STRATHCLYDE

DISCUSSION PAPERS IN ECONOMICS

TIME VARYING DIMENSION MODELS

BY

JOSHUA C.C. CHAN, GARY KOOP, ROBERT LEON-GONZALEZ AND RODNEY W. STRACHAN

NO. 11-16

DEPARTMENT OF ECONOMICS
UNIVERSITY OF STRATHCLYDE
GLASGOW

Time Varying Dimension Models

Joshua C.C. Chan
Australian National University

Gary Koop
University of Strathclyde

Roberto Leon-Gonzalez
National Graduate Institute for Policy Studies

Rodney W. Strachan
Australian National University

May 11, 2010

Abstract: Time varying parameter (TVP) models have enjoyed an increasing popularity in empirical macroeconomics. However, TVP models are parameter-rich and risk over-fitting unless the dimension of the model is small. Motivated by this worry, this paper proposes several Time Varying dimension (TVD) models where the dimension of the model can change over time, allowing for the model to automatically choose a more parsimonious TVP representation, or to switch between different parsimonious representations. Our TVD models all fall in the category of dynamic mixture models. We discuss the properties of these models and present methods for Bayesian inference. An application involving US inflation forecasting illustrates and compares the different TVD models. We find our TVD approaches exhibit better forecasting performance than several standard benchmarks and shrink towards parsimonious specifications.

Acknowledgements: Koop, Leon-Gonzalez and Strachan are Fellows of the Rimini Centre for Economic Analysis. We would like to thank the Leverhulme Trust and the Australian Research Council for financial support under Grant F/00 273/J and Grant DP0987170, respectively. Contact address: Gary Koop, Department of Economics, Strathclyde University, Glasgow, G4 0GE, United Kingdom, Gary.Koop@strath.ac.uk

1 Introduction

It is common for researchers to model variation in coefficients in time series models using state space methods. If, for $t = 1, .., T$, y_t is an $n \times 1$ vector of observations on the dependent variables, Z_t is an $n \times m$ matrix of observations on explanatory variables and θ_t is an $m \times 1$ vector of states, then such a state space model can be written as:

$$\begin{aligned} y_t &= Z_t \theta_t + \varepsilon_t \\ \theta_{t+1} &= \theta_t + \eta_t, \end{aligned} \tag{1}$$

where ε_t is $N(0, H_t)$ and η_t is $N(0, Q_t)$. The errors, ε_t and η_t, are assumed to be independent (at all leads and lags and of each other). This framework can used to estimate time-varying parameter (TVP) regression models, variants of which are commonly-used in macroeconomics (e.g., Groen, Paap and Ravazzolo, 2009, Koop and Korobilis, 2009). Furthermore, TVP-VARs (see among many others, Canova, 1993, Cogley and Sargent, 2005, D'Agostino, Gambetti and Giannone, 2009 and Primiceri, 2005) are obtained by letting Z_t contain deterministic terms and appropriate lags of the dependent variables, setting $Q_t = Q$ and giving H_t a multivariate stochastic volatility form.

Such TVP models allow for constant gradual evolution of parameters. However, they assume that the dimension of the model is constant over time in the sense that θ_t is always an $m \times 1$ vector of parameters. But there are theoretical and empirical reasons for being interested in TVP models where the dimension of the state vector changes over time. Macroeconomists are often interested in whether restrictions suggested by economic theory hold. For instance, Staiger, Stock and Watson (1997), show how, if the Phillips curve is vertical, a certain restriction is imposed on a particular regression involving inflation and unemployment. Koop, Leon-Gonzalez and Strachan (2009a) investigate this restriction in a TVP regression model and find that the probability that it holds varies substantially over time. As another example, consider the VARs of Amato and Swanson (2001) where interest centers on Granger causality restrictions that imply that money has no predictive power for output or inflation. It is possible (and empirically likely) that restrictions such as these hold at some points in time but not others. In such cases, the researcher would want to work with a TVP model, but where the parameters satisfy restrictions at certain points in time but not at others. In short, there are many theoretical reasons for wanting to work with a time-varying dimension (TVD) model where restrictions which reduce the dimension of the model are imposed only at some points in time.

There are also many econometrically-inspired reasons for being interested in TVD models. For instance, if lag length changes over time, then imposing different lag lengths at different points in time will lead to more precise estimates. In forecasting, the importance of shrinkage has been found in a myriad of studies (e.g. Groen, Paap and Ravazzolo, 2009 or Koop and Korobilis, 2009). In general, TVP models risk being over-parameterized. Allowing for the dimension of the model to change over time is potentially an effective way of reducing over-parameterization worries and ensuring shrinkage while minimizing the risk of model mis-specification.

The desire to work with a TVP model that falls into the familiar class of state space models, but allows for the dimension of the model to change over time motivates the present paper. To our knowledge, there are few existing papers which consider this question. There are, as discussed above, many papers which allow parameters to change over time and adopt state space methods. Furthermore, in previous work (Koop, Leon-Gonzalez and Strachan, 2009a), we have developed methods for calculating the probability that equality restrictions on states hold at any point in time (but without actually imposing the restrictions). Finally, there are some papers, such as Koop and Potter (2009), which develop methods for estimating state space models with inequality restrictions imposed. However, the aim of the present paper is different from all these approaches: we wish to develop methods for estimating models which impose equality restrictions on the states. In other words, the literature has considered the *testing* of *equality* restrictions on states in state space models and *estimation* of states under *inequality* restrictions. But the present paper is one of the few which considers *estimation* of state space models subject to *equality* restrictions on the states (where these restrictions may hold at some points in time but not others). Other papers which adopt different approaches to this problem include the dynamic model averaging approach of Raftery, Karny, Andrysek and Ettler (2007) and Koop and Korobilis (2009) and the combination of stochastic search variable selection methods with TVP models such as in Korobilis (2009). Our TVD models differ from these in that our framework involves the use of dynamic mixture models (see, e.g., Gerlach, Carter and Kohn, 2000) and associated posterior simulation algorithms. Such models have proved popular in several areas of macroeconomics (e.g. Giordani, Kohn and van Dijk, 2007). We consider several new ways of implementing the dynamic mixture approach which lead to models which allow for time-variation in both the parameters and the dimension of the model. We investigate these methods in an empirical application involving forecasting US inflation.

2 Time Varying Dimension Models

The models used in this paper are all dynamic mixture models (see, e.g., Gerlach, Carter and Kohn, 2000 or Giordani and Kohn, 2008). An advantage of adopting a dynamic mixture framework is that efficient methods of posterior simulation are available and well-understood. Thus, our discussion of Bayesian inference in these models can be very brief. It is the structure and justification for our particular dynamic mixture models that must be provided and this is what we do in this section. In the empirical section, we provide precise modelling details (including priors) for a TVD-regression application. But we outline the general ideas here first, since they can be used with other models such as TVD-VARs.

2.1 The Dynamic Mixture Model

The dynamic mixture model of Gerlach, Carter and Kohn (2000) adds to (1) the assumption that any or all of the system matrices, Z_t, Q_t and H_t, depend on an $s \times 1$ vector K_t. As a simple example, suppose θ_t contains regression coefficients, $s = 1$, $K_t \in \{0, 1\}$ and $Q_t = K_t Q$. Such a specification has been used with changepoint models (see, e.g., Giordani and Kohn, 2008 or Koop, León-González and Strachan, 2009b). That is, if $K_t = 0$, then $\theta_{t+1} = \theta_t$ and the regression coefficients do not change, but if $K_t = 1$ they do change.

Gerlach, Carter and Kohn (2000) discuss how this specification results in a mixtures of Normals representation for y_t and, hence, the terminology dynamic mixture model arises. The contribution of Gerlach, Carter and Kohn (2000) is to develop an efficient algorithm for posterior simulation for this class of models. The efficiency gains occur since the states are integrated out and $K = (K_1, .., K_T)'$ is drawn unconditionally (i.e. not conditional on the states). A simple alternative algorithm would involve drawing from the posterior for K conditional on $\theta = (\theta_1', .., \theta_T')'$ and then the posterior for θ conditional on K. Such a strategy can be shown to produce a chain of draws which is very slow to mix. The Gerlach, Carter and Kohn (2000) algorithm requires only that K_t be Markov (i.e. $p(K_t|K_{t-1}, .., K_1) = p(K_t|K_{t-1})$) and is particularly simple if K_t is a discrete random variable.

In this paper, we consider three different ways K_t can enter the system matrices so as to yield a TVD model.

2.2 A First TVD Model

We begin with a TVD model which adapts the approach of Gerlach, Carter and Kohn (2000) in a particular way such that θ_t remains an $m \times 1$ vector at all times, but there is a sense in which the dimension of the model can change over time. Since θ_t remains of full dimension at all times, our claim that the dimension of the model changes over time may sound odd. But we achieve our goal by allowing for explanatory variables to be included/excluded from the likelihood function depending on K_t. The basic idea can be illustrated quite simply in terms of (1). Suppose $Z_t = K_t z_t$ where z_t is an explanatory variable and $K_t \in \{0, 1\}$. If $K_t = 0$ then z_t does not enter the likelihood function and the coefficient θ_t does not enter the model. But if $K_s = 1$, then the coefficient θ_s does enter the model. Thus, the dimension of the model is different at time t than at time s.

An interesting and sensible implication of this specification can be seen by considering what happens if a coefficient is omitted from the model for h periods, but then is included again. That is, suppose we have $K_{t-1} = 1$,

$$K_t = K_{t+1} = ... = K_{t+h-1} = 0$$

but $K_{t+h} = 1$ and further assume $Q_t = Q$. Then (1) implies:

$$E\left(\theta_{t+h}\right) = \theta_{t-1}$$

but

$$var\left(\theta_{t+h}\right) = hQ.$$

In words, if an explanatory variable drops out of the model, but then reappears h periods later, then your best guess for its value is what it was when it was last in the model. However, the uncertainty associated with your best guess increases the longer the coefficient has been excluded from the model (since the variance increases with h).

It is worth stressing that, if $K_t = 0$ then θ_t does not enter the likelihood and, thus, it is not identified in the likelihood. However, because the state equation provides an informative hierarchical prior for θ_t, it will still have a proper posterior. To make this idea clear, let us revert to a general Bayesian framework. Suppose we have a model depending on a vector of parameters θ which are partitioned as $\theta = (\phi, \gamma)$. Suppose the prior is $p(\theta) = p(\phi, \gamma) = p(\gamma) p(\phi|\gamma)$ and the likelihood is $L(y|\theta)$. Now consider a second model which imposes the restriction that $\phi = 0$. Instead of directly imposing the restriction $\phi = 0$, consider what happens if we impose the restriction that ϕ does not enter the likelihood. That is, the likelihood for the second model is $L(y|\theta) = L(y|\gamma)$ and its posterior is

$$p(\theta|y) = \frac{L(y|\theta) p(\theta)}{\int L(y|\theta) p(\theta) d\theta} = \frac{L(y|\gamma) p(\gamma)}{\int L(y|\gamma) p(\theta) d\theta} p(\phi|\gamma) = p(\gamma|y) p(\phi|\gamma).$$

Since $p(\phi|\gamma)$ integrates to one (or assigns a point mass to $\phi = 0$) integrating $p(\theta|y)$ with respect to ϕ provides us with a valid posterior for the second model and the integral $\int L(y|\gamma) p(\theta) d\theta$ will result in the correct marginal likelihood. This is the strategy which underlies and justifies our approach.

Our TVD approach can be used with a wide range of time series models and with a wide range of restrictions on the coefficients. In our empirical section we describe a particular implementation of relevance for the TVD regression model. This simply defines K_t as being a vector of dummy variables which includes/excludes each explanatory variable. However, for other models slight differences in implementation might be appropriate. For instance, in the TVD-VAR, K_t could be restricted so that lag length can change only in a sequential manner.

2.3 A Second TVD Model

To explain our second approach to TVD modelling, we return to our general notation for state space models given in (1). The state equation can be interpreted as a hierarchical prior for θ_{t+1}, expressing a prior belief that it is similar to θ_t. In the empirical macroeconomics literature (see, among many others, Ballabriga, Sebastian and Valles, 1999, Canova and Ciccarelli, 2004, and Canova, 2007), there is a desire to combine such prior information with prior information of other sorts (e.g. the Minnesota prior). This can be done by replacing (1) by

$$\begin{aligned} y_t &= Z_t \theta_t + \varepsilon_t \\ \theta_{t+1} &= M\theta_t + (I-M)\,\overline{\theta} + \eta_t, \end{aligned} \tag{2}$$

where M is an $m \times m$ matrix, $\overline{\theta}$ is an $m \times 1$ vector and η_t is $N(0, Q_t)$. For instance, Canova (2007) sets $\overline{\theta}$ and Q_t to have forms based on the Minnesota prior and sets $M = gI$ where g is a scalar. If $g = 1$, then the traditional TVP-VAR prior is obtained, but as g decreases we move towards the Minnesota prior.

In the TVD model, alternative choices for M, $\overline{\theta}$ and Q_t suggest themselves. In particular, our second TVD model sets $\overline{\theta} = 0_m$,[1] M becomes M_t which is a diagonal matrix with diagonal elements $K_{tj} \in \{0, 1\}$ and $Q_t = M_t Q$. This model has the property that, if $K_{jt} = 1$ then the j^{th} coefficient is evolving according to a random walk in standard TVP-regression fashion. But if $K_{jt} = 0$, then the j^{th} coefficient is set to zero, thus reducing the dimension of the model.

To understand the implications of this specification for K_t, consider the simplest case where $m = 1$ and, thus θ_t and K_t are scalars and see what happens if a coefficient is omitted from the model for h periods. That is, suppose we have $K_{t-1} = 1$,

$$K_t = K_{t+1} = ... = K_{t+h-1} = 0$$

but $K_{t+h} = 1$. In this case, (2) implies:

$$E(\theta_{t+h}) = \overline{\theta}$$

but

$$var(\theta_{t+h}) = Q.$$

In words, in contrast to our first TVD model, our second TVD model implies that, if a coefficient drops out of the model, but then reappears h periods later, then your best guess for its value is 0 and the uncertainty associated with your best guess is Q (regardless of how long the coefficient has been

[1] If the dependent variable is in levels and the explanatory variables include lagged dependent variables, then the researcher may wish to set the element of $\overline{\theta}$ corresponding to the first lag to one, to reflect the common belief in random walk behavior.

excluded from the model). Thus, there is more shrinkage in this model than in our first TVD model and (in contrast to the first TVD model) it will always be shrinkage towards zero (assuming $\overline{\theta} = 0$). It is an empirical question as to whether this specification is more appropriate than the specification of the first TVD model.

As with our first TVD model, the posterior simulation algorithm of Gerlach, Carter and Kohn (2000) can be used directly and requires no further discussion here. Further details of how we implement this model are provided in the empirical section.

2.4 A Third TVD Model

To justify our third approach to TVD modelling, we begin by discussing the TVP-SUR approach of Chib and Greenberg (1995) which has been used in empirical macroeconomics in papers such as Ciccarelli and Rebucci (2002). If we return to our general notation for state space models in (1), the model of Chib and Greenberg (1995) adds another layer to the hierarchical prior:

$$\begin{aligned} y_t &= Z_t\theta_t + \varepsilon_t \\ \theta_{t+1} &= M\beta_{t+1} + \eta_t, \\ \beta_{t+1} &= \beta_t + u_t. \end{aligned} \tag{3}$$

where the assumptions about the errors are described after (1) with the additional assumptions that u_t is i.i.d. $N(0, R)$ and u_t is independent of the other errors in the model. Note that β_t can potentially be of lower dimension than θ_t, which is another avenue the researcher can use to achieve parsimony.

Note first that this specification retains the random walk evolution of the VAR coefficients since it can be written as:

$$\begin{aligned} y_t &= Z_t\theta_t + \varepsilon_t \\ \theta_{t+1} &= \theta_t + v_t, \end{aligned} \tag{4}$$

where $v_t = Mu_t + \eta_t - \eta_{t-1}$. In this sense, the difference between (1) and (4) is that the state equation errors of the latter have a particular MA(1) structure. However, if M is a square matrix, the hierarchical prior in (3) expresses the conditional prior belief that

$$E\left(\theta_{t+1}|\theta_t\right) = M\theta_t$$

and, thus, is a combination of the random walk prior belief of the conventional TVP model with the prior beliefs contained in M. M is typically treated as known.

Our third TVD model can be constructed by specifying M and Q_t to be exactly as in our second TVD model.

To understand the properties of the third TVD model, we can consider the same example as used previously (where a coefficient drops out of the model for h periods and then re-enters it). Remember that, in this case, the first TVD model implied $E\left(\theta_{t+h}\right) = \theta_{t-1}$ and $var\left(\theta_{t+h}\right) = hQ$ while the second TVD model implied $E\left(\theta_{t+h}\right) = 0$ and $var\left(\theta_{t+h}\right) = Q$. The third TVD model can be seen to have properties closer to those of the first approach and yields $E\left(\theta_{t+h}\right) = \beta_{t-1}$ and $var\left(\theta_{t+h}\right) = hR$ (if M is a square matrix).

The first and third TVD models, thus, can be seen to have similar properties. However, they differ in one important way. Remember that the first TVD model did not formally reduce the dimension of θ_t in that all of its elements were unrestricted (it constructed K_t in such a way so that some elements of θ_t did not enter the likelihood function). The third TVD model does formally reduce the dimension of θ_t since it allows for some of its elements at some points of time to be restricted to zero.

2.5 Posterior Computation in the TVD Models

The advantage of the TVD modelling framework outlined in this paper is that existing methods of posterior computation can be used to set up a fast and efficient Markov Chain Monte Carlo (MCMC) algorithm. Thus we can deal with computational issues quickly. For all our models, K is drawn using the algorithm described in Section 2 of Gerlach, Carter and Kohn (2000). Note that this algorithm draws K conditional on all the model parameters except for θ. The fact that θ is integrated out analytically greatly improves the efficiency of the algorithm. We draw θ (conditional on all the model parameters, including K) using the algorithm of Chan and Jeliazkov (2009), although any of the standard algorithms for drawing states in state space models (e.g. Carter and Kohn, 1994 or Durbin and Koopman, 2002) could be used. All our models have stochastic volatility and to draw the volatilities and all related parameters we use the algorithm of Section 3 of Kim, Shephard and Chib (1998). The remaining parameters are the error variances in the state equations and the parameters characterizing the hierarchical prior for K. These can be drawn using standard methods as discussed below in the context of the exact implementation of each approach.

3 Forecasting US Inflation

3.1 Data

To investigate the properties of the TVD models, we use a TVD regression model and investigate how the various approaches work in an empirical exercise involving US inflation forecasting. The literature on inflation forecasting is a voluminous one. Here we note only that there have been many papers which use regression-based methods in recursive or rolling forecast exercises (e.g. Ang, Bekaert and Wei, 2007 and Stock and Watson, 2007, 2008) and that recently papers have been appearing using TVP models for forecasting (e.g. D'Agostino, Gambetti and Giannone, 2009).

Our data runs from 1960:Q1 to 2008:Q2 and we use CPI inflation as our dependent variable. The explanatory variables are two lags of the dependent variable and the predictors listed in Table 1.[2]

Table 1: Predictors for Inflation

x_1	EMPLOY: the percentage change in employment
x_2	TBILL: three month Treasure bill rate
x_3	SPREAD: the spread between the 10 year and 3 month Treasury bill rates
x_4	MONEY: the percentage change in the money supply
x_5	INFEXP: University of Michigan measure of inflation expectations

We apply our TVD approach to the question of which of these explanatory variables is a good predictor for inflation at each point of time. Note that this means K_t is a vector of five dummy variables. At each point in time there are 2^5 values K_t could take. Although each of our TVD approaches defines a model, there is a sense in which they can be interpreted as automatically doing model averaging over a model space of 2^{5T} models (i.e.

[2] With the exception of the inflation expectations variable (which was obtained from the University of Michigan) all data was obtained from the FRED data base of the Federal Reserve Bank of St. Louis.

at each point in time each possible configuration of K_t can be thought of as defining a "model"). A related literature in the field of dynamic model averaging (see, e.g., Raftery et al, 2007 or Koop and Korobilis, 2009) also faces the computationally daunting task of working with model spaces of this order of magnitude. The dynamic model averaging literature typically uses approximate methods to make the computational burden manageable. Our TVD approaches can be thought of an alternative way of dealing with this problem which does not resort to approximations. However, it is the case that the computational burden can be overwhelming if K_t is of high dimension and the $K_{j,t}$ are, a priori, independent of one another. In such cases, the researcher may wish to put more structure on the hierarchical prior for K_t. For instance, in an AR(d) model an unrestricted approach would lead to K_t taking on 2^d possible values. But if we define a hierarchical prior which restricts K_t such that lags appear sequentially, this reduces to d the number of possible values K_t can take.

3.2 Details of the First TVD Model

Write the TVD regression model as:

$$y_t = \phi_{0,t} + \sum_{j=1}^{d} y_{t-j}\phi_{j,t} + \sum_{j=1}^{p} K_{j,t}x_{j,t-1}\gamma_{j,t} + \varepsilon_t, \tag{5}$$

where $K_{j,t} \in \{0,1\}$ is a binary variable that determines whether explanatory variable $x_{j,t-1}$ is included in the regression, and $\varepsilon_t \sim N(0, \exp(h_t))$.

Rewrite (5) as

$$y_t = w_t'\phi_t + (\widetilde{M}_t x_{t-1})'\gamma_t + \varepsilon_t, \tag{6}$$

where $w_t = (1, y_{t-1}, \ldots, y_{t-d})'$, $\phi_t = (\phi_{0,t}, \ldots, \phi_{d,t})'$, $\widetilde{M}_t = \mathrm{diag}(K_{1,t}, \ldots, K_{p,t})$, $x_{t-1} = (x_{1,t-1}, \ldots, x_{p,t-1})'$ is a $p \times 1$ vector of explanatory variables, and $\gamma_t = (\gamma_{1,t}, \ldots, \gamma_{p,t})'$. The states, $\theta_t = (\phi_t', \gamma_t')'$ are assumed to evolve according to a random walk as in (1) and we assume $\theta_1 \sim N(\theta_0, D)$ with relatively noninformative hyperparameter choices of $\theta_0 = 0$ and $D = 5 \times I$. This strategy of subjectively choosing proper but relatively noninformative priors is used for all of the parameters in all of our models.

We do not restrict the range of values that $K_t = (K_{1,t}, \ldots, K_{p,t})$ will take. Hence, it can take on 2^p values: $K_t \in \mathcal{I} = \{0,1\}^p$. But, we impose a Markov hierarchical prior which expresses the belief that, with probability

c the model will stay with its current set of explanatory variables and with probability $1-c$ it will switch to a new model. A priori, all of the 2^p-1 possible new models are equally likely. Thus we have:

$$\begin{aligned} \Pr(K_{t+1}=i \mid K_t=i) &= c, \qquad\qquad i \in \mathcal{I} \\ \Pr(K_{t+1}=j \mid K_t=i) &= \frac{1-c}{2^p-1}, \quad i \neq j, \quad i,j \in \mathcal{I} \end{aligned} \tag{7}$$

for $t=1,\ldots,T-1$. We assume that the prior for c follows a beta distribution with parameters $c_1^0=1.76$ and $c_2^0=2$, such that $E\left(c\right)=0.47$. With this assumption, the conditional posterior distribution is also beta (see, e.g., page 84 of Chib, 1996).

We also must specify a prior for the initial values, $K_{1,1},\ldots,K_{p,1}$. Each of these is, a priori, assumed to be an independent Bernoulli random variable: $\Pr(K_{j,1}=1)=b_j$, $j=1,\ldots,p$, where b_j has a beta distribution with hyperparameters $b_1^0=1.5$ and $b_2^0=1.5$, such that $E\left(b_j\right)=0.5$. The posterior for b_j (conditional on K) can also be calculated as described on page 84 of Chib (1996) and the parameters of the beta posterior are $K_{j,1}+b_1^0$ and $1-K_{j,1}+b_2^0$.

Next we assume $\eta_t \sim N(0,Q)$ where Q is a diagonal matrix. Each diagonal element of $Q=\operatorname{diag}(q_1,\ldots,q_{d+p+1})$ is assumed to follow, independently, an inverse-gamma distribution: $q_j \sim IG(\nu_j/2, S_j/2)$ with $\nu_1=\ldots=\nu_{d+p+1}=6$ and $S_1=\ldots=S_{d+p+1}=0.002$. The posterior for q_j (conditional on the states) then takes a familiar inverse-gamma form (see, e.g., Koop, 2003, page 201).

Finally, we assume a mean-reverting stochastic volatility process for $h_t=\ln\left(H_t\right)$:

$$h_{t+1}=\mu+\rho(h_t-\mu)+v_t, \tag{8}$$

for $t=1,\ldots,T-1$, where $v_t \sim N(0,\sigma_v^2)$. We restrict the log-volatility process to be stationary and impose $|\rho|<1$ with h_1 drawn from the stationary distribution, i.e., $h_1 \sim N(\mu, \sigma_v^2/(1-\rho^2))$.

The prior for μ is assumed to be non-informative, i.e., $p(\mu)$ is proportional to a constant. Following Kim, Shephard and Chib (1998), the prior for ρ is given by

$$\log p(\rho) \propto (\underline{\rho}_1-1)\log\left(\frac{1+\rho}{2}\right)+(\underline{\rho}_2-1)\log\left(\frac{1-\rho}{2}\right), \quad |\rho|<1,\ \underline{\rho}_1,\underline{\rho}_2>\frac{1}{2},$$

with $\underline{\rho}_1 = 20$ and $\underline{\rho}_2 = 1.5$, giving a prior mean of 0.86. The prior for σ_v^2 is assumed to be inverse-gamma: $\sigma_v^2 \sim IG(\nu_h/2, S_h/2)$, where $\nu_h = 6$ and $S_h = 0.04$. All of these prior hyperparameter choices are relatively noninformative but sensible, given the units of measurement of the data.

3.3 Details of the Second TVD Model

In our second approach to TVD modelling we work with:

$$y_t = \phi_{0,t} + \sum_{j=1}^{d} y_{t-j}\phi_{j,t} + \sum_{j=1}^{p} x_{j,t}\gamma_{j,t} + \varepsilon_t, \tag{9}$$

which can be rewritten as

$$y_t = w_t'\phi_t + x_{t-1}'\gamma_t + \varepsilon_t. \tag{10}$$

As suggested by (2), we assume $\theta_t = (\phi_t', \gamma_t')'$ evolves as follows:

$$\theta_{t+1} = M_{t+1}\theta_t + M_{t+1}\eta_t \tag{11}$$

for $t = 1, \ldots, T-1$, where $M_t = \text{diag}(\iota_{d+1}', K_{1,t}, \ldots, K_{p,t})$, ι_{d+1} is an $(d+1)\times 1$ column of ones. The initial condition θ_1 is modeled as $\theta_1 \sim N(M_1\theta_0, M_1DM_1)$, where θ_0 and D are known constants selected as in our first approach.

All other modelling assumptions are as for our first approach. Most importantly, our hierarchical prior for K is as specified in (7).

3.4 Details of the Third TVD Model

For the last formulation, we assume the same measurement equation given in (10), but alter the state equations as in (3) to:

$$\gamma_t = M_t\beta_t + M_t v_t, \tag{12}$$

where $M_t = \text{diag}(K_{1,t}, \ldots, K_{p,1})$ and

$$\widetilde{\theta}_{t+1} = \widetilde{\theta}_t + w_t, \tag{13}$$

where $\widetilde{\theta}_t = (\phi_t', \beta_t')'$ for $t = 1, \ldots, T-1$ and $\widetilde{\theta}_1 \sim N(\widetilde{\theta}_0, D_\theta)$ where $\widetilde{\theta}_0 = 0$, $D_\theta = 5 \times I$. Furthermore, we have $v_t \sim N(0, R_1)$ and $w_t \sim N(0, R_2)$ where

$R_1 = \text{diag}(r_{1,1}, \ldots, r_{1,p})$ and $R_2 = \text{diag}(r_{2,1}, \ldots, r_{2,d+p+1})$ are diagonal matrices. Each diagonal element of R_1 and R_2 is assumed to follow, independently, an inverse-gamma distribution: $r_{i,j} \sim IG(\nu/2, S/2)$ with $\nu = 6$ and $S = 0.002$ for all i and j. All other modelling details are as for the other two TVD approaches. With regards to posterior simulation, note that we draw $\widetilde{\theta}_t$ and γ_t (conditional on the other parameters) by first drawing $\widetilde{\theta}_t$ marginal of γ_t and then γ_t given $\widetilde{\theta}_t$. This differs from the approach of Chib and Greenberg (1995) who used the posterior of $\widetilde{\theta}_t$ conditional on γ_t.

3.5 Benchmark Models for Comparison

In addition to the three TVD models, we consider four benchmarks for comparison. These are a TVP model with stochastic volatility, two constant coefficient regression models (with and without stochastic volatility) and simple random walk forecasts. In both of the constant coefficients models we use a $N(0, 5I)$ prior for the regression coefficients (which is the same as the prior for the initial conditions in the TVD models). For the version with stochastic volatility we use the same specification as with the TVD models. For the homoskedastic version, the error variance has an $IG(3, 0.05)$ prior.

In order to make sure all our approaches are as comparable as possible, our TVP regression model is exactly the same as our TVD models (including having the same prior for all common parameters) except that we set $K_{jt} = 1$ for all j and t.

3.6 Results: Estimation using the Full Sample

Remember that all of our models are specified so that $K_{j,t} = 1$ if the j^{th} predictor is included. Thus, the probability that $K_{j,t} = 1$ sheds light on whether a predictor is included or excluded. In the latter case, a more parsimonious model is achieved. Figures 1 through 5 plot $p(K_{j,t} = 1 \,|\, y)$ for our three TVD models for $j = 1, .., 5$.

Note first that these figures show that parsimony is being achieved. For EMPLOY ($j = 1$) and SPREAD ($j = 3$), $p(K_{j,t} = 1 \,|\, y)$ is less than 0.5 for most or all the time. And for some of the other variables and other approaches, this probability is small for appreciable amounts of the time. Although it is interesting to note that (except for the first TVD approach) it is rare for $p(K_{j,t} = 1 \,|\, y)$ to be very close to zero. This suggests that shrinkage is being achieved, but not by completely excluding a predictor from the model. For instance, $p(K_{j,t} = 1 \,|\, y) = 0.2$ does not totally exclude a predictor from the model, but shrinks its effect towards zero.

Next consider the time varying aspect of the TVD models. Clearly there is substantial time variation in many of the lines in Figures 1 through 5, indicating that the impact of the corresponding predictor is changing over time. We can also be sure any patterns are not an artifact of the statistical methodology, since the patterns are so different in the different figures. For instance, one might expect that the lines in the figures would uniformly tend to increase over time since the available data is increasing, leading to more "significant" coefficients. This is not the case. Although there are some cases where $p(K_{j,t} = 1 \,|\, y)$ is increasing over time, there are many where it is not. Similarly, although our use of hierarchical priors will ensure shrinkage, the patterns in the figures do not solely reflect this shrinkage since some are shrunk much more than others. In short, these figures indicate that our TVD models are capturing time variation in the coefficients in a sensible and automatic fashion.

Next let us compare our three different TVD models. Broadly speaking, they are yielding similar results. However, our first TVD model is sometimes a bit different from the other two. It exhibits less time variation and, loosely speaking, tends to either include a variable or exclude it. This is not surprising in light of the properties of this model (see Section 2.2). That is, the longer a predictor is excluded from the model, the larger the variance in the state equation prior becomes (and the less shrinkage applies). This could be an attractive feature if there are big structural breaks for the coefficients. But, in our data set, this is a less attractive property. The third TVD approach shares these properties with the first TVD model, but this is partly counteracted by the added dimension reduction noted in Section 2.4. However, these considerations suggest that the second TVD approach might be most suitable for dealing with data sets with frequent small breaks.

Finally, it is hazardous to link economic stories based on reduced form regression results such as ours. But one point is worth noting. Figure 5 presents results for the inflation expectations variable. Previous studies (e.g. Koop and Korobilis, 2009) have found this to be a good predictor for inflation, but only in recent times (e.g. after the early 1980s). Before the Great Moderation, inflation was high and volatile and surveys of inflation expectations were often poor predictors of inflation. But subsequently, agents found it much easier to form accurate expectations of future inflation. Our findings in Figure 5 are consistent with this story (especially for our second and third

TVD models).

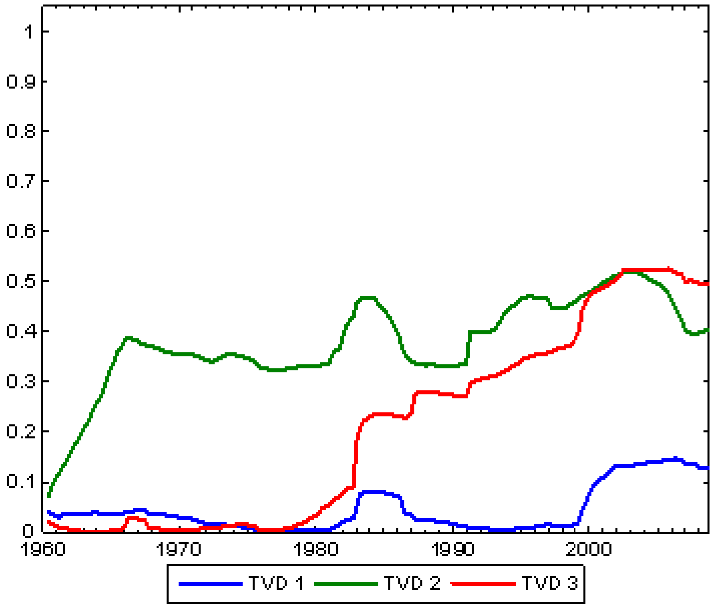

Figure 1: $p(K_{1,t} = 1 \,|\, y)$.

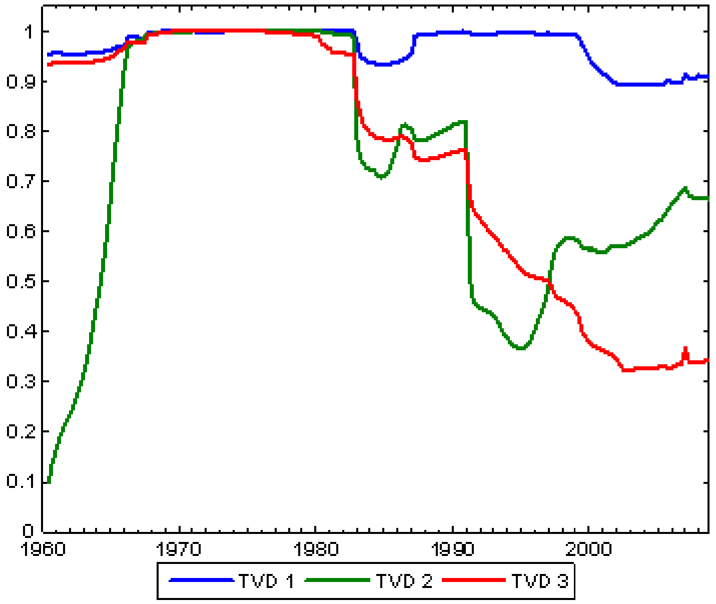

Figure 2: $p(K_{2,t} = 1 \,|\, y)$

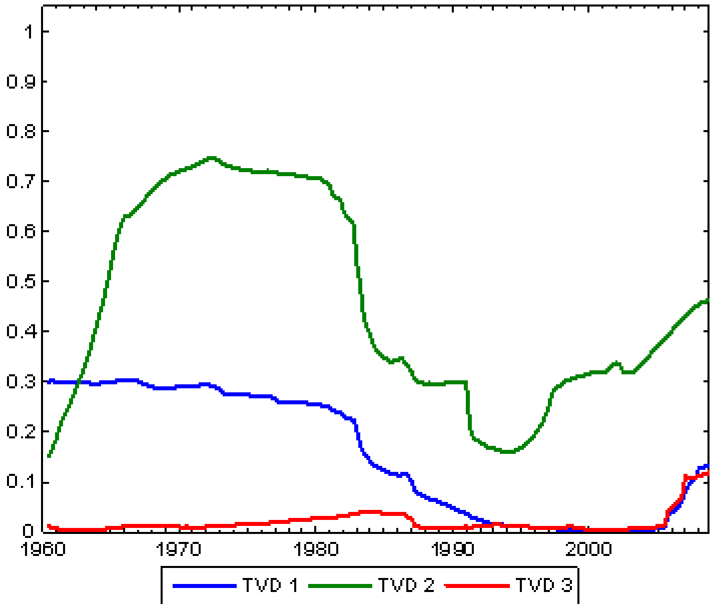

Figure 3: $p(K_{3,t} = 1 \,|\, y)$.

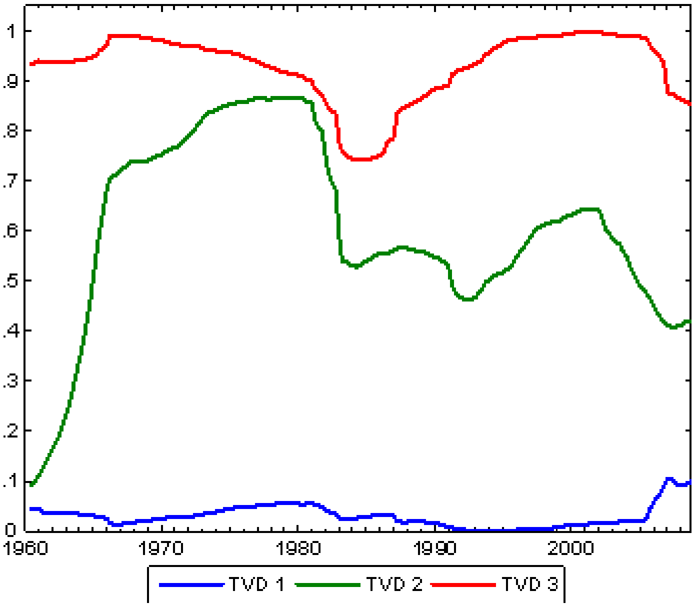

Figure 4: $p(K_{4,t} = 1 \,|\, y)$.

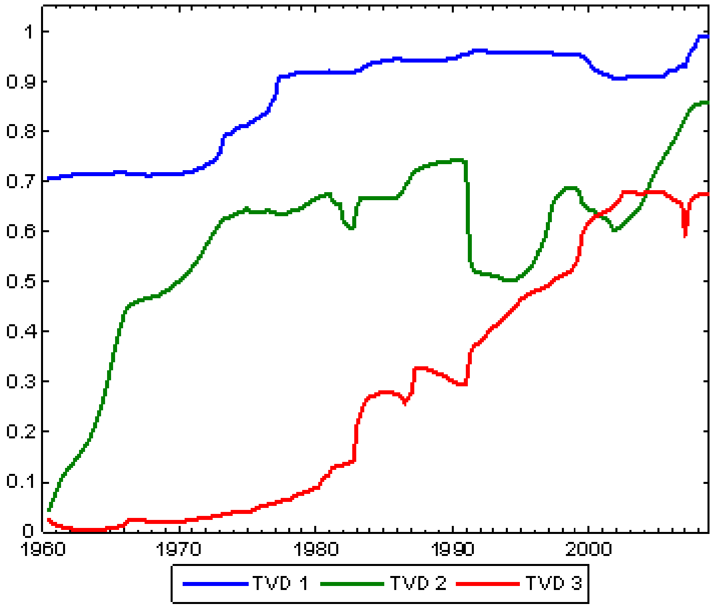

Figure 5: $p(K_{5,t} = 1 \,|\, y)$.

Figures 6 through 10 present posterior results for the coefficients themselves. That is, they plot $E\left(\gamma_{j,t}|y\right)$ for $j = 1, .., 5$. On the whole, they tell a similar story to Figures 1 through 5. For instance, Figure 10 shows the increasing role of the inflation expectations variable as time passes. There is also substantial evidence of time-variation in the marginal effects of the predictors in most cases. Previously we found that the first TVD approach exhibited less time variation in K_t and found it more difficult to switch between predictors. The impact of this can be clearly seen in several of the figures. For instance, in Figures 6, 8 and 9, the line corresponding to the first TVD approach is virtually flat (similar to what would be obtained using a recursive OLS estimate).

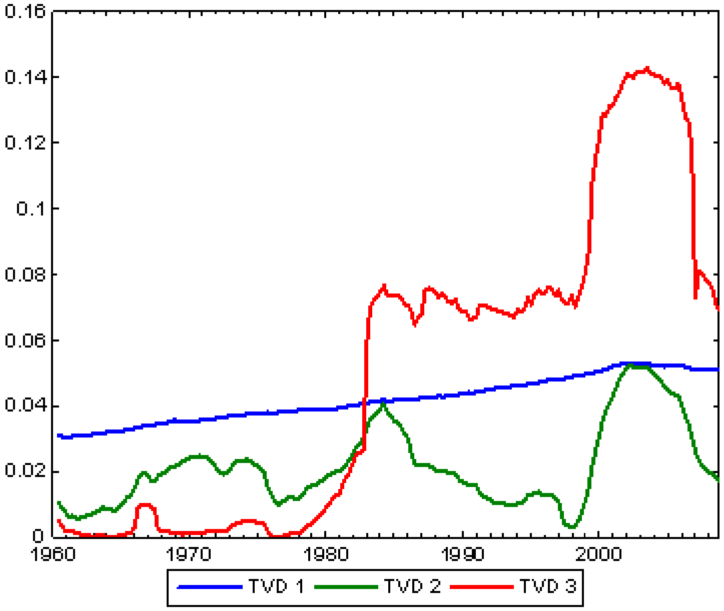

Figure 6: $E\left(\gamma_{1,t}|y\right)$

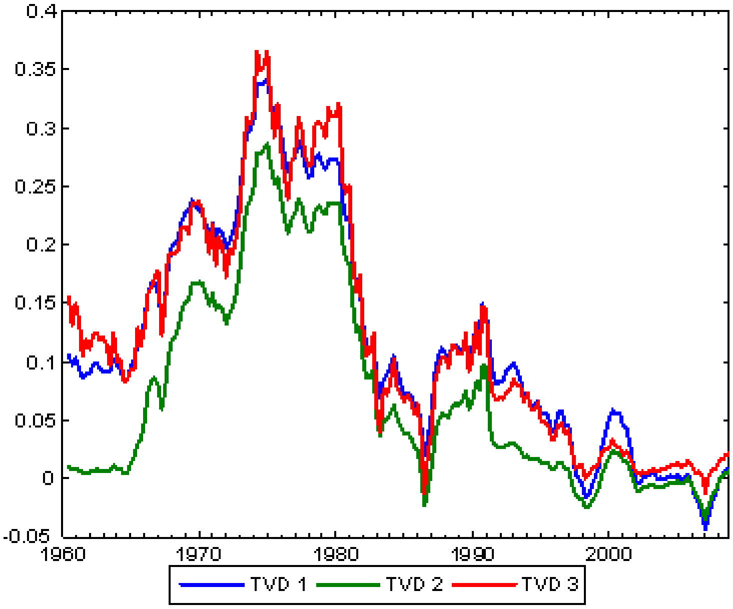

Figure 7: $E\left(\gamma_{2,t}|y\right)$

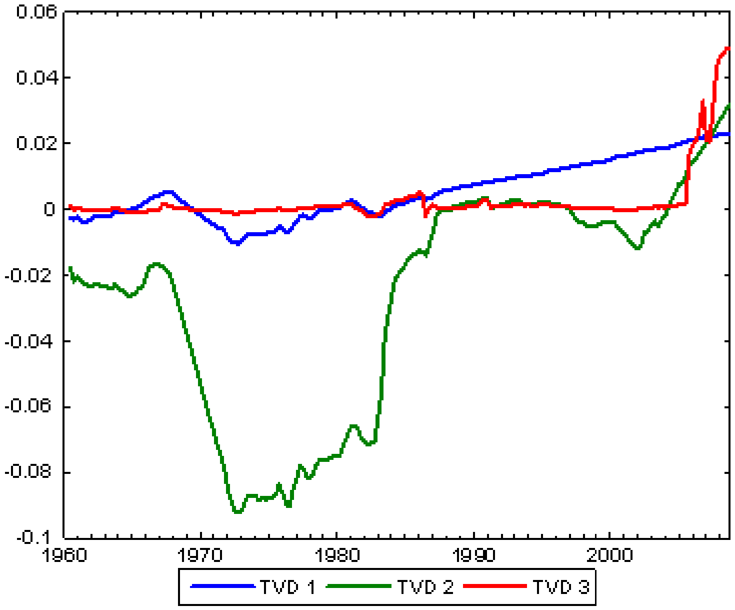

Figure 8: $E\left(\gamma_{3,t}|y\right)$

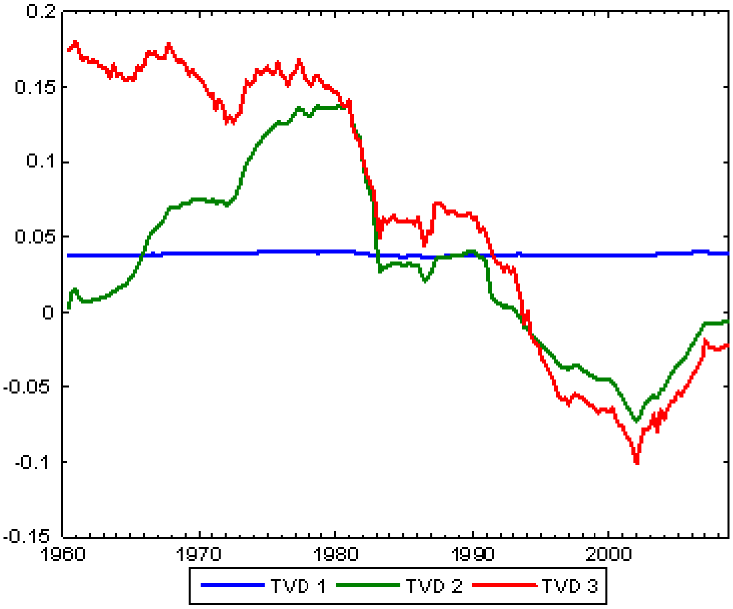

Figure 9: $E\left(\gamma_{9,t}|y\right)$

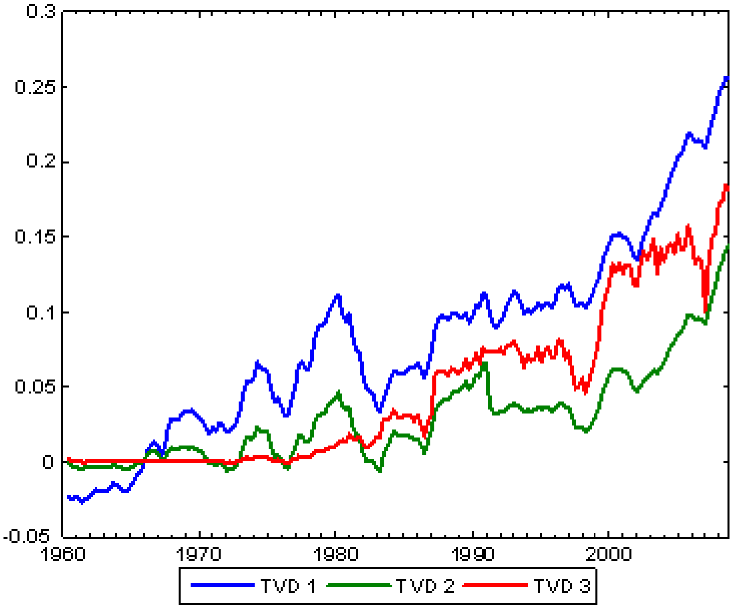

Figure 10: $E\left(\gamma_{5,t}|y\right)$

3.7 Results: Forecasting Exercise

Our models provide us with $p\left(y_{\tau+1}|Data_{\tau}\right)$, the predictive density for $y_{\tau+1}$ using data available through time τ. The predictive density is evaluated for $\tau=\tau_0,..,T-1$ where τ_0 is 1969Q4. Let $y^o_{\tau+1}$ be the observed value of $y_{\tau+1}$. Mean squared forecast error and mean absolute forecast error are common measures of forecast performance. These are defined as:

$$MSFE=\frac{\sum_{\tau=\tau_0}^{T-1}\left[y^o_{\tau+1}-E\left(y_{\tau+1}|Data_{\tau}\right)\right]^2}{T-\tau_0}$$

and

$$MAFE=\frac{\sum_{\tau=\tau_0}^{T-1}\left|y^o_{\tau+1}-E\left(y_{\tau+1}|Data_{\tau}\right)\right|}{T-\tau_0}$$

MSFE and MAFE only use the point forecasts and ignore the rest of the predictive distribution. For this reason, we also use the predictive likelihood to evaluate forecast performance. Note that a great advantage of predictive likelihoods is that they evaluate the forecasting performance of the entire predictive density. Predictive likelihoods are motivated and described in many places such as Geweke and Amisano (2010). The predictive likelihood is the predictive density for $y_{\tau+1}$ evaluated at the actual outcome $y^o_{\tau+1}$. We use the sum of log predictive likelihoods for forecast evaluation:

$$\sum_{\tau=\tau_0}^{T-1} \log\left[p\left(y_{\tau+1} = y^o_{\tau+1}|Data_\tau\right)\right].$$

Note that, if $\tau_0 = 0$ then this would be equivalent to the log of the marginal likelihood. Hence, the sum of log predictive likelihoods can also be interpreted as a measure similar to the log of the marginal likelihood, but made more robust by ignoring the initial $\tau_0 - 1$ observations in the sample (where prior sensitivity is most acute).[3]

Table 2 presents these forecast metrics for our three TVD approaches, the TVP model, a random walk model, the constant coefficient model (CC) and the constant coefficient model with stochastic volatility (CCSV). Regardless of which forecast metric we use, the evidence of Table 2 strongly indicates that all of our TVD approaches are forecasting substantially better than commonly-used benchmarks. Note that the researcher may wish to use the sum of log predictive likelihoods to construct posterior model probabilities for use in a Bayesian model averaging (BMA) exercise. If the two models under consideration were the second TVD model and the heteroskedastic constant coefficient model, then only 5% of the weight would be attached to the constant coefficient model. The homoskedastic constant coefficient model would receive only 1% in a similar exercise.

When we compare our three TVD approaches, we find that they perform similarly to one another. In terms of the predictive likelihoods (the metric preferred by most Bayesians), the second TVD approach forecasts best. However, in terms of MSFEs (MAFEs) the first (third) TVD approaches forecast best.

[3] To reduce the computational burden in this empirical illustration, the sums in our forecast metrics are taken at every seventh quarter (where seven is chosen so as to not to miss any seasonal differences).

Table 2: Measures of Forecast Performance			
Model	MSFE	MAFE	Sum of log Pred. like.
TVD 1	12.06	29.32	-8.96
TVD 2	13.13	30.02	-8.60
TVD 3	12.74	29.23	-8.62
Random Walk	27.37	41.08	–
CC	16.98	32.51	-13.10
CCSV	15.77	32.14	-11.50
TVP	13.74	31.01	-9.41

Table 2 establishes that, overall, the TVD approaches are forecasting better that some commonly used benchmarks. To gain insight on whether there are particular time periods they forecast particularly well, we present Figures 11 and 12. These are cumulative sums of the MSFE and log predictive likelihoods, respectively, for the various approaches. There is little evidence that there are particular periods of time where the TVD models obtain their overall lead in forecasting performance over the benchmark approaches. With the exception of the random walk model, which forecasts very poorly at the end of the sample, it seems that the TVD models are continually forecasting slightly better than the benchmark approaches and that, over time, this leads to substantial forecast improvements noted in Table 2.

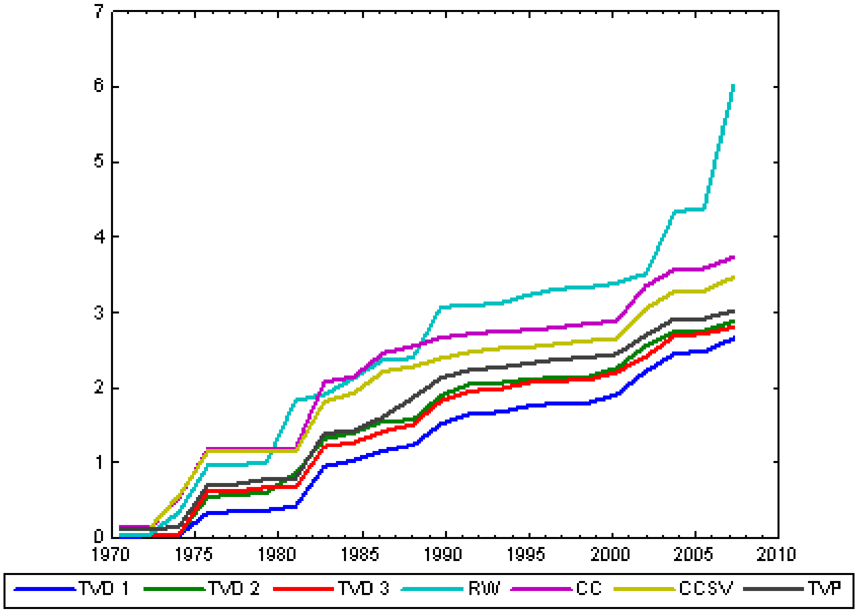

Figure 11: Cumulative Sums of MSFEs

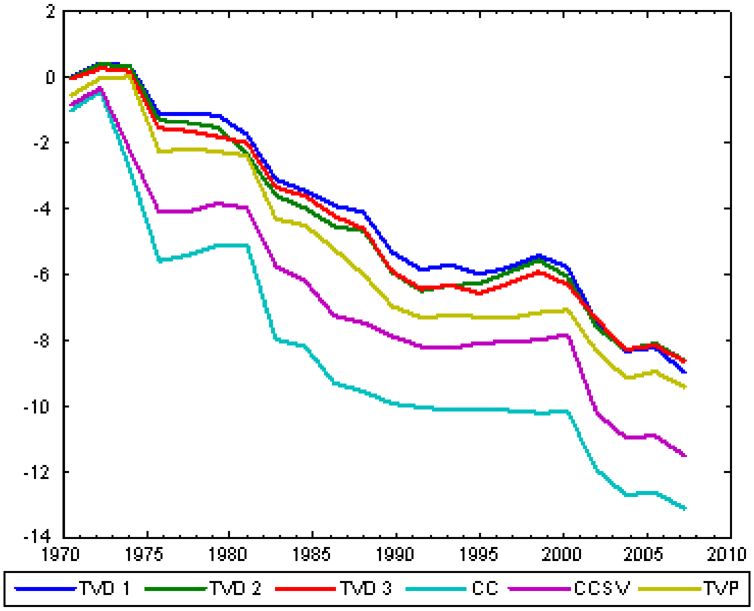

Figure 12: Cumulative Sums of Log Predictive Likelihoods

4 Conclusions

In this paper, we have presented a battery of theoretical and empirical arguments for the potential benefits of TVD models. Like TVP models, TVD models allow for the values of the parameters to change over time. Unlike TVP models, they also allow for the dimension of the parameter vector to change over time. Given the potential benefits of a TVD framework, the task is to build specific TVD models. This task was taken up in section 2 of this paper where three different TVD models were developed. These models all are dynamic mixture models and, thus, have the enormous benefit that we can draw on existing methods of posterior computation developed in Gerlach, Carter and Kohn (2000).

An empirical illustration involving forecasting US inflation illustrated the feasibility and desirability of the TVD approach. In-sample, we showed how TVD models can automatically and sensibly find parsimonious specifications within a flexible, but possibly over-parameterized, one. The benefits of such forms of shrinkage are shown in a forecasting exercise where TVD forecasting models out-perform benchmark alternatives.

References

Amato, J. and Swanson, N., 2001, The real-time predictive content of money for output, *Journal of Monetary Economics*, 48, 3-24.

Ang, A. Bekaert, G. and Wei, M., 2007, Do macro variables, asset markets, or surveys forecast inflation better?, *Journal of Monetary Economics* 54, 1163-1212.

Ballabriga, F., Sebastian, M. and Valles, J., 1999, European asymmetries, *Journal of International Economics*, 48, 233-253.

Canova, F., 1993, Modelling and forecasting exchange rates using a Bayesian time varying coefficient model, *Journal of Economic Dynamics and Control*, 17, 233-262.

Canova, F., 2007, *Methods for Applied Macroeconomic Research*, Princeton University Press: Princeton.

Canova, F. and Ciccarelli, M., 2004, Forecasting and turning point predictions in a Bayesian panel VAR model, *Journal of Econometrics*, 120, 327-359.

Carter, C. and Kohn, R., 1994, On Gibbs sampling for state space models, *Biometrika*, 81, 541–553.

Chan, J.C.C. and Jeliazkov, I., 2009, Efficient simulation and integrated likelihood estimation in state space models, *International Journal of Mathematical Modelling and Numerical Optimisation*, 1, 101-120.

Chib, S., 1996, Calculating posterior distributions and modal estimates in Markov mixture models, *Journal of Econometrics*, 75, 79-97.

Chib, S. and Greenberg, E., 1995, Hierarchical analysis of SUR models with extensions to correlated serial errors and time-varying parameter models, *Journal of Econometrics*, 68, 339-360.

Ciccarelli, M. and Rebucci, A., 2002, The transmission mechanism of European monetary policy: Is there heterogeneity? Is it changing over time?, International Monetary Fund working paper, WP 02/54.

Cogley, T. and Sargent, T., 2005, Drifts and volatilities: Monetary policies and outcomes in the post WWII U.S., *Review of Economic Dynamics*, 8, 262-302.

D'Agostino, A., Gambetti, L. and Giannone, D., 2009, Macroeconomic forecasting and structural change, ECARES working paper 2009-020.

Durbin, J. and Koopman, S., 2002, A simple and efficient simulation smoother for state space time series analysis, *Biometrika*, 89, 603-616.

Gerlach, R., Carter, C. and Kohn, R., 2000, Efficient Bayesian inference in dynamic mixture models, *Journal of the American Statistical Association*, 95, 819-828.

Geweke, J. and Amisano, G., 2010, Hierarchical Markov normal mixture models with applications to financial asset returns, *Journal of Applied Econometrics,* forthcoming.

Giordani, P. and Kohn, R., 2008, Efficient Bayesian inference for multiple change-point and mixture innovation models, *Journal of Business and Economic Statistics*, 12, 66-77.

Giordani, P., Kohn, R. and van Dijk, D., 2007, A unified approach to nonlinearity, structural change and outliers, *Journal of Econometrics*, 137, 112-133.

Groen, J., Paap, R. and Ravazzolo, F., 2009, Real-time Inflation Forecasting in a Changing World, Econometric Institute Report 2009-19, Erasmus University.

Kim, S., Shephard, N. and Chib, S., 1998, Stochastic volatility: likelihood inference and comparison with ARCH models, *Review of Economic Studies*, 65, 361-93.

Koop, G., 2003, *Bayesian Econometrics*, Wiley: Chichester.

Koop, G. and Korobilis, D., 2009, Forecasting inflation using dynamic model averaging, manuscript available at http://personal.strath.ac.uk/gary.koop/.

Koop, G., León-González, R. and Strachan R.W., 2009a, Dynamic probabilities of restrictions in state space models: An application to the Phillips curve, Journal of Business and Economic Statistics, forthcoming.

Koop, G., León-González, R. and Strachan R.W., 2009b, On the evolution of the monetary policy transmission mechanism, Journal of Economic Dynamics and Control, 33, 997–1017.

Koop, G. and Potter, S., 2009, Time varying VARs with inequality restrictions, manuscript available at http://personal.strath.ac.uk/gary.koop/koop_potter14.pdf.

Korobilis, D., 2009, VAR forecasting using Bayesian variable selection, manuscript available at http://mpra.ub.uni-muenchen.de/21124/.

Primiceri. G., 2005, Time varying structural vector autoregressions and monetary policy, *Review of Economic Studies*, 72, 821-852.

Raftery, A., Karny, M., Andrysek, J. and Ettler, P., 2007, Online prediction under model uncertainty via dynamic model averaging: Application to a cold rolling mill, Technical report 525, Department of Statistics, University of Washington.

Staiger, D., Stock, J. and Watson, M., 1997, The NAIRU, unemployment and monetary policy, *Journal of Economic Perspectives,* 11, 33-49.

Stock, J. and Watson, M., 2007. Why has US inflation become harder to forecast?, *Journal of Money, Credit and Banking*, 39, 3-33.

Stock, J. and Watson, M., 2008, Phillips curve inflation forecasts, NBER Working Paper No. 14322, 2008.