Journal of Econometrics 178 (2014) 779–793

Contents lists available at ScienceDirect

Journal of Econometrics

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

Time-varying sparsity in dynamic regression models

Maria Kalli[a,1], Jim E. Griffin[b,*]

[a] Canterbury Christ Church University Business School, Canterbury Christ Church University, Canterbury, UK

[b] School of Mathematics, Statistics & Actuarial Science, University of Kent, Canterbury, UK

ARTICLE INFO

Article history:
Received 21 February 2012
Received in revised form
23 August 2013
Accepted 31 October 2013
Available online 8 November 2013

JEL classification:
C11
C15
C52
C53
C58

Keywords:
Time-varying regression
Shrinkage priors
Normal-gamma priors
Markov chain Monte Carlo
Equity premium
Inflation

ABSTRACT

A novel Bayesian method for inference in dynamic regression models is proposed where both the values of the regression coefficients and the importance of the variables are allowed to change over time. We focus on forecasting and so the parsimony of the model is important for good performance. A prior is developed which allows the shrinkage of the regression coefficients to suitably change over time and an efficient Markov chain Monte Carlo method for posterior inference is described. The new method is applied to two forecasting problems in econometrics: equity premium prediction and inflation forecasting. The results show that this method outperforms current competing Bayesian methods.

1. Introduction

Forecasting, the estimation of a future value of a variable, plays an important role in both decision-making and strategic planning and has been extensively studied in econometrics. For example, forecasts of inflation affect the decisions of monetary and fiscal policymakers, investors who wish to hedge against the risk of nominal assets, and trade unions and management when they negotiate wage contracts, to name a few. Similarly, forecasts of equity premiums play an important role for investors who wish to diversify their equity portfolios to hedge against adverse market movements. The quality of the forecast depends on: the time scale involved (how far into the future we are trying to predict), the time period of the empirical sample, and the model used.

Regression models are a popular technique for forecasting since the values of other variables can be used to inform predictions. However, their use with observations made over time is complicated by several problems. Firstly, it has been found that these models can produce poor out-of-sample forecasts when the predictors' effects are assumed constant over time. This is generally taken as evidence that the effects of variables are time-varying. Sims (1980), Stock and Watson (1996), Cogley and Sargent (2001, 2005), Primicery (2005), Paye and Timmermann (2006), Ang and Bekaert (2007), Canova (2007), and Lettau and Van Nieuwerburgh (2008) are some studies providing evidence of time varying regressor effects in inflation and equity premium forecasting. Secondly, the increasing availability of large economic datasets has led to interest in using regression models with many regressors. It is well known that the estimation of regression models becomes more complicated when a large number of predictors is used due to the increased potential for over-fitting which can lead to poor out-of-sample forecasts or predictions. The problem of over-fitting can be alleviated by looking for sparse regression estimates where many regression coefficients are set to zero or values close to zero. This is usually achieved using regularisation of the regression coefficients or variable selection methods.

The problem of time-varying regressor effects can be addressed using dynamic regression models, which are a form of time-varying parameter models, where the regression coefficients are

* Corresponding author. Tel.: +44 1227 82 3865; fax: +44 1227 82 7932.
E-mail addresses: maria.kalli@canterbury.ac.uk (M. Kalli), J.E.Griffin-28@kent.ac.uk (J.E. Griffin).
1 Tel.: +44 1227863027; fax: +44 1227827932.

0304-4076/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jeconom.2013.10.012

assumed to evolve according to some stochastic process. This defines the dynamic linear models (DLM) discussed in West and Harrison (1999), or state-space models.

The problem of a large number of variables has been addressed in several ways. Initial work concentrated on models which assume a global measure of the importance of a variable. Groen et al. (2009) introduced a latent variable which indicates whether a variable is included in or excluded from the model. The approach is restricted so that the decision to include or exclude a predictor is irreversible. Belmonte et al. (2011) combined the Bayesian Lasso of Park and Casella (2008) with the model selection methods of Frühwirth-Schnatter and Wagner (2010) in order to have shrinkage in a dynamic regression setting. This approach allows some regression coefficients to be shrunk very close to zero for the whole time series and so effectively achieve variable selection. These methods have the potentially critical limitation that the importance of variables cannot change over time. For example, in some problems certain predictors could be useful for forecasting at particular times but not at others. In the Bayesian literature, this problem has been approached by allowing variables to enter and exit the model over time. Koop and Korobilis (2012) used the dynamic model averaging (DMA) method of Raftery et al. (2010) to select a suitable time-varying parameter model. However, the dynamics on model space are only implicitly defined by their approach. Alternatively, Chan et al. (2012) constructed the class of time-varying dimension (TVD) models using an explicitly constructed stochastic process for the subset of variables included in the model. This leads to a dynamic mixture model for which efficient posterior computational methods can be developed using the approach of Gerlach et al. (2000). However, both approaches are limited by the number of models that they can consider. The DMA approach uses full enumeration of posterior probabilities and the number of models with p regressors is 2^p precluding large values of p. Posterior computation in the TVD model also potentially involves all 2^p models but the authors suggest using a much restricted set of possible models.

The DMA and TVD approaches build on Bayesian variable selection techniques which explicitly consider all possible regression models. An alternative class of methods is Bayesian regularisation methods which use absolutely continuous priors and encourage small regression effects to be aggressively shrunk towards zero under the posterior (see Carvalho et al. (2010), and Polson and Scott (2011)). These authors have shown that these methods can lead to posteriors which place substantial mass on combinations of regression coefficients which are sparse (that is most of the regression coefficients have values very close to zero) if supported by the data. Belmonte et al. (2011) have already extended one such prior, the Bayesian Lasso, to the dynamic regression setting. Our methodological contribution differs from their work in two main respects. Firstly, our prior for the time-varying regression coefficients extends the more general normal-gamma (NG) prior (see Caron and Doucet (2008) and Griffin and Brown (2010)) to DR models and, secondly, our prior accounts for both time-varying regression coefficients and time-varying sparsity.

The paper is organised as follows: Section 2 introduces the normal-gamma autoregressive (NGAR) process prior, considers some of its properties and describes the full Bayesian model for dynamic regression with time-varying sparsity. Section 3 provides of an overview of the required Markov chain Monte Carlo (MCMC) method for fitting a dynamic regression model with an NGAR process prior (the full steps of the MCMC sampler are described in Appendix A). Section 4 applies the DR model with NGAR process priors to simulated data, while Section 5 considers empirical studies in equity premium prediction and inflation forecasting. Section 6 summarises our findings and conclusions.

2. A Bayesian dynamic regression model with time-varying sparsity

A dynamic regression (DR) model links a response y_t to regressors $x_{1,t}, \ldots, x_{m,t}$ (all observed at time t) by

$$y_t = \sum_{i=0}^{m} x_{i,t}\beta_{i,t} + \epsilon_t, \quad t = 1, \ldots, T,\ i = 0, \ldots, m \tag{1}$$

where $x_{0,t} = 1$ for all t (allowing for an intercept), $\beta_{i,t}$ is a vector of unknown coefficients for the ith regressor at time t, ϵ_t is the innovation term at time t generated from a normal distribution with zero mean and time-varying variance i.e. $\epsilon_t \sim \mathrm{N}(0, \sigma_t^2)$. The regressors $x_{1,t}, \ldots, x_{m,t}$ may include both lags of the response and exogenous variables. The DR model is usually completed by assuming that $\beta_{1,t}, \ldots, \beta_{m,t}$ follow a linear stochastic process (such as a random walk or vector autoregression).

In this paper we assume that the time-varying variances, $\sigma_1^2, \ldots, \sigma_T^2$ are generated by a gamma autoregressive (GAR) process using the method described in Pitt et al. (2002) and Pitt and Walker (2005) and later, independently, developed as the autoregressive gamma process by Gourieroux and Jasiak (2006). The process is specified using latent variables $\kappa_1^\sigma, \ldots, \kappa_{T-1}^\sigma$ by the recursion

$$\sigma_t^2 \sim \mathrm{Ga}\left(\lambda^\sigma + \kappa_{t-1}^\sigma, \lambda^\sigma / \left(\mu^\sigma \left(1 - \rho^\sigma\right)\right)\right) \quad \text{and}$$

$$\kappa_{t-1}^\sigma | \sigma_{t-1}^2 \sim \mathrm{Pn}\left(\lambda^\sigma \rho^\sigma \sigma_{t-1}^2 / \left(\left(1 - \rho^\sigma\right) \mu^\sigma\right)\right),$$

for $t = 2, \ldots, T$ and $\sigma_1^2 \sim \mathrm{Ga}\left(\lambda^\sigma, \lambda^\sigma/\mu^\sigma\right)$. This defines a first-order autoregressive model for $\sigma_1^2, \ldots, \sigma_T^2$ with autoregressive parameter ρ^σ and stationary distribution $\mathrm{Ga}(\lambda^\sigma, \lambda^\sigma/\mu^\sigma)$ where $x \sim \mathrm{Ga}(a, b)$ denotes that x follows a gamma distribution with shape parameter a and mean a/b. We discuss our choice of priors for $\lambda^\sigma, \mu^\sigma$, and ρ^σ in Section 2.2. The Bayesian model is completed by specifying a prior for $\beta_{i,t}$ for $i = 0, \ldots, m$ and $t = 1, \ldots, T$ which is discussed in the following section.

2.1. The NGAR process prior for $\beta_{i,t}$

In regression models with a large number of regressors, it is common to assume that only a subset of the regressors is important for prediction. In DR models, this assumption is naturally extended to subsets of important regressors that change over time. This assumption can be expressed in the prior by defining a stochastic process for $\beta_{1,t}, \ldots, \beta_{m,t}$ which allows a subset of $\beta_{1,t}, \ldots, \beta_{m,t}$ to be set equal to zero (or equivalently, some regressors to be removed from the model), or values close to zero at time t and allows the subset to change over time. We refer to the proportion of parameters $\delta = (\delta_1, \ldots, \delta_s)$ which are close to zero as the sparsity of δ with a larger proportion referred to as more sparsity. In DR models, there are two interesting forms of sparsity. Firstly, the sparsity of $\beta_i = (\beta_{i,1}, \ldots, \beta_{i,T})$ which is the proportion of time that $\beta_{i,t}$ is close to zero. Secondly, the sparsity of $\beta_{1,t}, \ldots, \beta_{m,t}$ which is the proportion of regression coefficients that are set close to zero at time t. The assumption of time-varying subsets of important variables can be expressed by time-varying sparsity of $\beta_{1,t}, \ldots, \beta_{m,t}$.

These forms of sparsity can be expressed by giving independent normal-gamma autoregressive (NGAR) process priors to the time series of regression coefficients $\beta_1, \ldots, \beta_m$. We define the NGAR process prior below. Let $x \sim \mathrm{Pn}(\mu)$ denote that x follows a Poisson distribution with mean μ.

Definition 1. The normal-gamma autoregressive (NGAR) process for β_i is defined by

$$\beta_{i,s} = \sqrt{\frac{\psi_{i,s}}{\psi_{i,s-1}}}\varphi_i \beta_{i,s-1} + \eta_{i,s},$$

$$\eta_{i,s} | \psi_{i,s} \sim \mathrm{N}\left(0, (1 - \varphi_i^2)\psi_{i,s}\right) \quad s = 2, \ldots, T,$$

which is a normal AR(1) process conditional on $\psi_i = (\psi_{i,1}, \ldots, \psi_{i,T})$, and where $\psi_{i,s}$ follows a first-order gamma autoregressive process specified using latent variables $\kappa_{i,1}, \ldots, \kappa_{i,s-1}$ via the recursion

$$\psi_{i,s}|\kappa_{i,s-1} \sim \mathrm{Ga}\left(\lambda_i + \kappa_{i,s-1}, \frac{\lambda_i}{\mu_i(1-\rho_i)}\right),$$

$$\kappa_{i,s-1}|\psi_{i,s-1} \sim \mathrm{Pn}\left(\frac{\rho_i\lambda_i\psi_{i,s-1}}{\mu_i(1-\rho_i)}\right),$$

with

$$\beta_{i,1}|\psi_{i,1} \sim \mathrm{N}(0, \psi_{i,1}), \qquad \psi_{i,1} \sim \mathrm{Ga}(\lambda_i, \lambda_i/\mu_i).$$

The normal-gamma autoregressive process will be written as $\beta_i \sim \mathrm{NGAR}(\lambda_i, \mu_i, \varphi_i, \rho_i)$.

The NGAR process can also be represented as the product of two independent stationary stochastic processes: $\psi_i = (\psi_{i,1}, \ldots, \psi_{i,T})$ and $\phi_i = (\phi_{i,1}, \ldots, \phi_{i,T})$. Under this representation, $\beta_{i,t} = \sqrt{\psi_{i,t}}\phi_{i,t}$ where $\phi_i = (\phi_{i,1}, \ldots, \phi_{i,T})$ is generated from an AR(1) process with autocorrelation parameter φ_i such that ϕ_i has the standard normal as its stationary distribution $\left(\text{i.e. } \phi_i = \varphi_i\,\phi_{i-1} + \varsigma_i \text{ where } \varsigma_i \sim \mathrm{N}(0, 1-\varphi_i^2)\right)$. The parameters $\psi_i = (\psi_{i,1}, \ldots, \psi_{i,T})$ are generated from a GAR process whose stationary distribution is $\mathrm{Ga}(\lambda_i, \frac{\lambda_i}{\mu_i})$. The conditional density of $\psi_{i,t}$ given $\psi_{i,t-1}$ is equal to

$$\sum_{\kappa=0}^{\infty} w_{\kappa,\psi_{i,t-1}} \mathrm{Ga}\left(\psi_{i,t}\,\middle|\,\lambda_i + \kappa, \frac{\lambda_i}{\mu_i(1-\rho_i)}\right) \tag{2}$$

which is a mixture of gamma distributions with parameters $\lambda_i + \kappa$, and $\frac{\lambda_i}{\mu_i(1-\rho_i)}$, and Poisson weights such that

$$w_{\kappa,\psi_{i,t-1}} = \frac{\exp\left\{-\frac{\rho_i\lambda_i}{\mu_i(1-\rho_i)}\psi_{i,t-1}\right\}\left(\frac{\rho_i\lambda_i}{\mu_i(1-\rho_i)}\psi_{i,t-1}\right)^{\kappa}}{\kappa!}. \tag{3}$$

The mean of $\psi_{i,t}$ given $\psi_{i,t-1}$ is

$$\mathrm{E}[\psi_{i,t}|\psi_{i,t-1}] = \mu_i(1-\rho_i) + \rho_i\psi_{i,t-1},$$

which has an autoregressive structure, and its conditional variance is

$$\mathrm{Var}[\psi_{i,t}|\psi_{i,t-1}] = \frac{\mu_i^2(1-\rho_i)^2}{\lambda_i} + \frac{2\rho_i\mu_i(1-\rho_i)\psi_{i,t-1}}{\lambda_i}.$$

The process β_i is stationary and has a normal-gamma stationary distribution since ψ_i and ϕ_i are independent and stationary. The unconditional variance of $\beta_{i,t}$ is $\mathrm{Var}(\beta_{i,t}) = \mu_i$ and the excess kurtosis is $\varkappa(\beta_{i,t}) = 3/\lambda_i$.

The set-up of the NGAR process prior is similar to that of the normal-gamma (NG) prior and can be considered as the natural extension of the NG prior to dynamic regression problems. Griffin and Brown (2010) study the use of a normal-gamma (NG) prior for regression problems where a sample $y_1, \ldots, y_n$ is observed with m-dimensional vectors of regressors $x_1, \ldots, x_n$ and is modelled by

$$y_j = x_j\beta + \epsilon_j, \quad \epsilon_j \sim \mathrm{N}(0, \sigma^2),\ j = 1, \ldots, n.$$

The prior assumes that $\beta_1, \ldots, \beta_m$ are independent with $\beta_i|\psi_i$ following a normal distribution with mean zero and variance ψ_i, where $\psi_i \sim \mathrm{Ga}(\lambda, \lambda/\mu)$. In this hierarchical set-up, the parameter ψ_i plays a key role in determining the shrinkage to zero of the least squares estimate $\hat{\beta}$ induced by the posterior mean $\mathrm{E}[\beta_i|y]$. Smaller values of ψ_i will lead to the posterior mean of β_i (conditional on ψ_i) being increasingly shrunk to zero. We can thus describe ψ_i as the relevance of the ith regressor with a smaller value of ψ_i implying less relevance for the ith regressor (as the posterior mean is increasingly shrunk to zero). This interpretation of the variance of a normal prior distribution for a regression coefficient dates back to, at least, Tipping (2000) and Bishop and Tipping (2000). Griffin and Brown (2010) show that the proportion of mass close to zero in the marginal prior distribution of β_i or, equivalently, the prior distribution of ψ_i controls the sparsity of the posterior means of the regression coefficients with smaller values of λ implying higher levels of sparsity. In other words, smaller values of λ imply that more regression coefficients will have posterior means close to zero if supported by the data.

The NGAR process prior has a similar structure with $\beta_{i,t}|\psi_{i,t}$ following a normal distribution with mean zero and variance $\psi_{i,t}$ and $\psi_{i,t}$ following a gamma distribution, marginally. Therefore, similar to the NG prior, $\psi_{i,t}$ plays the role of relevance of the ith regressor at time t. Small values of $\psi_{i,t}$ will lead to greater shrinkage of $\beta_{i,t}$. For a fixed prior mean, μ_i, as the value of the sparsity parameter λ_i decreases more prior mass for $\beta_{i,t}$ is placed close to zero and so the process tends to spend a greater proportion of time close to zero.

Figs. 1 and 2 display simulated paths of the NGAR process for both $\psi_{i,t}$ and $\beta_{i,t}$ with different combinations of λ_i, φ_i and ρ_i. These illustrate the ability of the prior to generate periods where the regression coefficients are close to zero and periods where the regression coefficients are away from zero.

The sparsity parameter λ_i clearly controls the proportion of time that the regression coefficient spends close to zero. This proportion becomes larger as λ_i decreases which is illustrated in Figs. 1 and 2 where $\lambda_i = 0.2$ and $\lambda_i = 1$ respectively. Smaller values of λ_i lead to "spikier" processes for $\psi_{i,t}$ and $\beta_{i,t}$ which favours increasingly rapid changes from small to large values. The autocorrelation parameter ρ_i controls the dependence between $\psi_{i,t-1}$ and $\psi_{i,t}$. Larger values of ρ_i lead to a larger autocorrelation and favour processes which spend longer periods close to zero or away from zero. Decreasing the value of ρ_i allows the regressors to increasingly jump from values near to zero to values away from zero (and vice versa). The autocorrelation parameter φ_i controls the dependence between $\beta_{i,t}$ and $\beta_{i,t-1}$ conditional on the $\psi_i = (\psi_{i,1}, \ldots, \psi_{i,T})$ process. Hyperpriors for the parameters of the NGAR process prior are described in the following section.

2.2. Bayesian inference for DR models and hyperpriors for the parameters of the NGAR process priors

We consider the DR model with independent NGAR processes priors for the regression coefficients. The NGAR process prior allows us to control the sparsity of the posterior distribution of the regression coefficients (*i.e.* the proportion of regression coefficients with mass close to zero at a time t), and assumes that regressors should rarely jump in and out of the model. Both assumptions are important to avoid over-fitting of the DR model.

The parameter μ_i acts as an overall relevance parameter for the ith regression coefficient since it controls the marginal variance of $\beta_{i,t}$. In particular, $\beta_{i,t}$ will be close to zero for all t if μ_i is small. Therefore, a hierarchical prior is specified for $\mu_1, \ldots, \mu_m$ with

$$\mu_i \sim \mathrm{Ga}(\lambda^\star, \lambda^\star/\mu^\star), \quad i = 0, \ldots, m$$

and

$$\lambda^\star \sim \mathrm{Ex}(1/s^\star), \qquad p(\mu^\star) \propto (\mu^\star + 2b^\star)^{-3}$$

where $\mathrm{Ex}(\gamma)$ represents an exponential distribution with mean $1/\gamma$. This introduces a second level of sparsity (at the level of the regressors rather than the time-varying regression coefficients). This is particularly important in problems with many regressors where some regressors have no regression effect across all observations. The hyperparameter $s^\star$ is the prior mean of $\lambda^\star$ and so gives an initial idea of the level of sparsity using the ideas described

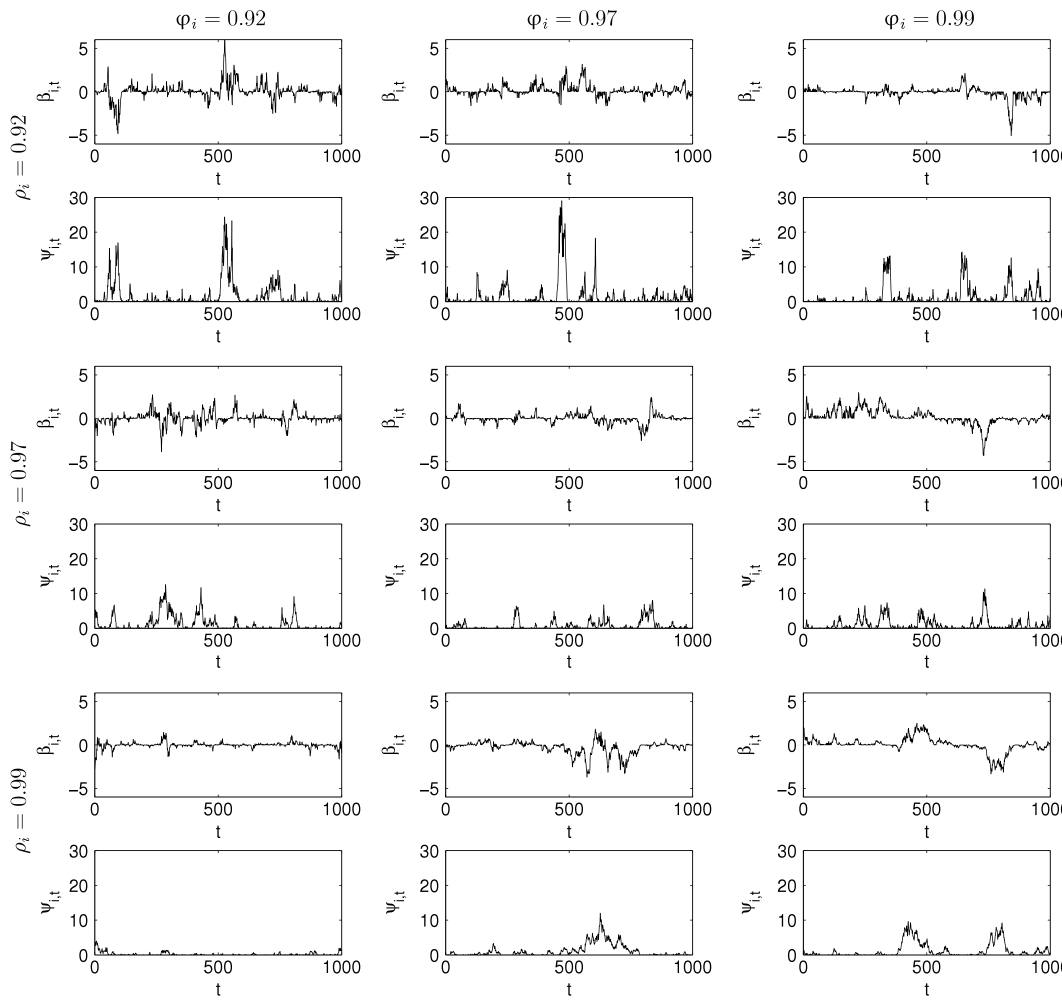

Fig. 1. Simulated paths of β_t and ψ_t with different values of ρ_i and φ_i with $\lambda_i = 0.2$ and $\mathrm{Var}[\beta_{i,t}] = \mu_i = 1$.

in Griffin and Brown (2010). The parameter $\mu^\star$ is given a heavy-tailed prior with prior mean $b^\star$, which is given a value suitable for the spread of the regression coefficients in the particular application.

The NGAR process prior now has two sparsity parameters. The parameter $\lambda^\star$ is the sparsity parameter for μ_i, with smaller values of $\lambda^\star$ indicating that more μ_i's are close to zero. This implies that $\beta_{i,t}$ is close to zero at all time t for more regressors, since μ_i is the variance of β_i under the stationary distribution. In contrast, λ_i controls the sparsity within the time series of the ith regression coefficient and a small value of λ_i would indicate that the regression coefficient is close to zero for a large proportion of observations. The sparsity parameter λ_i is given the prior

$$p(\lambda_i) \propto \lambda_i(0.5 + \lambda_i)^{-4}$$

which is a heavy-tailed prior giving values around 1. This centres the prior over the Lasso cases (which arises when $\lambda_i = 1$).

The flexibility of the NGAR process prior can lead to over-fitting when the values of φ_i and ρ_i are small. The problem of over-fitting is particularly acute in DR models since we have m regression coefficients at each time point. The realisations in Figs. 1 and 2 confirm that even a value of ρ_i close to 0.9 allows regressors to quickly be excluded from the DR model. Our prior is

$$\varphi_i \sim \mathrm{Be}(77.6, 2.4), \qquad \rho_i \sim \mathrm{Be}(77.6, 2.4), \quad i = 0, \ldots, m,$$

which gives a prior mean of 0.97 with most mass over 0.9 and implies that the processes for the regression coefficients and the relevances are strongly autocorrelated. This effectively excludes models which allow the regression coefficients to rapidly change over time (and lead to over-fitting).

The priors for the parameters of the volatility process σ_t^2 are chosen as

$$\lambda^\sigma \sim \mathrm{Ga}(3, 1), \qquad p(\mu^\sigma) \propto (1 + \mu^\sigma)^{-3/2}, \qquad \rho^\sigma \sim \mathrm{Be}(38, 2).$$

The choice for λ^σ signifies that the volatility process will have stationary distribution which is less heavy tailed than a Laplace distribution. The mean μ^σ is given a very heavy tailed prior to allow for a wide-range of possible values. The dependence parameter ρ^σ is given an informative prior that enforces stationarity and places most of its mass on values greater than 0.85. This seems reasonable given the value usually associated with stochastic volatility models.

3. Computation

MCMC methods to fit a DR model, as in (1), with an NGAR process prior are described in this section. It is possible to use a standard Gibbs sampler which simulates from each full conditional distribution in turn. However, this approach leads to highly autocorrelated draws. The problem is mainly caused by the correlation between the process $(\beta_i, \psi_i, \kappa_i)$ and its parameters $\theta_i = (\lambda_i, \mu_i, \rho_i, \varphi_i)$ and the process $(\sigma^2, \kappa^\sigma)$ and its parameters $\theta^\sigma = (\lambda^\sigma, \mu^\sigma, \rho^\sigma)$. Our sampling scheme involves jointly updating (ψ_i, κ_i) with θ_i and (σ^2, κ) with θ^σ whilst integrating out a subset

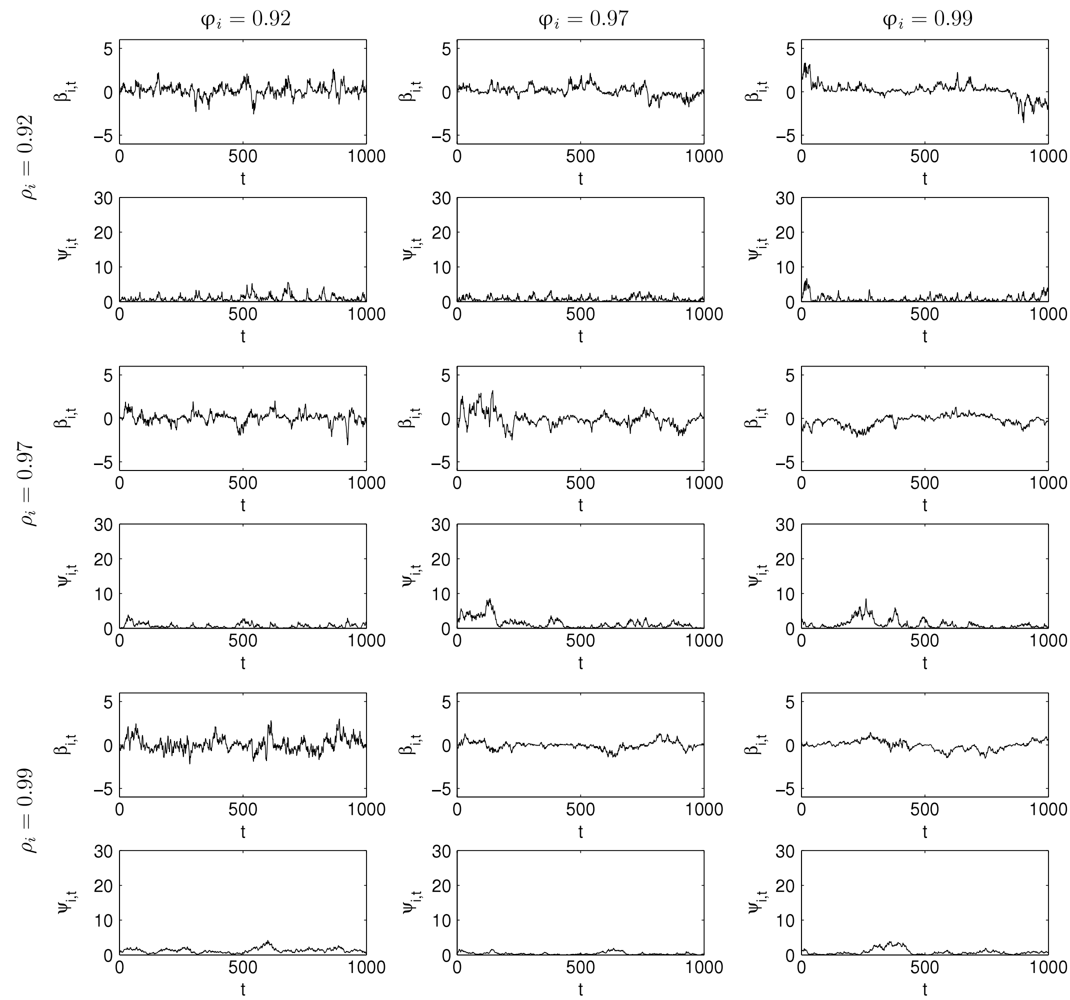

Fig. 2. Simulated paths of $\beta_{i,t}$ and $\psi_{i,t}$ with different values of ρ_i and φ_i with $\lambda_i = 1$ and $\mathrm{Var}[\beta_{i,t}] = \mu_i = 1$.

β_M of $\beta_1, \ldots, \beta_m$ (which changes at each iteration). The full conditional densities can be calculated using the Kalman filter since the regression model conditional on ψ is a Gaussian state-space model. Realisations of β_M can subsequently be generated using standard forward-filtering backwards-sampling (Frühwirth-Schnatter, 1994; Carter and Kohn, 1994).

Many full conditional distributions in the sampler have non-standard forms and so adaptive MCMC methods are used (see, *e.g.* Andrieu and Thoms, 2008; Griffin and Stephens, 2013, for reviews). These methods allow the proposal in a Metropolis–Hastings sampler to depend on previous samples and so break the Markov assumption which underlies the use of MCMC. Roberts and Rosenthal (2007) gave relatively weak conditions for the convergence of such methods. In our sampler, we assume that the state space is bounded at extremely large and small values. This has no practical effect on the implementation of the MCMC methods. Two particular forms of adaptive update are used: the adaptive Metropolis–Hastings random walk (AMHRW) step (Atchadé and Rosenthal, 2005) and the adaptive scale within adaptive Metropolis (ASWAM) step (Haario et al., 2001; Atchade and Fort, 2010). In the AMHRW step for a parameter ξ, at the γth iteration, a new value $\xi' \sim \mathrm{N}\left(\xi^{(\gamma-1)}, \tau_\gamma^\xi\right)$ is proposed where $\xi^{(j)}$ is the value of ξ at the jth iteration. The acceptance probability α_γ^ξ is calculated in the usual way and the proposal is either accepted or rejected. The variance of the proposal is updated using

$$\log \tau_{\gamma+1}^\xi = \log \tau_\gamma^\xi + \gamma^{-\eta}\left(\alpha_\gamma^\xi - \hat{\alpha}\right) \tag{4}$$

where $0.5 < \eta \leq 1$ and $0 < \hat{\alpha} < 1$. Under the scheme, the variance τ_γ^ξ converges to a value with average acceptance rate $\hat{\alpha}$ (which we set to 0.3 in our examples). This method is restricted to univariate parameters. Multivariate parameters can be updated using ASWAM. Let ξ be a p-dimensional parameter, a Metropolis–Hastings random walk step is used with the proposal $\xi' \sim \mathrm{N}\left(\xi^{(\gamma-1)}, s_\gamma^\xi S_\gamma^\xi\right)$ where $s_\gamma^\xi > 0$ is a scale parameter and S_γ^ξ is the sample variance–covariance of ξ calculated using the first $(\gamma - 1)$ draws from the MCMC sampler. The proposed value is accepted or rejected as usual in a Metropolis–Hastings step and the value of s_γ^ξ is updated using $\log s_{\gamma+1}^\xi = \log s_\gamma^\xi + \gamma^{-\eta}\left(\alpha_\gamma^\xi - \hat{\alpha}\right)$.

The joint updating of (ψ_i, κ_i) with θ_i and (σ^2, κ) with θ^σ uses the following method which can be seen as a form of retrospective sampling (see, e.g. Papaspiliopoulos et al., 2007). Suppose a reversible Markov chain is constructed for $(\psi_\theta, \kappa_\theta)$ whose stationary distribution is the marginal distribution of ψ, κ given θ and with transitions of the form $q_{\psi,\kappa}\left((\psi_\theta, \kappa_\theta), (\psi'_{\theta'}, \kappa'_{\theta'})\right)$. A Metropolis–Hastings sampler can be constructed with a proposal of the form

$$q\left((\psi, \kappa, \theta)\right) = q_{\psi,\kappa}\left((\psi, \kappa), (\psi', \kappa')\right) q_\theta\left(\theta, \theta'\right).$$

The acceptance probability in the Metropolis–Hastings step is

$$\min\left\{1, \frac{p(y|\psi', \kappa')q(\theta', \theta)}{p(y|\psi, \kappa)q(\theta', \theta)}\right\}$$

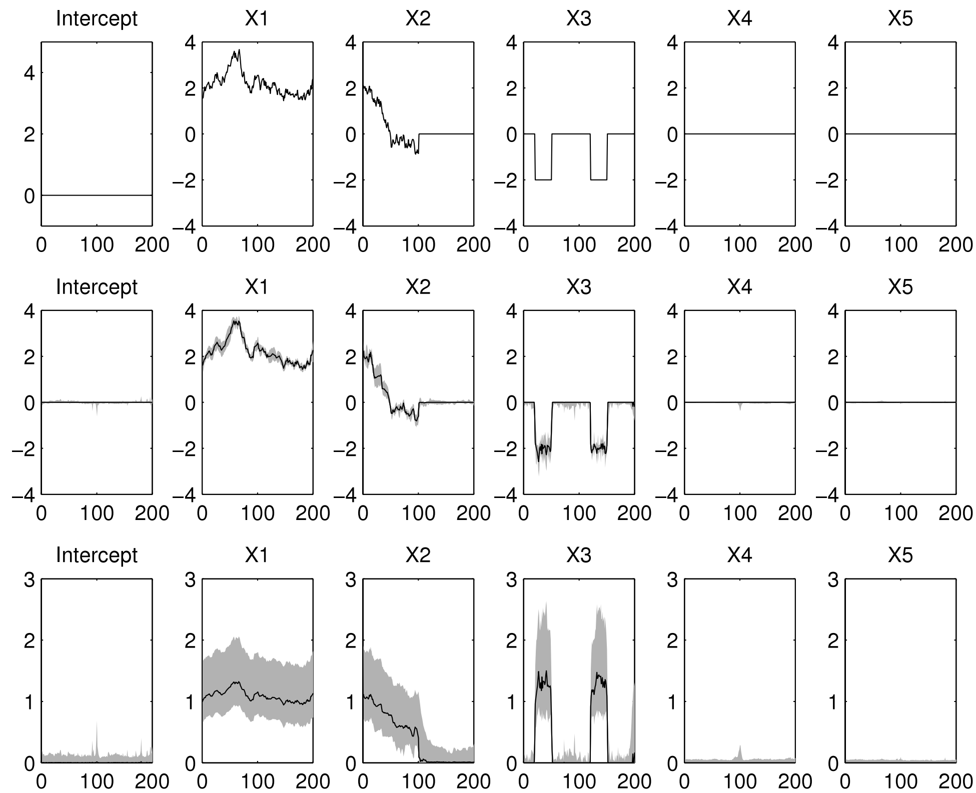

Fig. 3. Simulated: First row—true regression coefficients $\beta_{i,t}$, Second row—the posterior median (solid line) and 95% credible interval (grey shading) of $\beta_{i,t}$, and Third row—the posterior median (solid line) and 95% credible interval (grey shading) of $\sqrt{\psi_{i,t}}$.

since the chain is reversible and so obeys detailed balance. A suitable Markov chain for the NGAR process can be constructed using

$$\psi'_{1,i} = \begin{cases} \dfrac{\lambda_i \mu_i'}{\lambda_i' \mu_i} \psi_{i,1} + \mathrm{Ga}(\lambda_i' - \lambda_i, \lambda_i'/\mu_i') & \text{if } \lambda_i' > \lambda_i \\ \dfrac{\lambda_i \mu_i'}{\lambda_i' \mu_i} \psi_{i,1}\, \mathrm{Be}(\lambda_i', \lambda_i - \lambda_i') & \text{if } \lambda_i' < \lambda_i \end{cases}$$

$$\kappa'_{i,t} = \begin{cases} \kappa_{i,t} + \mathrm{Pn}\left(\dfrac{\rho_j'}{1-\rho_j'} \dfrac{\psi'_{i,t}\lambda_i'}{\mu_i'} - \dfrac{\rho_j}{1-\rho_j} \dfrac{\psi_{i,t}\lambda_i}{\mu_i} \right) \\ \qquad \text{if } \dfrac{\rho_i'}{1-\rho_i'} \dfrac{\psi'_{i,t}\lambda_i'}{\mu_i'} > \dfrac{\rho_i}{1-\rho_i} \dfrac{\psi_{i,t}\lambda_i}{\mu_i} \\ \mathrm{Bi}\left(\kappa_{i,t}, \dfrac{\rho_i'(1-\rho_i)}{\rho_i(1-\rho_i')} \dfrac{\mu_i \psi'_{i,t}\lambda_i'}{\mu_i' \psi_{i,t}\lambda_i} \right) \\ \qquad \text{if } \dfrac{\rho_i'}{1-\rho_i'} \dfrac{\psi'_{i,t}\lambda_i'}{\mu_i'} < \dfrac{\rho_i}{1-\rho_i} \dfrac{\psi_{i,t}\lambda_i}{\mu_i} \end{cases}$$

for $t = 1, \ldots, T-1$ and

$$\psi'_{i,t} = \begin{cases} \dfrac{\lambda_i \mu_i'(1-\rho_i')}{\lambda_i \mu_i (1-\rho_i)} \psi_{i,t} \\ \qquad + \mathrm{Ga}\left(\lambda_i' + \kappa'_{i,t-1} - \lambda_i - \kappa_{i,t-1}, \dfrac{\lambda_i'}{\mu_i'(1-\rho_i')} \right) \\ \qquad \text{if } \lambda_i' + \kappa'_{i,t-1} > \lambda_i + \kappa_{i,t-1} \\ \dfrac{\lambda_i \mu_i'(1-\rho_i')}{\lambda_i \mu_i (1-\rho_i)} \psi_{i,t}\, \mathrm{Be}(\lambda_i' + \kappa'_{i,t-1}, \lambda_i \\ \qquad + \kappa_{i,t-1} - \lambda_i' - \kappa'_{i,t-1}) \quad \text{if } \lambda_i' + \kappa'_{i,t-1} < \lambda_i + \kappa_{i,t-1} \end{cases}$$

for $t = 2, \ldots, T$.

The steps of the MCMC sampler are described in Appendix A.

4. Simulated example

The following simulated example illustrates the ability of the NGAR process prior to allow time-varying sparsity in dynamic regression. We generated the data from Eq. (1) with $m = 5$, $x_{i,t} \sim \mathrm{N}(0, I_5)$ and $x_{i,1}, \ldots, x_{i,T}$ independent. We introduced five regression coefficients: $\beta_{1,t}$ followed an AR(1) process with AR parameter 0.97 and a normal marginal distribution with mean 2 and variance 0.25, $\beta_{2,t}$ followed an AR(1) process with AR parameter 0.97 and a normal marginal distribution with mean 0 and variance 0.25 for $t < 100$ and $\beta_{2,t} = 0$ for $t > 100$ with $\beta_{2,1} \sim \mathrm{N}(2, 0.25)$,

$$\beta_{3,t} = \begin{cases} 0 & \text{if } t \le 20,\ 51 \le t \le 120, \text{ and } 151 \le t \le 200 \\ -2 & \text{if } 21 \le t \le 50, \text{ and } 121 \le t \le 150, \end{cases}$$

and $\beta_{0,t}$ (the intercept), $\beta_{4,t}$ and $\beta_{5,t}$ were zero for all times. The innovation variance σ_t^2 was generated using an AR(1) process on the log scale

$$\log \sigma_t^2 = \log(0.01) + 0.97(\log \sigma_{t-1}^2 - \log(0.01)) + \sqrt{\frac{0.01}{1-0.97^2}}\,\nu_t$$

where $\nu_t \sim \mathrm{N}(0, 1)$. The initial value of each parameter was drawn from its stationary distribution. The generated values of the regression coefficients are shown in the first row of Fig. 3 where $\beta_{1,t}$ is always important, $\beta_{4,t}$ and $\beta_{5,t}$ are never important, the importance of $\beta_{2,t}$ tends to decrease until $t = 100$ after which the value of $\beta_{2,t}$ is zero, and $\beta_{3,t}$ enters and exits the model abruptly on two occasions.

Inference for the DR model with NGAR process prior, $s^\star = 0.1$ and $b^\star = 0.1$ are shown in Fig. 3. The MCMC sampler was run

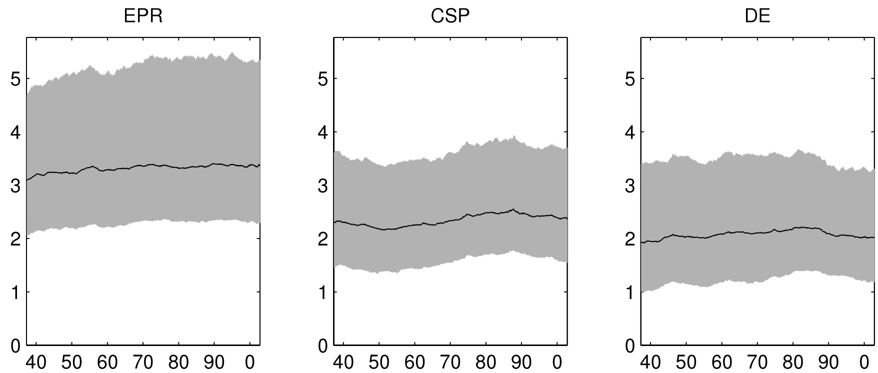

Fig. 4. Equity Premium: Posterior medians (solid line) of the time-varying regression relevance, $\sqrt{\psi_{i,t}}$. (y_{axis} = value of $\sqrt{\psi_{i,t}}$, x_{axis} = time) with 95% credible interval (grey shading).

for 100,000 iterations, 10,000 were discarded as an initial burn-in and every 18th draw was saved. An iMac with a 2.66 GHz Intel Core i5 processor, and memory 4 GB 1067 MHz DDR3 was used, and the computing time was 6 h. The second row of Fig. 3 shows the estimated regression coefficients which follow the true values closely. The posterior median and 95% credible interval of $\beta_{0,t}$, $\beta_{4,t}$ and $\beta_{5,t}$ are very close to zero. The NGAR process prior is also able to adapt to the changing importance of $\beta_{2,t}$ and the abrupt behaviour of $\beta_{3,t}$ in the model. The third row of Fig. 3 shows the posterior inference on the regressor time-varying relevance factor $\sqrt{\psi_{i,t}}$. The values for the intercept, x_4 and x_5 are fairly constant and close to zero. The posterior median of the relevance factor for x_2 is decreasing until about $t = 100$ and then takes a value close to zero, whereas the posterior median of the relevance factor for x_3 correctly replicates the abrupt entries and exits of the regressor from the model. The results of this simulation illustrate the ability of the NGAR process prior to shrink values close to zero when the data supports this.

5. Empirical examples

In this section we apply the dynamic regression model with an NGAR process prior for the regression coefficients to both equity premium and inflation datasets. Our aim is to provide evidence that this model adequately accounts for the time-varying effect of the regression coefficients, identifies variables that are most relevant at each time point and produces good out-of-sample forecasts. The value $s^\star = 0.1$ was used in all examples, which leads to substantial selection of the regressors. We used $b^\star = 2$ in the equity premium prediction example and $b^\star = 0.1$ in the inflation forecasting examples. This represented the different scales of the regression effects in the two examples. The MCMC sampler was run for 100,000 iterations, 10,000 were discarded as an initial burn-in and every 18th draw was saved. The computing time was 15 h for the inflation forecasting examples and 26 h for the equity premium prediction example.

In Section 2, we discussed the NGAR process prior for the time-varying regression coefficient, $\beta_{i,t}$, and the time-varying relevance of the ith regressor, $\psi_{i,t}$. Recall that a smaller value of $\psi_{i,t}$ implies that the ith regressor is less important at time t. We present two plots for each dataset. The first plot displays the posterior median of $\sqrt{\psi_{i,t}}$ as it changes over time and shows the importance of each predictor over time (including periods where it has most impact). The second plot displays the posterior median of $\beta_{i,t}$ over time, which illustrates the effect of each relevant predictor. It is natural to expect that when a predictor is not relevant (when the posterior median of $\sqrt{\psi_{i,t}}$ is zero), then the value of $\beta_{i,t}$ should be very close to zero. We also plot the time-varying innovation variance, σ_t^2, to identify the periods when σ_t^2 changes. All plots are displayed with 95% credible intervals (CI).

5.1. Equity premium prediction

The set of variables relevant to equity premium forecasting is large. It ranges from variables relating to dividends and earnings such as dividend yield and price earnings ratio to interest rates, bond yields, and inflation. For our empirical study we use the same dataset as Goyal and Welch (2008). The response variable is the value weighted monthly return of the S&P 500 obtained from the CRSP database. For our illustration we considered all the twelve predictors (see Appendix B for the complete list), including cross sectional beta premium (CSP) (see Roll and Ross, 1994). For this reason the sample period is restricted from May 1937 to December 2002, as it is the period where values of CSP are available.

From the original list of thirteen predictors, ten had posterior medians and 95% CI's for both $\sqrt{\psi_{i,t}}$ and $\beta_{i,t}$ that were very close to zero for the whole observation period and therefore have been excluded from our plots. These excluded variables were: B/M, TBL, LTY, NTIS, INFL, LTR, D/Y, SVAR, DFY, and DRF (see Appendix B for full details). The plots of the posterior medians and 95% credible intervals (CI's) of $\sqrt{\psi_{i,t}}$ for EPR (earnings price ratio), CSP (relative valuation of high and low beta stocks), and DE (dividend payout ratio), the three relevant predictors, and the posterior medians and 95% CI's of $\beta_{i,t}$ for these predictors are displayed in Figs. 4 and 5 respectively. The relevance of EPR is relatively constant over time. The same is true for its regression coefficient, which has a posterior median around 7 for the whole period. EPR has a positive effect on the equity premium for the whole period which is expected as it signals a firm's profitability. The regression coefficients for CSP and DE show more fluctuation in their relevance over time. However, their relevance is still relatively constant over time. The coefficient of CSP is almost always positive. It increases from the mid 1950's up to the late 1980's and then it decreases. In addition we can also observe an oscillating pattern within this gradual increase and decrease of the CSP effect. One possible explanation is that within each decade there are years of high economic growth followed by years of slow growth. The beta of the firm is a measure of the firm risk that is attributed to the market and cannot be diversified. The beta will be high in times of recession and will affect equity premiums more than during periods of high growth. The coefficient of DE is also positive for all time periods, with its effect being largest during the 1980's, the period when the Reagan administration began the deregulation of US financial markets. The effect decreases from the 1990's to 2000's and this could be due to the shift of emphasis in investment decisions from DE to firm growth prospects. Goyal and Welch (2008) do not provide estimates for the regression coefficients but look at the importance of predictors by running simple regressions for different periods within the sample.

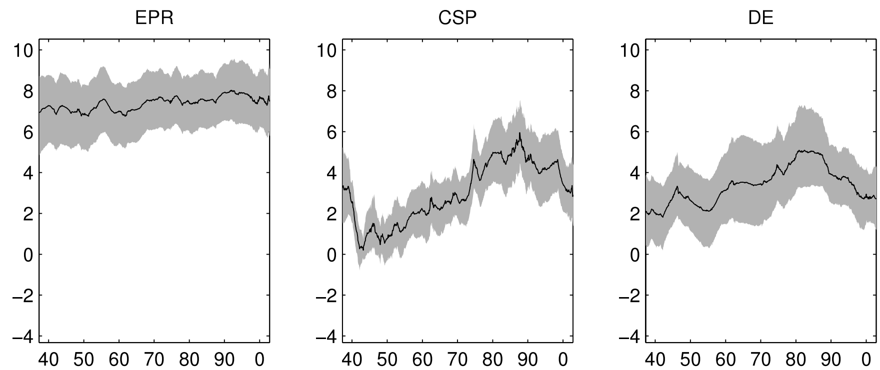

Fig. 5. Equity Premium: Posterior median (solid line) of the time-varying regression coefficients, $\beta_{i,t}$. (y_{axis} = value of $\beta_{i,t}$, x_{axis} = time) with 95% credible interval (grey shading).

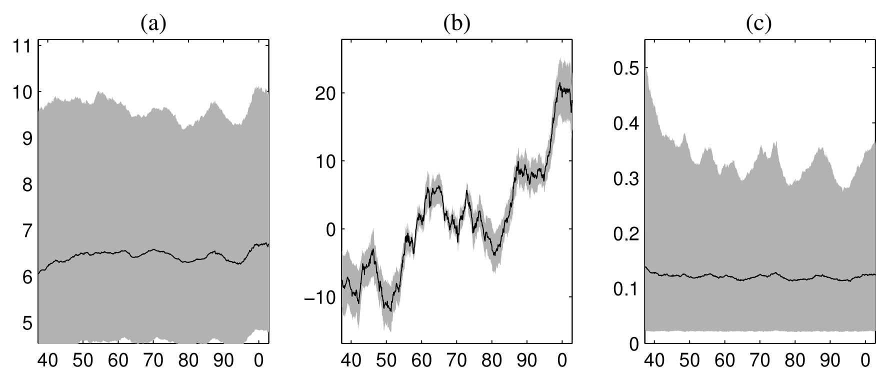

Fig. 6. Equity Premium: (a) the posterior median (solid line) and 95% CI (grey shading) of the relevances of the intercept $\sqrt{\psi_{0,t}}$, (b) the posterior median (solid line) and 95% CI (grey shading) of the intercept $\beta_{0,t}$, and (c) posterior median (solid line) and 95% CI (grey shading) of the time-varying innovation volatility, σ_t^2. (y_{axis} = values of $\sqrt{\psi_{0,t}}$, $\beta_{0,t}$ and σ_t^2 respectively, x_{axis} = time).

Fig. 6 displays the time-varying relevance and the effect of the intercept (the first two plots) and the behaviour of the innovation variance, σ_t^2 over time. The importance of the intercept is fairly constant. Its effect is increasing over time and is positive with the exception of the period of the second World War and the beginning of the 1950's which was a period of reorganisation following the War. The innovation variance is fairly constant over time, around 0.12.

5.2. Inflation forecasting

Forecasts of inflation are usually classified according to the type of explanatory variables used. The size of the set of potential variables is huge and is usually split into four subsets: past inflation forecasts, where the explanatory variables are previous lags of inflation; Phillips curve forecasts, which involve activity variables, such as economic growth rate or output gap, unemployment rate, and lagged inflation; forecasts based on variables which are themselves forecasts of asset prices (combination indices), term structures of nominal debt, and consumer surveys; and forecasts based on other exogenous variables such as government investment, the number of new private houses built, etc.

For our inflation forecasting study we constructed a dataset using data series obtained from: FRED, the economic database of the Federal Reserve Bank of St. Louis, the consumer survey database of the University of Michigan, the Federal Reserve Bank of Philadelphia, and the Institute of Supply Management. We use two different quarterly measures of US inflation as the response variable, the personal consumption expenditure (PCE) deflator and the gross domestic product (GDP) deflator. We therefore fit two separate models. The sample period for both is from the second quarter of 1965 to first quarter of 2011. Our dataset includes 31 predictors, from activity and term structure variables to survey forecasts and previous lags. The full list with details of each is included in Appendix B.

5.2.1. PCE deflator results

We first discuss the results based on the PCE deflator. From the thirty one predictors sixteen had a posterior median for $\sqrt{\psi_{i,t}}$ and $\beta_{i,t}$ greater than zero. The plots of the posterior median and 95% CI of $\sqrt{\psi_{i,t}}$ are displayed in Fig. 7 and the plots for the posterior median and 95% CI of $\beta_{i,t}$ are displayed in Fig. 8. Since the posterior median of $\beta_{i,t}$ is rarely far from zero, we interpret the 95% credible interval as a set of plausible values for the regression coefficients. This allows us to identify times when a large absolute value of the regression coefficient is implausible and also variables for which it is implausible that the regression coefficient takes a particular sign (either positive or negative).

Of these sixteen predictors, the ones that are noticeably relevant in forecasting the PCE deflator are: IMGS (import of goods and services) growth, INF EXP (inflation expectation), T-Bill 3m rate, Materials, Lags 2 and 4, DJIA (S&P 500 returns), GDP and PCE growth.

The most relevant predictor of PCE deflator is IMGS growth. The value of $\sqrt{\psi_{i,t}}$ is fairly constant over time, being higher in the 1970's (oil shock) and late 2000's (financial crisis) which were periods of economic slowdown. Its coefficient is always positive, with higher values during these two periods. The importance of INF EXP is slightly decreasing up to the early 1990's and from

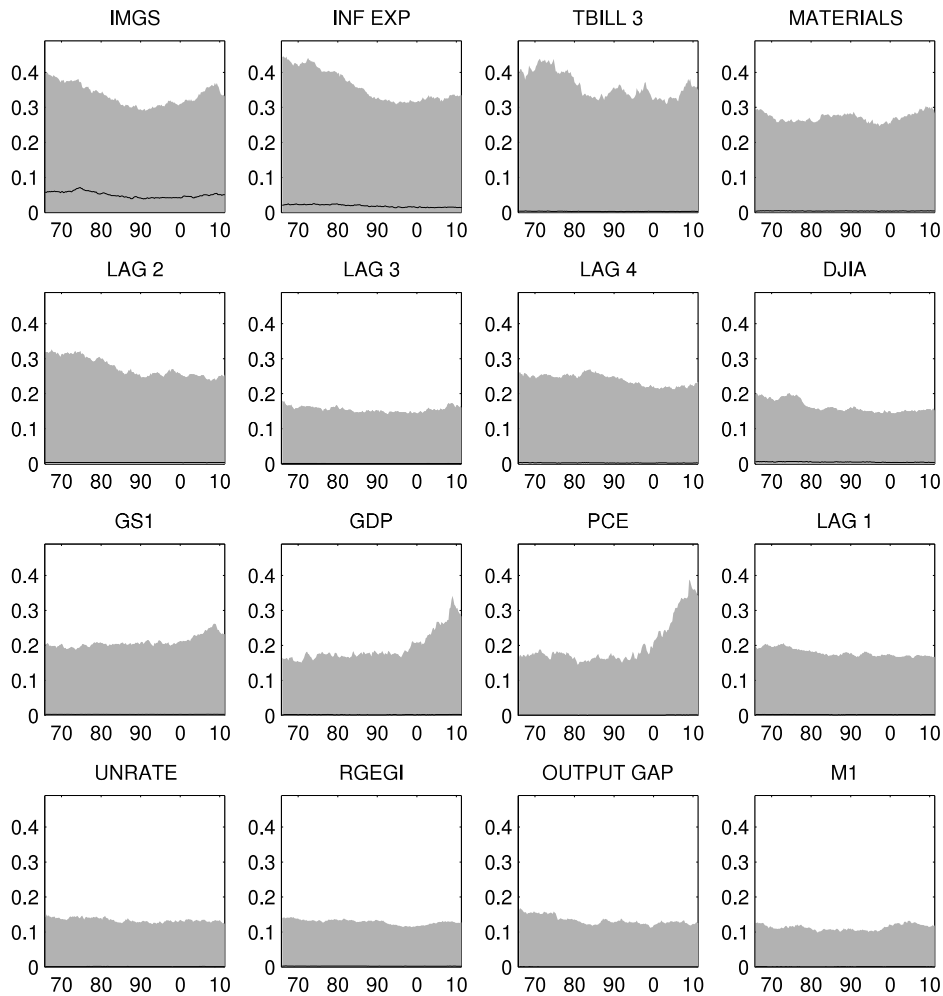

Fig. 7. PCE deflator: Posterior medians (solid line) of the time-varying regression relevance, $\sqrt{\psi_{i,t}}$. (y_{axis} = value of $\sqrt{\psi_{i,t}}$, x_{axis} = time) with 95% CI (grey shading).

then on it stabilises. Overall, it has a positive coefficient, reaching its peak value (around 0.55) in the mid 1970's and mid 1980's. From then on its value decreases, possibly turning negative in the late 2000's. The relevance of T-Bill 3m rate, Materials, Lags 2 and 4, and DJIA (S&P 500 returns) is fairly constant over time, but the value of their coefficients is not, it fluctuates. For the T-Bill 3m rate it jumps from 0.2 to 0.6 in the mid 1970's it then falls to around 0.1 and remains there until the late 2000's when it increases to around 0.45. For Materials the coefficient is negative in the 1980's and 1990's but not the other periods. The coefficients of the two Lags appear to move in the opposite direction. Lag 2's coefficient is positive up to the early 2000's and then decreases becoming negative in the late 2000's, whereas that of Lag 4 starts off positive, turns negative between 1980's and 1990's and then turns positive. The coefficient of DJIA is negative up to the mid 1990's and then turns positive, and the coefficients of GDP and PCE growth are very close to zero up to the mid 2000's and then become positive.

The time-varying relevance and effect of the intercept are shown in the first two plots of Fig. 9, whereas the third plot displays the behaviour of the innovation variance σ_t^2. The relevance of the intercept is higher during the 1970's oil shock when compared to all other periods. Its coefficient is positive in the 1970's and 1980's and turns negative in the early 1990's. The innovation variance oscillates over time and reaches its peak around 2008, the start of the recent financial crisis.

5.2.2. GDP deflator results

The posterior median and 95% CI plot of $\sqrt{\psi_{i,t}}$ for each of the predictors of the GDP deflator are displayed in the plots of Fig. 10. The posterior median and 95% CI plot of $\beta_{i,t}$ for each predictor are displayed Fig. 11.

Sixteen predictors are identified with posterior median for $\sqrt{\psi_{i,t}}$ substantially greater than zero. Almost double the number compared to the PCE deflator. This is reasonable as GDP reflects the value of all finished goods and services produced within the country whereas the PCE reflects personal consumption of goods and services. These predictors were: INF EXP, Lag 3, Materials, RGEGI growth, Private employment, Non farm payroll (NFP), T-Bill 3m rate, Lag 4, IMGS growth, Output gap, DJIA, M1 (money supply), Lag 2, Housing starts, T-Bill spread, and GS 1 (Treasury constant maturity rate).

The most relevant predictor of GDP deflator is INF EXP. Its relevance is highest during the 1970's and then decreases. Its coefficient is positive with highest values form the 1970's to early 1980's when it starts to decrease and settles around 0.1. The relevance of the other predictors is fairly constant but the values of their coefficients differ. For T-Bill 3m rate, Lag 4, IMGS growth, Output gap, DJIA, M1, Lag 2, Housing starts, T-Bill spread, and GS 1 the coefficients are more or less stable. The 95% credible interval bands suggest that there are periods were a change in the regression coefficient sign is possible. For example for M1, around the early 2000's, its 95% credible interval band hints to a sign change from positive to negative, but this sign change is more

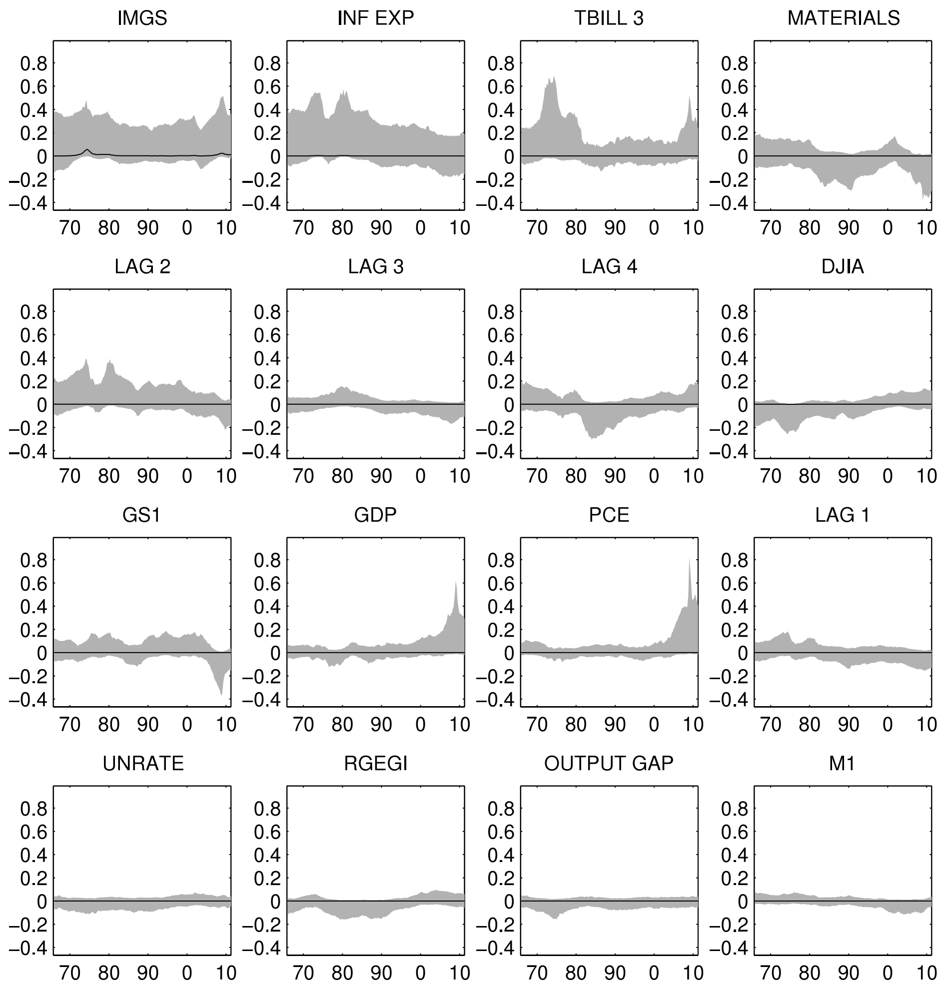

Fig. 8. PCE deflator: Posterior medians (sold line) of the time-varying regression coefficients, $\beta_{i,t}$. (y_{axis} = value of $\beta_{i,t}$, x_{axis} = time) with 95% CI (grey shading).

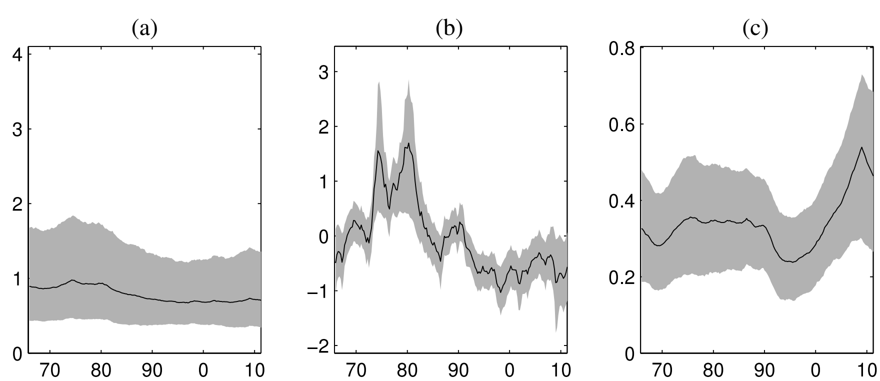

Fig. 9. PCE deflator: (a) the posterior median (solid line) and 95% CI (grey shading) of the relevances of the intercept $\sqrt{\psi_{0,t}}$, (b) the posterior median (solid line) and 95% CI (grey shading) of the intercept $\beta_{0,t}$, and (c) posterior median (solid line) and 95% CI (grey shading) of the time-varying innovation volatility, σ_t^2. (y_{axis} = values of $\sqrt{\psi_{0,t}}$, $\beta_{0,t}$ and σ_t^2 respectively, x_{axis} = time).

clearly displayed in the case of Materials, RGEGI, PRIVATE EMP, NFP. For MATERIALS the coefficient is positive between the 1970's and 1980's possibly due the oil shock which led to increases in prices and thus to higher inflation. From the 1980's onwards this effect becomes smaller. It starts fairly small (possibly negative) and it then becomes positive (from the 1990's to late 2000's). Periods of economic growth always provide more employment opportunities in the private sector. The coefficient of NFP (non farm payroll) also has the same effect on the GDP deflator. The coefficient of RGEGI growth is negative in the 1980's and then turns positive but it is very close to zero. Finally the coefficient of the third lag of GDP deflator is almost always positive as is the case with that of IMGS growth.

The NGAR process prior can induce bimodal posterior distributions with one mode at zero and one mode away from zero. Fig. 12 illustrates this interesting aspect of the posterior for the INF

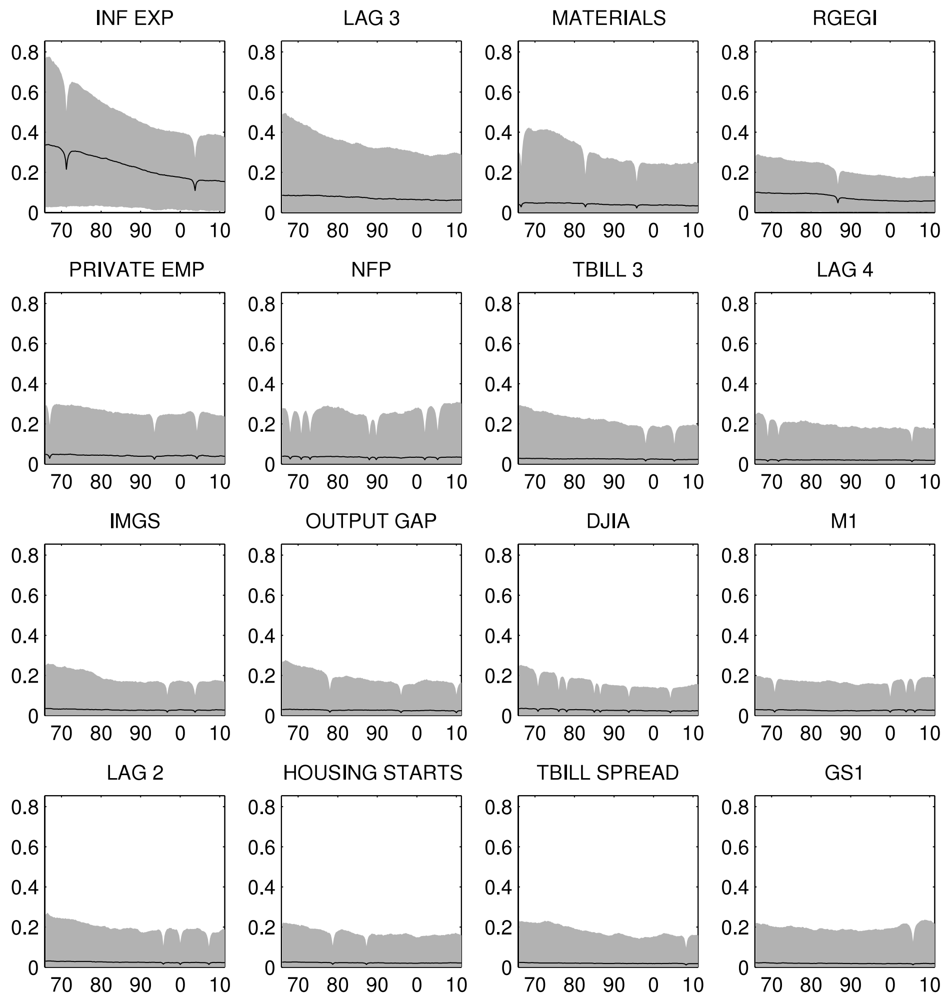

Fig. 10. GDP deflator, $h = 1$: Posterior medians (solid line) of the time-varying regression relevance, $\sqrt{\psi_{i,t}}$. (y_{axis} = values of $\sqrt{\psi_{i,t}}$, x_{axis} = time) with 95% CI (grey shading).

EXP regressor. The posterior median is much larger in 1973 and 1983 with one mode at about 0.5. In 1993 and 2003, the secondary mode disappears and a single mode at zero remains. These posteriors can be interpreted as expressing our uncertainty about the importance of a particular regressor.

The time-varying relevance and effect of the intercept are shown in the first two plots of Fig. 13, whereas the third plot displays the behaviour of the innovation variance σ_t^2. The relevance of the intercept is higher during the 1970's oil shock when compared to all other periods. Its coefficient is positive in the 1970's and 1980's and turns negative in the late 1990's. The innovation variance oscillates over time and reaches its peak in the mid 1970's and around 2008, two recessionary periods.

5.3. Comparison to other methods

We compare the predictive performance of the DR model with the NGAR process prior to other Bayesian variable selection and regularisation methods that have recently been proposed for DR models with a large number of potential predictors. These methods are: Time Varying Dimension (TVD) models (Chan et al., 2012), the dynamic model average (DMA) approach (Koop and Korobilis, 2012), and the hierarchical shrinkage (HierShrink) prior of Belmonte et al. (2011). In the case of the TVD and the HierShrink methods, the priors suggested in the related papers were used and in the case of the DMA we set the "forgetting" parameters $\lambda = \alpha = 0.99$, as the paper suggested. We also use a rolling window Bayesian Model Averaging (BMA) approach with a g-prior for prediction. We use the default choices of Fernandez et al. (2001) for the g-prior with the previous k observations, i.e. $y_{t-k}, \ldots, y_{t-1}$, to predict y_t at each time point. Finally, we use the random walk model of Atkeson and Ohanian (2001) as a Benchmark model. We focus on one step ahead forecasts and our comparison metric is the root mean square error (RMSE) using the posterior mean as our estimate calculated on the second half of the data

$$\sqrt{\frac{1}{T-s}\sum_{t=s+1}^{T}\left(y_t - \mathrm{E}\left[y_t | y_1, \ldots, y_{t-1}, x_1, \ldots, x_t\right]\right)^2}$$

where $x_t = (x_{0,t}, x_{1,t}, \ldots, x_{m,t})$ and $s = \lfloor \frac{T}{2} \rfloor$ (*i.e.* the largest integer less than or equal to $T/2$). It could be argued that the posterior median and the mean absolute error are more appropriate here given the multi-modal posterior distributions of $\beta_{i,t}$ (Gneiting, 2011) but we concentrate on the more commonly used RMSE. The posterior predictive means of y_t for $t = s+1, \ldots, T$ were calculated using particle filtering methods and includes uncertainty in all parameters.

Table 1 displays the RMSE for all three datasets under the different models. There are three versions of TVD which make

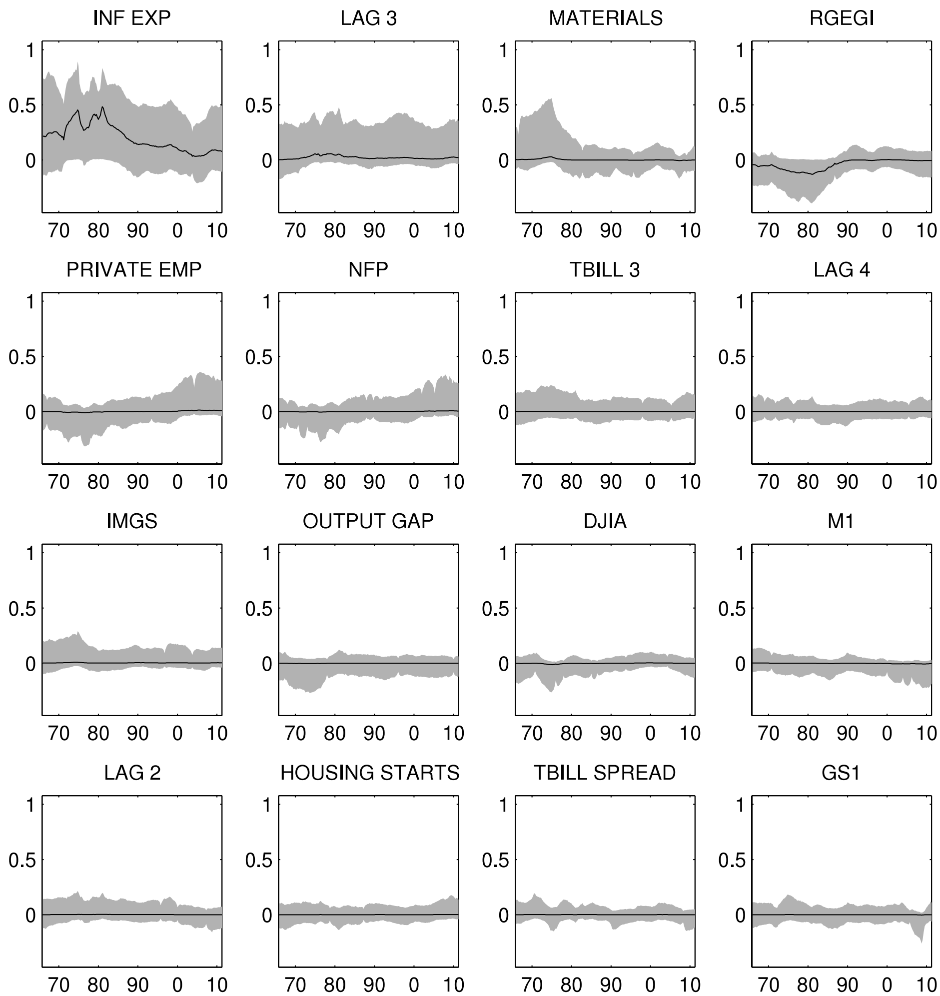

Fig. 11. GDP deflator: Posterior medians (solid line) of the time-varying regression coefficients, $\beta_{i,t}$. (y_{axis} = values of $\beta_{i,t}$, x_{axis} = time) with 95% CI (grey shading).

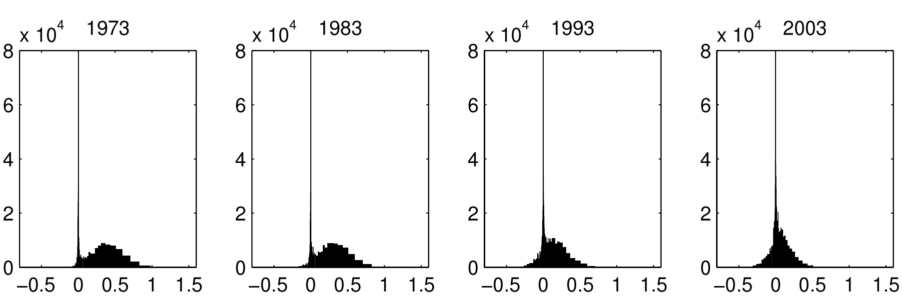

Fig. 12. GDP deflator: Histograms of the posterior distribution of the time-varying regression coefficients of INF EXP in the second quarter of 1973, 1983, 1993 and 2003.

Table 1
RMSE of out-of-sample prediction with different priors for the three datasets. The smallest RMSE for each dataset is written in bold.

	Equity premium	PCE inflation	GDP inflation
RW	1.100	0.635	**0.373**
NGAR	**0.977**	**0.611**	0.410
DMA	1.01	0.660	0.422
TVD1	2.193	2.688	2.688
TVD2	0.986	0.623	0.481
TVD3	0.992	0.628	0.500
HierShrink	1.547	1.131	2.556
gprior^1	2.822	0.796	0.660
gprior^2	1.648	0.712	0.588
gprior^3	1.282	0.681	0.516

different assumptions about the evolution of the regression coefficients and which are fully described in Chan et al. (2012). The window lengths for the three g-priors were 100 (gprior^1), 200 (gprior^2) and 300 (gprior^3) for the equity premium data and 50 (gprior^1), 70 (gprior^2) and 90 (gprior^3) for the inflation data. The choice of the window is controlled by the number of regressors included (which must be less than the window length) and the number of observations in the sample. The DR model with NGAR process prior is the best performing approach for two datasets (equity premium and PCE inflation) and the second best performing for the GDP inflation data (with only the random walk giving better predictions). The TVD2 and TVD3 model and DMA also perform well across the three datasets. In general,

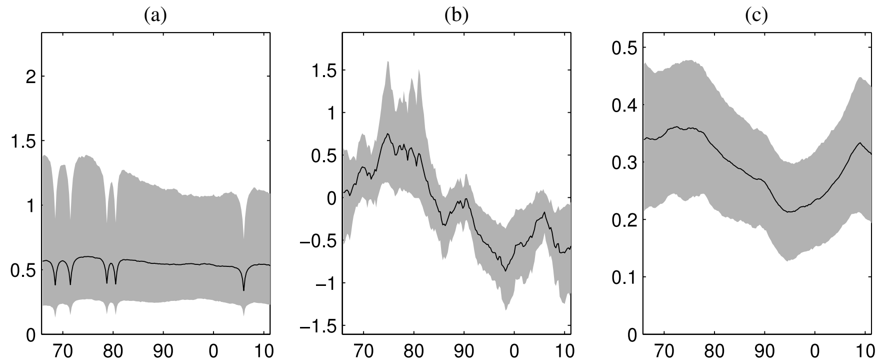

Fig. 13. GDP deflator: (a) the posterior median (solid line) and 95% CI (grey shading) of the relevances of the intercept $\sqrt{\psi_{0,t}}$, (b) the posterior median (solid line) and 95% CI (grey shading) of the intercept $\beta_{0,t}$, and (c) posterior median (solid line) and 95% CI (grey shading) of the time-varying innovation volatility, σ_t. (y_{axis} = values of $\sqrt{\psi_{0,t}}$, $\beta_{0,t}$ and σ_t^2 respectively, x_{axis} = time).

the approaches which allow the complexity of the regression model to change over time (NGAR, TVD and DMA) outperform the other approaches (HierShrink and rolling window g-prior). This illustrates the importance of allowing time-variation in the relevance of regression coefficients. The poor performance of the HierShrink prior suggests that the double exponential prior may be unsuitable with these data and imply too little sparsity. This illustrates the importance of allowing for time-varying sparsity in these data.

6. Discussion

This paper introduces a new approach to time-varying sparsity in dynamic regression models. The time-varying regression coefficients follow a stochastic process with a normal-gamma marginal distribution and smaller values of the shape parameter imply that the process will spend more time at values close to zero. This allows us to identify periods when regression coefficients are very close to zero and so are effectively removed from the model. A normal-gamma prior on the variance of the marginal distribution of $\beta_{i,t}$ encourages shrinkage of the whole path of $\beta_{i,t}$ close to zero. The empirical examples illustrate that the method leads to a smaller out-of-sample predictive RMSE than several recently proposed approaches to dynamic regression models with many regressors.

Acknowledgements

We would like to thank the Editor, Associate Editor and two referees for their valuable comments.

Appendix A. Gibbs sampler

Updating ψ

The full conditional density of $\psi_{i,1}$ is proportional to

$$\psi_{i,1}{}^{\lambda_i+\kappa_{i,1}-3/2}\left(\frac{\lambda_i\rho_i}{(1-\rho_i)\mu_i}\right)^{\kappa_{i,1}}\exp\left\{-\frac{\psi_{i,1}\lambda_i}{\mu_i\,(1-\rho_i)}\right\}$$
$$\times\ \exp\left\{-\frac{1}{2}\left[\frac{\beta_{i,1}^2}{\psi_{i,1}}+\frac{\left(\beta_{i,2}-\varphi_i\sqrt{\frac{\psi_{i,2}}{\psi_{i,1}}}\beta_{i,1}\right)^2}{\psi_{i,2}\left(1-\varphi_i^2\right)}\right]\right\},$$

the full conditional density of $\psi_{i,t}$ is proportional to

$$\psi_{i,t}{}^{\lambda_i+\kappa_{i,t-1}+\kappa_{i,t}-3/2}\left(\frac{\rho_i\lambda_i}{(1-\rho_i)\mu_i}\right)^{\kappa_{i,t}}(1-\varphi^2)^{1/2}$$
$$\times\ \exp\left\{-\frac{\lambda_i\psi_{i,t}}{\mu_i}\left(1+\frac{2\rho_i\lambda_i}{(1-\rho_i)\mu_i}\right)\right\}$$
$$\times\ \exp\left\{-\frac{1}{2}\left[\frac{\left(\beta_{i,t}-\varphi_i\sqrt{\frac{\psi_{i,t}}{\psi_{i,t-1}}}\beta_{i,t-1}\right)^2}{\psi_{i,t}\left(1-\varphi_i^2\right)}\right.\right.$$
$$\left.\left.+\frac{\left(\beta_{i,t+1}-\varphi_i\sqrt{\frac{\psi_{i,t+1}}{\psi_{i,t}}}\beta_{i,t}\right)^2}{\psi_{i,t+1}\left(1-\varphi_i^2\right)}\right]\right\}$$

if $1<t<T$ and the full conditional density of $\psi_{i,T}$ is proportional to

$$\psi_{i,T}^{\lambda_i+\kappa_{i,T-1}-3/2}(1-\varphi_i^2)^{-1/2}\exp\left\{-\frac{\lambda_i\psi_{i,T}}{(1-\rho_i)\mu_i}\right\}$$
$$\times\ \exp\left\{-\frac{1}{2}\left[\frac{\left(\beta_{i,T}-\varphi_i\sqrt{\frac{\psi_{i,T}}{\psi_{i,T-1}}}\beta_{i,T-1}\right)^2}{\psi_{i,T}\left(1-\varphi_i^2\right)}\right]\right\}.$$

Each parameter can be updated using an AMHRW step where the new value $\psi_{i,t}'$ is proposed according to $\log\psi_{i,t}{}'\sim\mathrm{N}\left(\log\psi_{i,t},\tau_{i,t}^{\psi}\right)$ with the variances updated as in (4).

Updating κ

The full conditional distribution of $\kappa_{i,t}$ is proportional to

$$\left(\frac{\lambda_i}{(1-\rho_i)\mu_i}\right)^{\lambda_i+\kappa_{i,t}}\left(\frac{\psi_{i,t}\lambda_i\rho_i}{(1-\rho_i)\mu_i}\right)^{\kappa_{i,t}}\psi_{i,t+1}^{\lambda_i+\kappa_{i,t}-1}\frac{1}{\kappa_{i,t}!\,\Gamma(\lambda_i+\kappa_{i,t})}$$

for $1\le t\le T-1$. We update this parameter using an AMHRW step, which is a variation of the method of Atchadé and Rosenthal (2005). At the γth iteration, the proposed value $\kappa_{i,t}'=\kappa_{i,t}^{\gamma}+d\epsilon$ where $p(d=-1)=p(d=1)=1/2$ and $\epsilon\sim\mathrm{Ge}\left(1\Big/\left(1+z_{i,t}^{\gamma}\right)\right)$. Here $x\sim\mathrm{Ge}(q)$ denotes that x follows a geometric distribution whose probability mass function is

$$p(x)=q(1-q)^x,\quad x=0,1,2,\ldots.$$

The proposed value is rejected is $\kappa_{i,t}'<0$ and otherwise accepted using the standard Metropolis–Hastings acceptance probability for a random walk proposal. The parameter $z_{i,t}^{\gamma}$ is updated using $z_{i,t}^{\gamma+1}=z_{i,t}^{\gamma}+\gamma^{-\eta}(\alpha_i-\hat{\alpha})$.

Table 2
Equity return data.
Source: Goyal and Welch (2008).

Name	Description
B/M	Ratio of book to market value for the Dow Jones Industrial Average
TBL	3m Treasury Bill: Secondary Market Rate
LTY	Difference between the long term yield on government bonds and treasury bill
NTIS	Ratio of 12m moving sums of net issues by NYSE listed stocks to total year end market cap
INFL	Consumer Price Index
LTR	Long term government bond yield
SVAR	Sum of squared daily returns of S&P500
CSP	Cross-sectional beta premium (relative valuation of high and low beta firms)
D/Y	Dividend yield: difference between the log of dividends and the log of lagged prices (S&P500)
EPR	Earnings price ratio: difference between the log of earnings and the log of prices (S&P500)
DE	Dividend payout ratio: difference between the log of dividends and the log of earnings (S&P 500)
DFY	Default yield spread: difference between BAA and AAA-rated corporate bond yields
DRF	Default return spread: difference between long term corporate bond and long term government bond returns

Updating σ^2

The full conditional distribution of σ_t^2 follows a generalised inverse Gaussian distribution which has density

$$\frac{(c/d)^{h/2}}{2K_h(\sqrt{cd})}(\sigma_t^2)^{h-1}\exp\left\{-\frac{1}{2}\left(c\sigma_t^2+\frac{d}{\sigma_t^2}\right)\right\},$$

where K_h is a modified Bessel function of the second kind. The parameter values of the full distribution for σ_t^2 are

$$d=\left(y_t-\sum_{i=0}^{m}\beta_{i,t}x_{i,t}\right)^2,\quad t=1,\dots,T,$$

$$c=\begin{cases}\dfrac{2\lambda^\sigma}{\mu^\sigma(1-\rho^\sigma)}, & t=1,T\\ 2\dfrac{\lambda^\sigma+\rho^\sigma\lambda^\sigma}{\mu^\sigma(1-\rho^\sigma)}, & 1<t<T\end{cases}$$

and

$$h=\begin{cases}\kappa_t^\sigma+\lambda^\sigma-0.5, & t=1\\ \kappa_t^\sigma+\kappa_{t-1}^\sigma+\lambda^\sigma-0.5, & 1<t<T\\ \kappa_{t-1}^\sigma+\lambda^\sigma-0.5, & t=T.\end{cases}$$

Updating κ^σ

The full conditional distribution of κ_t^σ is proportional to

$$\left(\frac{\lambda^\sigma}{(1-\rho^\sigma)\mu^\sigma}\right)^{\lambda^\sigma+\kappa_t^\sigma}\left(\frac{\sigma_t^2\lambda^\sigma\rho^\sigma}{(1-\rho^\sigma)\mu^\sigma}\right)^{\kappa_t^\sigma}(\sigma_{t+1}^2)^{\lambda^\sigma+\kappa_t^\sigma-1}\times\frac{1}{\kappa_t^\sigma!\Gamma(\lambda^\sigma+\kappa_t^\sigma)}$$

for $1\le t\le T-1$. We update this parameter using an adaptive Metropolis–Hastings random walk step. At the γth iteration, the proposed value $\kappa_t^{\sigma\prime}=\kappa_t^{\sigma\gamma}+d\epsilon$ where $p(d=-1)=p(d=1)=1/2$ and $\epsilon\sim\text{Ge}\left(1/(1+z_t^{\sigma\gamma})\right)$. The proposed value is rejected is $\kappa_t^{\sigma\prime}<0$ and otherwise accepted using the standard Metropolis–Hastings acceptance probability for a random walk proposal. The parameter $z_t^{\sigma\gamma}$ is updated using $z_t^{\sigma\gamma}=z_t^{\sigma\gamma}+\gamma^{-\eta}(\alpha_\gamma-\hat{\alpha})$.

Updating θ, θ^σ and β

At each iteration, a vector $s_1,\dots,s_m$ is generated where $p(s_i=1)=\frac{5}{m}$ and $p(s_i=0)=1-\frac{5}{m}$ for $i=1,\dots,m$. Let $\beta_M=\{\beta_i|s_i=1\}$ and $\beta_C=\{\beta_j|s_j=0\}$. The parameter θ and θ^σ are updated conditional on β_C and marginalising over β_M. The value of $p(y|X,\psi,\sigma^2,\varphi,\beta_C)$ can be easily calculated using the Kalman filter.

The parameter $\{\theta_j|s_j=1\}$ are update in turn using a Gibbs sampler. The transformation $\zeta_i=(\log\lambda_i,\log\mu_i,\log\rho_i-\log(1-\rho_i),\log\varphi_i-\log(1-\varphi_i))$ is used with an ASWAM step. The proposed values of ψ_i and κ_i are denoted ψ_i' and κ_i' and are simulated using the method of Section 3. The acceptance ratio for the Metropolis–Hastings algorithm is

$$\alpha_\theta=\min\left\{1,\frac{p\left(y|X,\psi',\sigma^2,\varphi,\beta_C\right)p(\theta)q(\theta',\theta)}{p\left(y|X,\psi,\sigma^2,\varphi,\beta_C\right)p(\theta)q(\theta,\theta')}\right\}.$$

The parameters θ^σ are also updated using an ASWAM step with the transformation $\zeta^\sigma=(\log\lambda^\sigma,\log\mu^\sigma,\log\rho^\sigma-\log(1-\rho^\sigma))$. The proposed values of σ^2 and κ^σ are denoted $\sigma^{2\prime}$ and $\kappa^{\sigma\prime}$ and are sampled using the method of Section 3. The acceptance ratio for the Metropolis–Hastings algorithm is

$$\alpha_{\theta^\sigma}=\min\left\{1,\frac{p\left(y\left|X,\psi,\sigma^{2\prime},\varphi,\beta_C\right.\right)p(\theta^\sigma)q\left(\theta^{\sigma\prime},\theta^\sigma\right)}{p\left(y\left|X,\psi,\sigma^2,\varphi,\beta_C\right.\right)p(\theta^\sigma)q\left(\theta^\sigma,\theta^{\sigma\prime}\right)}\right\}.$$

Updating $\mu^\star$

The full conditional density of $\mu^\star$ is proportional to

$$(\mu^\star+2b^\star)^{-3}\left(\frac{\lambda^\star}{\mu^\star}\right)^{m\lambda^\star}\exp\left\{-\frac{\lambda^\star}{\mu^\star}\sum_{i=1}^{m}\mu_i\right\}.$$

The parameter can be updated using an AMHRW step where the proposed value $\mu^{\star\prime}$ is simulated according to $\log\mu^{\star\prime}\sim\text{N}\left(\log\mu^\star,\tau^{\mu^\star}\right)$ with the variance updated as in (4).

Updating $\lambda^\star$

The full conditional density of $\lambda^\star$ is proportional to

$$\exp\left\{-\lambda^\star/s^\star\right\}\left(\frac{\lambda^{\star\lambda^\star}}{\mu^{\star\lambda^\star}\Gamma(\lambda^\star)}\right)^m\exp\left\{-\frac{\lambda^\star}{\mu^\star}\sum_{i=1}^{m}\mu_i\right\}\prod_{i=1}^{m}\mu_i^{\lambda^\star}.$$

The parameter can be updated using an AMHRW step where the proposed value $\lambda^{\star\prime}$ is simulated according to $\log\lambda^{\star\prime}\sim\text{N}\left(\log\lambda^\star,\tau^{\lambda^\star}\right)$. The variance is updated as in (4).

Appendix B. Data appendix

See Tables 2 and 3.

Table 3
Inflation data. Sources: FRED database, Federal Reserve Bank of St.Louis, University of Michigan Consumer Survey Database, Federal Reserve Bank of Philadelphia, and Institute of Supply Management.

Name	Description
GDP	Difference in logs of real gross domestic product
PCE	Difference in logs of real personal consumption expenditure
GPI	Difference in logs of real gross private investment
RGEGI	Difference in logs of real government consumption expenditure and gross investment
IMGS	Difference in logs of imports of goods and services
NFP	Difference in logs non farm payroll
M2	Difference in logs M2 (commercial bank money)
ENERGY	Difference in logs of oil price index
FOOD	Difference in logs of food price index
MATERIALS	Difference in logs of producer price index (PPI) industrial commodities
OUTPUT GAP	Difference in logs of potential GDP level
GS10	Difference in logs of 10yr Treasury constant maturity rate
GS5	Difference in logs of 5yr Treasury constant maturity rate
GS3	Difference in logs 3yr Treasury constant maturity rate
GS1	Difference in logs 1yr Treasury constant maturity rate
PRIVATE EMPLOYMENT	Log difference in total private employment
PMI MANU	Log difference in PMI-manufacturing index
AHEPNSE	Log difference in average hourly earnings of private non management employees
DJIA	Log difference in Dow Jones Industrial Average Returns
M1	Log difference in M1 (narrow-commercial bank money)
ISM SDI	Institute for Supply Management (ISM) Supplier Deliveries Inventory
CONSUMER	University of Michigan consumer sentiment (level)
UNRATE	Log of the unemployment rate
TBILL3	3m Treasury bill rate
TBILL SPREAD	Difference between GS10 and TBILL3
HOUSING STARTS	Private housing (in thousands of units)
INF EXP	University of Michigan inflation expectations (level)
LAG1, LAG2, LAG3, LAG4	The first, second, third and fourth lags of the response variable

References

Andrieu, C., Thoms, J., 2008. A tutorial on adaptive MCMC. Stat. Comput. 18, 343–373.
Ang, A., Bekaert, G., 2007. Stock return predictability: is it there? Rev. Financ. Stud. 20 (3), 657–707.
Atchadé, Y.F., Rosenthal, J.S., 2005. On adaptive Markov chain Monte Carlo algorithms. Bernoulli 11, 815–828.
Atchade, Y., Fort, G., 2010. Limit theorems for some adapative mcmc algorithms with subgeometric kernels. Bernoullli 16, 116–154.
Atkeson, A., Ohanian, L., 2001. Are Philips Curves Useful for Forecasting Inflation? Federal Reserve Bank of Minneapolis, (Quarterly review).
Belmonte, M., Koop, G., Korobilis, D., 2011. Hierarchical Shrinkage in Time Varying Parameter Models. Universite Catholique de Louvain, (CORE Discussion Papers 2011036).
Bishop, C.M., Tipping, M.E., 2000. Variational relevance vector machines, In: Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence, pp. 46–53.
Canova, F., 2007. G-7 inflation forecasts: random walk, Phillips curve, or what else? Macroeconom. Dynam. 11, 1–30.
Caron, F., Doucet, A., (2008), Sparse Bayesian nonparametric regression, In: Proceedings of the 25th International Conference on Machine Learning, Helsinki, Finland.
Carter, C.K., Kohn, R., 1994. On Gibbs sampling for state space models. Biometrika 81, 541–553.
Carvalho, C., Polson, N., Scott, J., 2010. The horseshoe estimator for sparse signals. Biometrika 97, 465–480.
Chan, J., Koop, G., Leon-Gonzales, R., Strachan, R., 2012. Time varying dimension models. J. Bus. Econom. Statist. 30, 358–367.
Cogley, T., Sargent, T., 2001. Evolving post world war II US inflation dynamics. In: NBER Macroeconomics Annual. MIT Press/NBER, pp. 331–373.
Cogley, T., Sargent, T., 2005. Drifts and volatilities: monetary policies and outcomes in the post world war US. Rev. Econom. Dynam. 8 (2), 262–302.
Fernandez, C., Ley, E., Steel, M., 2001. Model uncertainty in cross-country growth regressions. J. Appl. Econom. 16, 563–576.
Frühwirth-Schnatter, S., 1994. Data augmentation and dynamic linear models. J. Time Ser. Anal. 15, 183–202.
Frühwirth-Schnatter, S., Wagner, H., 2010. Stochastic model specification search for Gaussian and partial non-Gaussian state space models. J. Econometrics 154, 85–100.
Gerlach, R., Carter, C., Kohn, R., 2000. Efficient Bayesian inference in dynamic mixture models. J. Amer. Statist. Assoc. 95, 819–828.
Gneiting, T., 2011. Making and evaluating point forecasts. J. Amer. Statist. Assoc. 106, 746–762.
Gourieroux, C., Jasiak, J., 2006. Autoregressive gamma processes. J. Forecast. 25, 129–152.
Goyal, A., Welch, I., 2008. A comprehensive look at the empirical performance of equity premium prediction. Rev. Financ. Stud. 21, 1455–1508.
Griffin, J., Brown, P., 2010. Inference with normal-gamma prior distributions in regression models. Bayesian Anal. 5, 171–188.
Griffin, J.E., Stephens, D.A., 2013. Advances in Markov chain Monte Carlo. In: Damien, P., Dellaportas, P., Polson, N.G., Stephens, D.A. (Eds.), Bayesian Theory and Applications. pp. 104–144.
Groen, J., Paap, R., Ravazzolo, F., 2009. Real Time Inflation Forecasting in A Changing World. Technical Report 19. Economic Institute, Erasmus University, Rotterdam.
Haario, H., Saksman, E., Tamminen, J., 2001. An adaptive Metropolis algorithm. Bernoulli 7, 223–242.
Koop, G., Korobilis, D., 2012. Forecasting inflation using dynamic model averaging. Internat. Econom. Rev. 53, 867–886.
Lettau, M., Van Nieuwerburgh, S., 2008. Reconciling the return predictability evidence. Rev. Financ. Stud. 21, 1607–1652.
Papaspiliopoulos, O., Roberts, G.O., Skold, M., 2007. A general framework for the parameterization of hierarchical models. Statist. Sci. 22 (1), 59–73.
Park, T., Casella, G., 2008. The Bayesian Lasso. J. Amer. Statist. Assoc. 103, 672–680.
Paye, B., Timmermann, A., 2006. Instability of return prediction models. J. Empir. Finance 13, 274–315.
Pitt, M., Chatfield, C., Walker, S., 2002. Constructing first order stationary autoregressive models via latent processes. Scand. J. Stat. 29 (4), 657–663.
Pitt, M., Walker, S., 2005. Constructing stationary time series models using auxiliary variables with applications. J. Amer. Statist. Assoc. 100 (470), 554–564.
Polson, N.G., Scott, J.G., 2011. Shrink globally, act locally: sparse Bayesian regularization and prediction. In: Bernardo, J.M., Bayarri, M.J., Berger, J.O., Dawid, A.P., Heckerman, D., Smith, A.F.M., West, M. (Eds.), Bayesian Statistics, Vol. 9. Clarendon Press, pp. 501–538.
Primicery, G., 2005. Time varying structural vector autoregressions and monetary policy. Rev. Econom. Stud. 72, 821–852.
Raftery, A., Karny, M., Ettler, P., 2010. Online prediction under model uncertainty via dynamic model averaging: application to a cold rolling mill. Technometrics 52, 52–66.
Roberts, G.O., Rosenthal, J.S., 2007. Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms. J. Appl. Probab. 44, 458–475.
Roll, R., Ross, S., 1994. On the cross sectional relation between expected returns and betas. J. Finance 49, 101–121.
Sims, C., 1980. Macroeconomics and reality. Econometrica 48 (1), 1–48.
Stock, J., Watson, M., 1996. Evidence on structural instability in macroeconomic time series relations. J. Bus. Econom. Statist. 14, 11–30.
Tipping, M.E., 2000. The relevance vector machine. In: Advances in Neural Information Processing Systems, Vol. 12. Cambridge, US, pp. 652–658.
West, M., Harrison, J., 1999. Bayesian forecasting and dynamic models. In: Springer Series in Statistics, second ed. Springer.