06	0.89	1	0.37	0.78		0.50
MKT	1	0.04	0.98	0.03	0.93	1	1		0.97
SMB	0	0.02	0	0.04	0.90	0.28	0		0.05
HML	0.01	0.03	0	0.02	0.92	0.09	0		0.06
STR	0	0.05	0	0.06	0.92	0.09	0		0.53
TBL	0	0.02	0	0.10	0.99	0.71	0.03		0.10
RTB	0	0.02	0	0.09	0.99	0.02	0		0.04
LTR	0.01	0.03	0	0.03	1	0.01	0		0.05
RBR	0	0.03	0	0.03	1	0.01	0		0.05
TMS	0.01	0.06	0.17	0.26	0.89	0.23	0		0.05
DEF	1	0.02	1	0.79	1	1	1		1
INF	0	0.02	0.11	0.03	0.92	0.47	0		0.04

Table 4: Posterior probabilities of inclusion for the competing models. Predictors with inclusion probability above 0.5 (median probability model) are marked with bold values. M E stands for Mean Equation and T M for Transition Matrix.

6.2.2 Forecasting

As far as forecasting is concerned, in Table 5, we report the values of the forecasting criteria that we used for all the competing models. We conclude that NHHMM_1 performs better than all the other models, since it has better scores in the probabilistic criterion mean Ranked Probability Score ($E(\text{CPRS})$) and also in the point forecasting criteria MAFE and MSFE.

	Forecasting Criteria				
	NHPG ₁	NHPG ₀	M&D	HHMM	CSS
E(CRPS)	-0.1971	-0.2175	-0.2191	-0.2118	-0.2238
MAFE	0.3821	0.4643	0.4172	0.4534	0.4787
MSFE	0.2467	0.3426	0.2678	0.3449	0.3813

Table 5: This table presents a summary of the results regarding the forecasting performance of the five competing models obtained from the log-realized volatility dataset. NHPG₀ and NHPG₁ denote the proposed methodology without autoregressive terms and with one autoregressive term respectively, M&D the methodology proposed by Meligkotsidou and Dellaportas [2011], HHMM the Homogeneous model with variable selection using a reversible jump step and CSS the model of Christiansen et al. [2012]. The best performance for each criterion is denoted by bold values.

7 Conclusions

In this paper we have considered inference on predictive Non-Homogeneous Hidden Markov models. We allowed different predictors affecting the mean equation and the time-varying transition probabilities. Transition probabilities were parametrized via a logistic link. Even though inference for logistic regression coefficients is considered to be an easy task, in the case of using logistic functions for modeling the transition probabilities of NHHMMs, many proposed algorithms face serious convergence issues. Applying the recently proposed Pólya-Gamma data augmentation scheme of Polson et al. [2013] we built a stable and accurate MCMC scheme to make inferences on the parameters of our model. Furthermore, we performed stochastic variable selection using a double reversible jump step in order to determine which covariates affect the series linearly and/or in a non-linear fashion. To summarize our approach, in each MCMC iteration we simulated the hidden states using the scaled Forward-Backward algorithm of Scott [2002], simulated the mean equation parameters using a Gibbs step, then we used the Pólya-Gamma augmentation scheme to simulate the logistic regression coefficients for the transition probabilities and finally we performed a double reversible jump step to choose the covariates that affect the mean equation and the transition probabilities. When convergence was achieved we then made one-step-look-ahead predictions.

To assess the performance of the proposed algorithm, as well as the predictive ability of our model and methods, we have tested our methodology quite extensively by conducting a number of simulation experiments. Our results have shown that our algorithm mixes and converges well and provides accurate estimates of the model’s parameters. Moreover, we showed that our model outperforms competing models/approaches used as benchmarks, such as the approach of Meligkotsidou and Dellaportas [2011], the BASAD model of Narisetty and He [2014] and the homogeneous hidden Markov model, in terms of variable selection and forecasting ability, according to the continuous ranked probability score, the mean absolute forecasting error and the mean square forecasting error. The proposed methodology was applied to a realized volatility dataset—analytically presented in Christiansen et al. [2012]—for predicting future observations and for predictor selection. We identified which predictors affect the

series linearly and/or non-linearly. The median probability model derived using our model/method identified three predictors, one affecting the analyzed series only linearly, one non-linearly and the remaining one affecting the series both linearly and non-linearly. In a comparison with the included variables of Christiansen et al. [2012] we found that the three identified predictors of our approach are in common with the three out of four predictors of their model, who however can not distinguish if the predictors affect the series linearly or non-linearly. Finally, we have reported that our model exhibits improved predictive ability. We believe that our empirical findings pave the way for applying complex Non-Homogeneous Hidden Markov models for predicting univariate and multivariate, financial and economic time series.

Appendix

7.1 Benchmark models

For the sake of completeness we provide the definitions of two of our benchmark models. In 5.1 we compare our findings with the BASAD model of Narisetty and He [2014]. In that paper the authors introduce shrinking and diffusing priors as a spike and slab priors model, with prior parameters depending on the sample size to achieve appropriate shrinkage. Narisetty and He [2014] work with orthogonal design matrices and use binary latent variables U_i to indicate if a covariate is active or not. In our setting, the BASAD model for the mean equation is defined as:

$$\begin{aligned}
Y_t \mid (X_{t-1}B_s, \sigma_s^2) &\sim \mathcal{N}(X_{t-1}B_s, \sigma_s^2), \quad s = 1, 2, \quad t = 1, \dots, T, \\
B_{k,s} \mid (\sigma_s^2, U_{k,s} = 0) &\sim \mathcal{N}(0, \sigma_s^2 \tau_{0B,n}^2), \quad B_{k,s} \mid (\sigma_s^2, U_{k,s} = 1) \sim \mathcal{N}(0, \sigma_s^2 \tau_{1B,n}^2), \quad k = 1, \dots, r, \\
P(U_{k,s} = 1) &= 1 - P(U_{k,s} = 0) = q_n, \quad k = 1, \dots, r, \\
&\text{and} \\
\sigma_s^2 &\sim \mathcal{IG}(\alpha_1, \alpha_2).
\end{aligned}$$

The transition probabilities are parametrized as:

$$\begin{aligned}
\tilde{Z}_{t+1}^s &\sim \text{Bin}\left(1, \frac{\exp(x_t \beta_s)}{1 + \exp(x_t \beta_s)}\right), \\
\beta_k^s \mid (U_{k,s} = 0) &\sim \mathcal{N}(0, \sigma_s^2 \tau_{0\beta,n}^2), \quad \beta_i^s \mid (U_{k,s} = 1) \sim \mathcal{N}(0, \sigma_s^2 \tau_{1\beta,n}^2), \quad k = 1, \dots, r, \\
\omega_s &\sim \mathcal{PG}(b_\omega, 0) \\
&\text{and} \\
P(U_{k,s} = 1) &= 1 - P(U_{k,s} = 0) = q_n, \quad k = 1, \dots, r.
\end{aligned}$$

Although considered standard, for the sake of completeness, we give the definition of the HHMM model. In the Homogeneous Hidden Markov model, covariates affect the mean equation but the transition probability matrix is constant. That is,

$$\begin{aligned}
Y_t \mid Z_t = s &\sim \mathcal{N}\left(X_{t-1}^{(1)}B_s, \sigma_s^2\right), \quad s = 1, 2, \quad t = 1, \dots, T, \\
P(z_t = j \mid z_{t-1} = i) &= p_{ij}, \quad i, j = 1, 2 \quad \forall t = 1, \dots, T.
\end{aligned}$$

7.2 Case study for the fixed model

We simulated data from a NHHMM of size $T = 1500$. We used three covariates $X^{(1)} = \{1, X_1, X_2, X_3\}$ affecting the mean equation and three covariates $X^{(2)} = \{1, X_1, X_2, X_4\}$ affecting the transition matrix, with X independently normally distributed covariates with means $\mu_x = [4, 3, -2, -5]$ and variances $\sigma_x^2 = [1, 1, 0.5, 1]$. The mean equation parameters were $B_1 = [2, -0.3, 2, 2]'$, $\sigma_1^2 = 1.5$ and $B_2 = [1, 3, 4, 3]'$, $\sigma_2^2 = 0.8$ whereas the logistic regression coefficients were $\beta_1 = [1.5, 1, 2, 3]'$ and $\beta_2 = [3, -2.5, 4, 1]'$ for the two states, respectively. We kept $L = 100$ out-of-sample observations and we computed a sequence of one-step-ahead forecasts of the real observed process. We used non-informative priors for the unknown parameters σ_s, B_s, β_s , $s = 1, 2$, that is $\sigma_s^2 \sim \mathcal{IG}(0.1, 0.1)$, $B_s \mid \sigma_s^2 \sim \mathcal{N}(0, 100\sigma_s^2 \times I)$ and finally $\beta_s \sim \mathcal{N}(0, 100 \times I)$.

We used several metrics for assessing the efficiency of our algorithm (see Appendix 7.4). The quality of the sample is measured with the effective sample size (ESS), multivariate effective sample size (mESS). To assess the convergence and mixing of the algorithm we use the Potential Scale Reduction factor (PSRF) of Brooks and Gelman [1998], Gelman et al. [2013] and the multivariate mixing diagnostic of Paye [2012], hereinafter referred to as mCM.

Inferences are based on an MCMC sample of 25000 iterations after a burn in period of 10000 iterations. A summary of the results of this experiment is reported in Table 6. Specifically, we show that our algorithm has converged to the stationary distribution and has good mixing properties, using the aforementioned diagnostic criteria. Also, the univariate and multivariate effective sample sizes for all the methodologies/models are large, implying an efficient algorithm. In addition, the NHPG has best forecasting performance, since it has the best score among the benchmarks' scores in all forecasting criteria. In Figure 5 we visualize the empirical continuous approximation of the posterior predictive densities of NHPG, M&D, HHMM, for the three randomly selected out-of-sample periods, $L = 15, 85, 100$. These plots provide additional evidence that the NHPG gives at least as good estimates of the predictions as the other two considered models.

		NHPG	M&D	HHMM
Forecasting criteria	E(CRPS)	-2.0794	-2.1920	-4.4323
	MAFE	4.2661	4.3907	8.6745
	MSFE	60.9978	63.1617	143.1512
Sample Quality	ESS	11936	11934	12153
	mESS	24391	24343	24578
Convergence & Mixing	PSRF	1	1	1.0005
	mCM	18(18)	18(18)	8(10)

Table 6: Summary of the results of the experiment: Forecasting, Sample quality, Convergence and Mixing criteria. E(CRPS) is the mean continuous rank probability score, MAFE is the mean absolute forecast error and MSFE is the mean square forecast error for the 100 out-of-sample predictions. ESS: is the minimum effective size of among the ESS for all parameters and mESS the multivariate effective size, for an MCMC run of 25000 iterations. PSRF is the maximum potential scale reduction factor among all the PSRF for the studied model and mCM is the multivariate convergence and mixing diagnostic. In the mCM line we report the number of the components of the parameters out of the total components– in parenthesis– that fall into the 95% confidence interval of the test. NHPG is the proposed model, M&D is the model of Meligkotsidou and Dellaportas [2011], HHMM is the homogeneous model. Bold values denote the best values for the corresponding criterion among all the competing models.

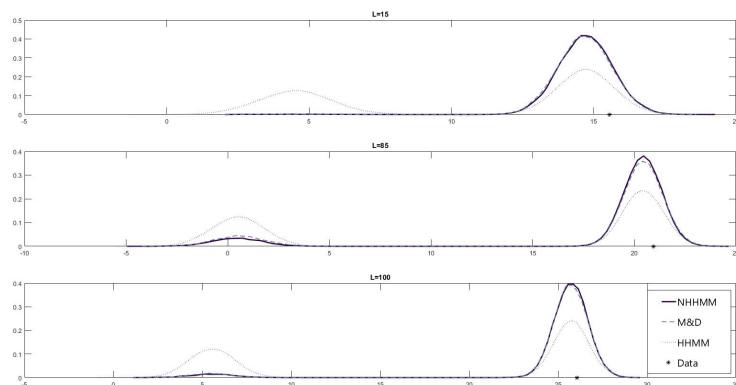


Figure 5: Conjointly plotted empirical continuous approximations of the posterior predictive distribution (based on a normal kernel function) for three out of sample periods ($L = 15, 85, 100$) using the NHPG (black continuous line), M&D (gray dashed line) and the HHMM (gray dotted line). Actual out-of-sample values are marked with asterisks.

7.3 The Continuous Rank Probability Score and

Let y_t be the real observed values of the forecasts, \mathbf{y} the history of the predictive quantity and \hat{y}_t the estimated forecasts. Using the notation f for the distribution

of the true model, $f_p(\hat{y}_l)$ for the posterior predictive density of the new data, and $F_p(x) = \int_{-\infty}^x f_p(\hat{y} | \mathbf{y}) d\hat{y}$ for the posterior predictive cumulative density function. The CRPS for y_i is defined as,

$$CRPS(F_{p,l}, y_l) = - \int_{-\infty}^{\infty} (F_{p,l}(\hat{y}_l) - F_{y_i}(\hat{y}_l))^2 d\hat{y}_l,$$

where $I(x \geq y)$ denotes a step function along the real line that attains the value 1 if $x \geq y$ and the value 0 otherwise, $F_{y_i} = H(\hat{y}_l - y_l)$ is the cumulative distribution of the real value y_l and H is the Heaviside function (Hersbach [2000]), $H(x) = 0$, if $x \leq 0$ and 1 otherwise.

7.4 Metrics of Comparison

As a primary metric of comparison, following Holmes and Held [2006] and Polson et al. [2013], we calculated the effective sample size (ESS). For each dimension of the parameter vector, ESS_i is the number of independent samples needed to obtain a parameter estimate with the same standard error as the MCMC estimate based on M dependent samples (see Neal [1993], Kass et al. [1998]). If θ is the p -dimensional parameter of interest, and $\theta_n = 1/n \sum_{t=1}^n g(x_t)$ is an estimate of θ based on a Markov chain $\{X_t\}$, with $\theta_n \rightarrow \theta$ the Monte Carlo error, $\theta_n - \theta$ is described asymptotically by the Central Limit Theorem (CLT), $\sqrt{n}(\theta_n - \theta) \xrightarrow[n \rightarrow \infty]{d} \mathcal{N}(0, \Sigma_p)$. The idea of the EES lies on the univariate CLT for each component of θ and it is defined, for $i = 1, \dots, p$, as

$$ESS_i = \frac{M}{1 + 2 \sum_{j=1}^k \rho(j)} = M \frac{\lambda_i}{\sigma_i},$$

where $\rho(k)$ is the sample autocorrelation of lag k of the parameter θ_i , λ_i the diagonal element of the sample covariance matrix Λ , σ_i the diagonal element of Σ_p and M the number of post-burn in samples.

Vats et al. [2018] argue that a univariate approach ignores cross-correlation across components, leading to an inaccurate picture of the quality of the sample. Thus, they define a multivariate version of the ESS. Specifically,

$$\text{mESS} = M \left(\frac{|\Lambda|}{|\Sigma_p|} \right)^{1/p}.$$

When there is no correlation, then $\Sigma = \Lambda$ and $\text{mESS} = M$.

To assess the convergence and mixing of our algorithm we use the Potential Scale Reduction factor (PSRF) of Brooks and Gelman [1998], Gelman et al. [2013] and the multivariate convergence and mixing diagnostic proposed by Paul et al. [2012], hereinafter referred to as mCM. In brief, implementation of PSRF requires sample runs from multiple chains (alternatively a very long chain can be divided into two or more subchains). The key quantity is the ratio of the resulting between- and within-chain variances. If the within-chain variance dominates the between-chain variance, the ratio approaches 1, which suggests that the chains have approximately reached stationarity. Desirable values for PSRF are the values below 1.1 for every component of the parameters. Paul et al. [2012] obtain MCMC-based estimators of posterior expectations by combining

different subgroup (subchain) estimators using stratification and poststratification methods. Variance estimates of the limiting distributions of such estimators are developed. Based on these variance estimates, they propose a statistic test to aid in the assessment of convergence and mixing of chains.

References

- H. Akaike, B. Petrov, and F. Csaki. Information theory and an extension of the maximum likelihood principle. *Proceedings of the Second International Symposium on Information Theory*, pages 267–281, 1973.
- T. G. Andersen, T. Bollerslev, F. X. Diebold, and P. Labys. Modeling and forecasting realized volatility. *Econometrica*, 71(2):579–62, 2003. doi: 10.1111/1468-0262.00418.
- R. T. Baillie, F. Calonaci, D. Cho, and S. Rho. Long memory, realized volatility and heterogeneous autoregressive models. *Journal of Time Series Analysis*, 2019. doi: 10.1111/jtsa.12470.
- M. M. Barbieri and J. O. Berger. Optimal predictive model selection. *The Annals of Statistics*, 32(3):870–897, 2004. doi: 10.1214/009053604000000238.
- L. E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The Annals of Mathematical Statistics*, 41(1):164–171, 1970. doi: 10.1214/aoms/1177697196.
- J. Besag, P. Green, D. Higdon, and K. Mengersen. Bayesian computation and stochastic systems. *Statistical Science*, 10(1):3–41, 1995. doi: 10.1214/ss/1177010123.
- M. Billio, A. Monfort, and C. Robert. Bayesian estimation of switching arma models. *Journal of Econometrics*, 93(2):229 – 255, 1999. ISSN 0304-4076. doi: 10.1016/S0304-4076(99)00010-X.
- G. Boero, J. Smith, and K. F. Wallis. Scoring rules and survey density forecasts. *International Journal of Forecasting*, 27(2):379 – 393, 2011. doi: 10.1016/j.ijforecast.2010.04.003.
- S. P. Brooks and A. Gelman. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4):434–455, 1998. doi: 10.1080/10618600.1998.10474787.
- O. Cappé, E. Moulines, and T. Ryden. *Inference in Hidden Markov Models (Springer Series in Statistics)*. Springer-Verlag New York, Inc., 2005. ISBN 0387402640.
- C. Christiansen, M. Schmeling, and A. Schrimpf. A comprehensive look at financial volatility prediction by economic variables. *Journal of Applied Econometrics*, 27(6):956–977, 2012. doi: 10.1002/jae.2298.
- P. Dellaportas, J. Forster, and I. Ntzoufras. On bayesian model and variable selection using mcmc. *Statistics and Computing*, 12(1):27–36, 2002. doi: 10.1023/A:1013164120801.

- F. X. Diebold, J.-H. Lee, and G. C. Weinbach. *Regime switching with time-varying transition probabilities*. In: *Hargreaves, C.P. (ed.) Nonstationary Time Series Analysis and Cointegration*. Oxford: Oxford University Press, 1994.
- E. F. Fama and K. R. French. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1):3 – 56, 1993. doi: 10.1016/0304-405X(93)90023-5.
- A. J. Filardo and S. F. Gordon. Business cycle durations. *Journal of Econometrics*, 85(1):99 – 123, 1998. doi: 10.1016/S0304-4076(97)00096-1.
- S. Frühwirth-Schnatter and R. Frühwirth. *Data Augmentation and MCMC for Binary and Multinomial Logit Models*. Physica-Verlag HD, 2010. doi: 10.1007/978-3-7908-2413-1_7.
- A. Fussl, S. Frühwirth-Schnatter, and R. Frühwirth. Efficient mcmc for binomial logit models. *ACM Trans. Model. Comput. Simul.*, 23(1):3:1–3:21, 2013. doi: 10.1145/2414416.2414419.
- A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2013.
- A. Gelman, J. Hwang, and A. Vehtari. Understanding predictive information criteria for bayesian models. *Statistics and Computing*, 24(6):997–1016, 2014. doi: 10.1007/s11222-013-9416-2.
- J. Geweke and C. Whiteman. Chapter 1 bayesian forecasting. In G. Elliott, C. Granger, and A. Timmermann, editors, *Handbook of Economic Forecasting*, volume 1, pages 3 – 80. Elsevier, 2006. doi: 10.1016/S1574-0706(05)01001-3.
- T. Gneiting. Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762, 2011. doi: 10.1198/jasa.2011.r10138.
- T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.
- N. Gospodinov, A. Gavala, and D. Jiang. Forecasting volatility. *Journal of Forecasting*, 25(6):381–400, 2006. doi: 10.1002/for.993.
- P. J. Green. Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711–732, 1995. doi: 10.1093/biomet/82.4.711.
- E. P. Gritti, T. Gneiting, V. J. Berrocal, and N. A. Johnson. The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification. *Quarterly Journal of the Royal Meteorological Society*, 132(621C):2925–2942, 2006. doi: 10.1256/qj.05.235.
- S. Gschlößl and C. Czado. Spatial modelling of claim frequency and claim size in non-life insurance. *Scandinavian Actuarial Journal*, 2007(3):202–225, 2007. doi: 10.1080/03461230701414764.

- D. I. Hastie and P. J. Green. Model choice using reversible jump markov chain monte carlo. *Statistica Neerlandica*, 66(3):309–338, 2011. doi: 10.1111/j.1467-9574.2012.00516.x.
- H. Hersbach. Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15(5):559–570, 2000. doi: 10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2.
- C. C. Holmes and L. Held. Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis*, 1(1):145–168, 2006. doi: 10.1214/06-BA105.
- T. Holsclaw, A. M. Greene, A. W. Robertson, and P. Smyth. Bayesian nonhomogeneous markov models via pólya-gamma data augmentation with applications to rainfall modeling. *The Annals of Applied Statistics*, 11(1):393–426, 2017. doi: 10.1214/16-AOAS1009.
- R. E. Kass, B. P. Carlin, A. Gelman, and R. M. Neal. Markov chain monte carlo in practice: A roundtable discussion. *The American Statistician*, 52(2):93–100, 1998. doi: 10.1080/00031305.1998.10480547.
- R. L. Machete. Contrasting probabilistic scoring rules. *Journal of Statistical Planning and Inference*, 143(10):1781 – 1790, 2013. doi: 10.1016/j.jspi.2013.05.012.
- M. McAleer and M. C. Medeiros. Realized volatility: A review. *Econometric Reviews*, 27(1-3):10–45, 2008. doi: 10.1080/07474930701853509.
- L. Meligkotsidou and P. Dellaportas. Forecasting with non-homogeneous hidden markov models. *Statistics and Computing*, 21(3):439–449, 2011. doi: 10.1007/s11222-010-9180-5.
- L. Meligkotsidou, E. Panopoulou, I. D. Vrontos, and S. D. Vrontos. Quantile forecast combinations in realised volatility prediction. *SSRN Electronic Journal*, 2018. doi: 10.2139/ssrn.2673529.
- S. Mittnik, N. Robinzonov, and M. Spindler. Stock market volatility: Identifying major drivers and the nature of their impact. *Journal of Banking & Finance*, 58:1 – 14, 2015. doi: 10.1016/j.jbankfin.2015.04.003.
- N. N. Narisetty and X. He. Bayesian variable selection with shrinking and diffusing priors. *Ann. Statist.*, 42(2):789–817, 2014. doi: 10.1214/14-AOS1207.
- N. N. Narisetty, J. Shen, and X. He. Skinny gibbs: A consistent and scalable gibbs sampler for model selection. *Journal of the American Statistical Association*, 0(0):1–13, 2018. doi: 10.1080/01621459.2018.1482754.
- R. M. Neal. Probabilitstic inference using markov chain monte carlo methods. Technical report, t RG-TR-93-1, University of Toronto, Department of Computer Science, 1993. URL <https://www.cs.toronto.edu/~radford/ftp/review.pdf>.
- D. B. Nelson. Conditional heteroskedasticity in asset returns: A new approach. *Econometrica*, 59(2):347–370, 1991. doi: 10.2307/2938260.

- S. M. O'Brien and D. B. Dunson. Bayesian multivariate logistic regression. *Biometrics*, 60(3):739–746, 2004. doi: 10.1111/j.0006-341X.2004.00224.x.
- R. O'Hara and M. J. Sillanpää. A review of bayesian variable selection methods: what, how and which. *Bayesian Analysis*, 4(1):85–117, 2009. doi: 10.1214/09-BA403.
- R. Paul, S. N. MacEachern, and L. M. Berliner. Assessing convergence and mixing of mcmc implementations via stratification. *Journal of Computational and Graphical Statistics*, 21(3):693–712, 2012. doi: 10.1080/10618600.2012.663293.
- B. S. Paye. 'déjà vol': Predictive regressions for aggregate stock market volatility using macroeconomic variables. *Journal of Financial Economics*, 106(3):527–546, 2012. doi: 10.1016/j.jfineco.2012.06.005.
- N. G. Polson, J. G. Scott, and J. Windle. Bayesian inference for logistic models using polya-gamma latent variables. *Journal of the American Statistical Association*, 108(504):1339–1349, 2013. doi: 10.1080/01621459.2013.829001.
- G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 03 1978. doi: 10.1214/aos/1176344136.
- S. L. Scott. Bayesian methods for hidden markov models. *Journal of the American Statistical Association*, 97(457):337–351, 2002. doi: 10.1198/016214502753479464.
- L. Spezia. Bayesian analysis of non-homogeneous hidden markov models. *Journal of Statistical Computation and Simulation*, 76(8):713–725, 2006. doi: 10.1080/10629360500108798.
- D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639, 2002. doi: 10.1111/1467-9868.00353.
- G. J. Székely and M. L. Rizzo. A new test for multivariate normality. *Journal of Multivariate Analysis*, 93(1):58–80, 2005. doi: 10.1016/j.jmva.2003.12.002.
- D. Vats, J. M. Flegal, and G. L. Jones. Multivariate output analysis for markov chain monte carlo. *Biometrika, to be appear*, 2018.
- R. Waagepetersen and D. Sorensen. A tutorial on reversible jump mcmc with a view toward applications in qtl-mapping. *International Statistical Review*, 69(1):49–61, 2001. doi: 10.1111/j.1751-5823.2001.tb00479.x.
- S. Watanabe. A widely applicable bayesian information criterion. *J. Mach. Learn. Res.*, 14(1):867–897, 2013. ISSN 1532-4435. URL <http://dl.acm.org/citation.cfm?id=2567709.2502609>.
- I. Welch and A. Goyal. A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies*, 21(4):1455–1508, 2008. doi: 10.1093/rfs/hhm014.