Large Time-Varying Parameter VARs: A Non-Parametric Approach*

George Kapetanios
Queen Mary, University of London
g.kapetanios@qmul.ac.uk

Massimiliano Marcellino
Bocconi University, IGIER and CEPR
massimiliano.marcellino@unibocconi.it

Fabrizio Venditti
Banca d'Italia
fabrizio.venditti@bcandaditalia.it

June 2016

Abstract

In this paper we introduce a nonparametric estimation method for a large Vector Autoregression (VAR) with time-varying parameters. The estimators and their asymptotic distributions are available in closed form. This makes the method computationally efficient and capable of handling information sets as large as those typically handled by factor models and Factor Augmented VARs (FAVAR). When applied to the problem of forecasting key macroeconomic variables, the method outperforms constant parameter benchmarks and large Bayesian VARs with time-varying parameters. The tool can also be used for structural analysis. As an example, we study the time-varying effects of oil price innovations on sectoral U.S. industrial output. We find that the changing interaction between unexpected oil price increases and business cycle fluctuations is shaped by the durable materials sector, rather by the automotive sector on which a large part of the literature has typically focused.

JEL classification: C14, C32, C53, C55

Keywords: Large VARs, time-varying Parameters, Non Parametric Estimation, Forecasting, Impulse Response Analysis

*The views expressed in this paper are those of the authors and do not necessarily reflect those of the Banca d'Italia. We thank for comments seminar participants at the Banca d'Italia, Deutsche Bundesbank, Bank of Canada, Queen Mary University of London, Norges Bank and European University Institute.

1 Introduction

In recent years macro-econometric research has been particularly active on two fronts. First, increasing availability of economic time series has prompted the development of methods capable of handling large dimensional datasets. Second a number of changes in the economic landscape (a renewed stream of oil price shocks, the Great Recession, unconventional monetary policy in most advanced countries) further stimulated work on models with time-varying parameters.

On the large models front typical solutions include data reduction and parameter shrinkage. Data reduction reduces the data space through linear combinations (factors) of the observed variables. This parsimonious representation of the data typically yields benefits in terms of estimation precision and forecasting. Shrinkage, on the other hand, constraints the parameter space within values that are (a priori) plausible. It therefore reduces estimation uncertainty, providing an alternative solution to the over-fitting problem. In the context of large Vector Autoregressions (VARs), for example, Banbura, Giannone, and Reichlin (2010) show that progressively tightening shrinkage as the cross-sectional dimension of the VAR increases, results in more accurate forecasts than those obtained on the basis of unrestricted VARs. Despite different premises, data reduction and shrinkage go in the same direction since, as shown by De Mol, Giannone, and Reichlin (2008), both methods stabilize OLS estimation by *regularising* the covariance matrix of the regressors.

Turning to time-varying parameters (TVP) models, a prolific line of research has grown in the Bayesian context, starting from the seminal work on VARs with time-varying coefficients and variances by Cogley and Sargent (2005) and Primiceri (2005). The estimation procedure of these models rests on the assumption that the VAR coefficients follow a random walk (or autoregressive) plus noise process. The assumed law of motion for the model parameters, coupled with the VAR equations, form a State Space system. Given the presence of time-varying second moments, a combination of Kalman filtering and Metropolis Hasting sampling is then used to deal with such models. The need to use the Kalman filter, however, limits the scale of the models, so that the numerous empirical applications that have followed this approach usually model a relatively small number of time series, see for example Benati and Surico (2008), Canova and Gambetti (2010) and Benati and Mumtaz (2007). Furthermore, in settings where the nature of the structural change is uncertain, methods based on simple data discounting could be more robust than Kalman filter based models.

These motivations are behind a stream of papers that in recent years have explored the performance of non parametric estimation methods for TVP-models. The viewpoint of this line of research is that the nature of time variation in the co-movement across time series is itself evolving, i.e. large infrequent breaks could coexist with periods of slow gradual time variation. Given this complexity, adaptive methods can deliver good forecasts and an accurate description of the structural relationships among macroeconomic variables at a relatively low computational

cost.[1] In this framework Giraitis, Kapetanios, and Yates (2014) and Giraitis, Kapetanios, and Yates (2012) have developed non-parametric estimators for univariate and multivariate dynamic models. They show that, for a wide class of models in which the coefficients evolve stochastically over time, the path of the parameters can be consistently estimated by suitably discounting distant data and provide details on how to choose the degree of such discounting. Furthermore, being available in closed form, the estimator proposed by Giraitis, Kapetanios, and Yates (2012) in the context of VARs partly addresses the curse of dimensionality, as systems of seven variables are easily handled, see Giraitis, Kapetanios, Theodoridis, and Yates (2014).

The paths traced by the large model literature and by the TVP model literature have seldom crossed. Connections have been established for Factor Augmented VAR (FAVAR) models, see for example Eickmeier, Lemke, and Marcellino (2015) and Mumtaz and Surico (2012), while they are still scant in the VAR literature. A notable exception is represented by the paper by Koop and Korobilis (2013) where the restricted Kalman filter by Raftery, Karny, and Ettler (2010) is used to make a TVP-VAR suitable for large information sets. This approach, while solving some issues, presents some shortcomings. First, the curse of dimensionality is only partially addressed since the parametric nature of the model implicitly limits its size. In practice, this framework cannot handle the large information sets employed in factor models (or FAVARs) or the large number of lags that are used when fitting medium-size VARs to monthly data like in Banbura, Giannone, and Reichlin (2010). Second, if the true data generating process is different from the postulated random walk type variation, the robustness of the Kalman filter to model misspecification is an obvious concern.

In this paper we propose an estimator that addresses, in a nonparametric context, both of these problems. Our idea is to start from the nonparametric estimator proposed by Giraitis, Kapetanios and Yates (2012), and adapt it to handle large information sets. To solve the issue of over fitting that arises when the size of the VAR increases, we recur to the mixed estimator by Theil and Goldberger (1960), which imposes stochastic constraints on the model coefficients, therefore mimicking in a classical context the role of the prior in Bayesian models. The resulting estimator, for which we derive asymptotic properties, mixes sample and non sample information to shrink the model parameters. It can be seen both as a generalization to a time-varying parameter structure of the model by Banbura, Giannone, and Reichlin (2010) and as a penalized regression version of the estimator by Giraitis, Kapetanios and Yates (2012). The proposed method is, given its nonparametric nature, robust to changes in the underlying data generating process and for popular shrinkage methods delivers equation by equation estimation. This implies that the estimator can cope with systems as large as those analyzed in the FAVAR and factor model literature.

Our estimator depends crucially on two parameters, the tuning constant that regulates the width of the kernel window used to discount past data, and the penalty parameter that deter-

[1] A crucial issue in this framework is how to select the degree of data discounting. The problem is addressed by Giraitis, Kapetanios, and Price (2013), who show how to make this choice data dependent using cross-validation methods.

mines the severity of the constraints imposed to control over fitting, akin to the prior tightness in Bayesian estimators. In the paper we explore a variety of cross-validation techniques to set these two parameters based on past model performance.[2] We also consider model averaging as an alternative strategy to deal with model uncertainty.

We next assess in Monte Carlo experiments the finite sample performance of our estimator, which turns out to be good, and compare it with the parametric estimator for TVP-VARs proposed by Koop and Korobilis (2013). We find that when the data generating process matches exactly the one assumed in the parametric setup, the two estimators give broadly similar results. Yet, as we move away from this assumption, the performance of the parametric estimator deteriorates, while our non-parametric estimator proves quite robust to changes in the underlying data generating process.

After discussing the theoretical and finite sample properties of the non-parametric estimators, we examine their use through a number of applications. First, we explore whether time variation is indeed a necessary feature of the model to successfully forecast key macroeconomic variables using a large panel (up to 78 variables) of U.S. monthly time series. We organize the forecast exercise around three questions that have been central to the forecasting literature in recent years. The first one is whether time variation actually improves forecast accuracy. The second one is whether the performance of *medium-sized* VARs with time-varying parameters can be approximated by that of *large* VARs with *constant coefficients*. This question is motivated by the contrasting findings in Stock and Watson (2012), who find little evidence of parameter changes during the financial crisis in the context of a factor model, and those reported by Aastveit, Carriero, Clark, and Marcellino (2014), who provide substantial evidence of parameter changes in smaller dimensional VARs. This conflicting evidence suggests that parameter time variation can be due, at least partly, to omitted variables, so that enlarging the information set makes parameters' time variation unnecessary. Once we have established that time variation is indeed beneficial to forecast accuracy the third question is whether it pays off to go beyond a medium size system, i.e. if going from a 20 to a 78 TVP-VAR improves forecast accuracy for the small set of key variables that we are interested in.

The analysis indicates that the introduction of time variation in the model parameters yields an improvement in prediction accuracy over models with constant coefficients, in particular when forecast combination is used to pool forecasts obtained with models with different degrees of time variation and shrinkage. Our findings also indicate that, especially at longer horizons, medium-sized TVP-VARs perform better than a VAR with constant parameters that uses a large information set. Finally, we find that, in the context of TVP-VARs, going beyond 20 variables is not beneficial to forecast accuracy, in line with the results for the constant parameter case in studies such as Banbura, Giannone, and Reichlin (2010) and Koop (2013).

[2]The use of data discounting in regression models as a way to handle structural breaks is the focus of a large literature, see in particular Pesaran and Timmermann (2007), Pesaran and Pick (2011), and Rossi, Inoue, and Jin (2014). All these papers, however, are concerned with single equation regressions rather than with large models.

Our non-parametric large TVP-VAR is also useful for structural analysis. As an illustration we revisit, in the context of a large information set, the issue of the diminished effects of oil price shocks on economic activity, a question that has spurred a large number of studies in the applied macro literature in recent years, see for example Hooker (1999), Edelstein and Kilian (2009), Hamilton (2009), Blanchard and Gali (2007), Blanchard and Riggi (2013) and Baumeister and Peersman (2013). The use of a large information set allows us to take a more granular view, allowing us to uncover some interesting findings on the evolving impact of oil price innovations on the output of different sectors of the U.S. industry. Specifically, we find that the declining role of oil prices in shaping U.S. business cycle fluctuations stems from lower effects on the production of durable materials, rather than on the automotive sector on which part of the literature has traditionally focused.

The paper is structured as follows. In Section 2 we describe the estimation method and derive its theoretical properties. In Section 3, we discuss cross-validation and model averaging. In Section 4 we assess the finite sample properties of our nonparametric method in Monte Carlo experiments and compare it with available parametric methods. In Section 5 we present the main forecasting exercise. In Section 6 we present an analysis of the time-varying impact of unexpected increases in the price of oil on U.S. industrial production. In Section 7 we summarize our main findings and conclude. Additional details are provided in Appendixes.

2 Setup of the problem

Let us consider a p-order VAR with n variables and time-varying (stochastic) coefficients:

$$\begin{aligned} \underset{1\times n}{y_t'} &= x_t'\Theta_t + \underset{1\times n}{u_t'}, \ t = 1, ..., T \\ \underset{1\times k}{x_t'} &= [y_{t-1}', y_{t-2,...,}' y_{t-p}', 1] \\ \underset{k\times n}{\Theta_t} &= [\Theta_{t,1}', \Theta_{t,2}', ..., \Theta_{t,p}', A_t']' \end{aligned} \tag{1}$$

where $k = (np + 1)$ is the number of random coefficients to be estimated in each equation so that at each time t there are nk parameters to be estimated, collected in the matrix Θ_t. For the time being, we assume that u_t is a martingale difference process with finite variance Σ_n.[3] A further crucial assumption is that Θ_t changes rather slowly, i.e., that:

$$\sup_{j\leq h} \|\Theta_t - \Theta_{t+j}\| = O_p\left(\frac{h}{t}\right). \tag{2}$$

A number of classes of models satisfy (2). For example, one such model is obtained by setting $\Theta_t = [\theta_{ij,t}]$, $\tilde{\Theta}_t = [\tilde{\theta}_{ij,t}]$ and letting $\tilde{\theta}_{ij,t} = \tilde{\theta}_{ij,t-1} + \epsilon_{\tilde{\theta},ij,t}$ and $\theta_{ij,t} = \theta_{ij}\frac{\tilde{\theta}_{ij,t}}{\max_{1\leq i\leq t}\tilde{\theta}_{ij,t}}$ for some bounded set of constants θ_{ij} and some set of stochastic processes $\epsilon_{\tilde{\theta},ij,t}$. This is an example

[3]The issue of heteroschedasticity is discussed in Section 2.5.

of a bounbed random walk model. We can allow for a wide variety of processes, $\epsilon_{\tilde{\theta},ij,t}$, making this class suitably wide.

Applying the *vec* operator to both sides of (1) we obtain:

$$\underset{n\times 1}{y_t} = \underset{n\times nk}{(I_n \otimes x_t')} \underset{nk\times 1}{\beta_t} + \underset{n\times 1}{u_t}, \tag{3}$$

where $\beta_t = vec(\Theta_t)$. Assuming persistence and boundedness[4] of the coefficients in Θ_t, Giraitis, Kapetanios, and Yates (2012, henceforth GKY) show that the path of the random coefficients is consistently estimated by the following kernel estimator:

$$\beta_t^{GKY} = \left[I_n \otimes \sum_{j=1}^{T} w_{j,t}(H) x_j x_j' \right]^{-1} \left[\sum_{j=1}^{T} w_{j,t}(H) vec\left(x_j y_j'\right) \right], \tag{4}$$

where the generic j^{th} element $w_{j,t}(H)$ is a kernel, function with bandwidth H, used to discount distant data. Throughout the paper we use a Gaussian kernel:[5]

$$w_{j,t}(H) = \frac{K_{j,t}(H)}{\sum_{j=1}^{T} K_{j,t}(H)}, \tag{6}$$

$$K_{j,t}(H) = (1/\sqrt{2\pi}) exp\left[-\frac{1}{2} \left(\frac{j-t}{H} \right)^2 \right]. \tag{7}$$

One appealing feature of the estimator in (4) is that, given the Kronecker structure of the first term, it only requires the inversion of the $k \times k$ matrices $\sum_{j=1}^{T} w_{j,t} x_j x_j'$. In other words, estimation can be performed equation by equation that, as emphasized by Carriero, Clark, and Marcellino (2016) in a Bayesian context, substantially reduces the computing time.

A more compact notation is obtained by introducing the following notation: $X_{w,t} = W_{H,t} X$, where $W_{H,t} = diag(w_{1t}^{1/2}(H), ..., w_{Tt}^{1/2}(H))$ and the $T \times k$ matrix X is formed by stacking over t the vectors x_t'. Also, let us define $X_{ww,t} = W_{H,t} X_{w,t}$ and denote with Y the $T \times n$ matrix formed by stacking over t the vectors y_t'. The GKY estimator can now be cast in the following matrix form:

$$\Theta_t^{GKY} = \left[X_{w,t}' X_{w,t} \right]^{-1} \left[X_{ww,t}' Y \right]. \tag{8}$$

[4]More specifically, writing the VAR in companion form as a VAR(1) model, $Y_t = \Psi_t Y_{t-1}$, GKY assume that the spectral norm (that is the maximum absolute eigenvalue) of Ψ_t is strictly lower than 1.

[5]When forecasting, in order to preserve the pseudo real time nature of the exercise, we introduce an indicator function that assigns zero weight to the out of sample observations, so that only in sample information is used to estimate the parameters:

$$K_{j,t}(H) = (1/\sqrt{2\pi}) exp\left[-\frac{1}{2} \left(\frac{j-t}{H} \right)^2 \right] I(j \leq t) \tag{5}$$

2.1 Shrinkage through stochastic constraints

When the dimension of the system grows, it is desirable to impose some shrinkage on the model parameters to avoid an increase in the estimation variance (Hastie, Tibshirani, and Friedman, 2003). While in a Bayesian framework this can be achieved through the prior distribution, in a classical framework shrinkage can be performed by using the mixed estimator of Theil and Goldberger (1960). This is obtained by adding a set of stochastic constraints (i.e., constraints that hold with some degree of uncertainty) to model (3). The constraints are written as linear combinations of the parameter vector β_t plus a vector of noises, where the latter ensures that the constraints do not hold exactly. The complete model can be written as:

$$\underset{n\times 1}{y_t} = \underset{n\times nk}{(I_n \otimes x_t')}\ \underset{nk\times 1}{\beta_t} + \underset{n\times 1}{u_t} \tag{9}$$

$$\sqrt{\lambda}\ \underset{nk\times 1}{r} = \sqrt{\lambda}\ \underset{nk\times nk}{R}\ \underset{nk\times 1}{\beta_t} + \underset{nk\times 1}{u_t^r}\ . \tag{10}$$

We assume that the errors u_t^r are a martingale difference process with finite variance and that their variance is proportional to that of the data, that is $var(u_t^r) = I_k \otimes \Sigma_n$. In other words, when the noise in the dynamic relationship between y_t and x_t is high, uncertainty about the constraints on the coefficients β_t also increases. As for expected value of u_t^r, for the moment we leave it unspecified since it plays a crucial role in determining the bias of the estimator, as we show further below. Notice that both sides of equation (10) are pre-multiplied by a constant $\sqrt{\lambda}$. It is easy to see that this constant acts a scaling factor of the variance of the stochastic constraints u_t^r.[6] Hence, low values of λ imply that the coefficient vector β_t is left relatively unrestricted; vice versa, high values of λ imply that the constraints in (10) hold relatively more tightly. Regarding the structure of the matrix R, we consider two cases. In the first case we assume that R has a Kronecker structure: $\underset{nk\times nk}{R} = (I_n \otimes \underset{k\times k}{\overline{R}})$. This case is of particular interest for two reasons. First, it holds for a number of popular shrinkage methods, like the Ridge regression and the Litterman prior. Second, it results in an estimator that can be cast in matrix form, hence being very efficient from a computational point of view and directly comparable to its unconstrained counterpart, i.e. the GKY estimator. Next, we consider the more general case where R does not have this particular structure.

2.2 Case 1: R has a Kronecker structure

If R has a Kronecker structure, the analysis of the estimator can proceed equation by equation. First let us state the following definitions: $\underset{kn\times 1}{u_t^r} = vec(\underset{k\times n}{\overline{u}_t^r})$ and $\underset{kn\times 1}{r} = vec(\underset{k\times n}{\overline{r}})$. Also, since $\underset{nk\times nk}{R} = (I_n \otimes \underset{k\times k}{\overline{R}})$ and $\beta_t = vec(\Theta_t)$, it follows that $R\beta_t = (I_n \otimes \overline{R})vec(\Theta_t) = vec(\overline{R}\Theta_t)$.

[6] Notice that, by multiplying both sides of (10) by $\frac{1}{\sqrt{\lambda}}$ the variance of the noise in (10) becomes $\frac{1}{\lambda}(I_k \otimes \Sigma_n)$.

Hence the joint model in (9) and (10) can be expressed in matrix form as:

$$\underset{1\times n}{y_t'} = \underset{1\times k}{x_t'}\underset{k\times n}{\Theta_t} + \underset{1\times n}{u_t'}, \tag{11}$$

$$\underset{k\times n}{\sqrt{\lambda}\,\overline{r}} = \underset{k\times k}{\sqrt{\lambda}\,\overline{R}}\underset{k\times n}{\Theta_t} + \underset{k\times n}{\overline{u}_t^r}, \tag{12}$$

or more compactly:

$$\underset{(k+1)\times n}{y_t^*} = x_j'^*\Theta_t + u_t^*, \tag{13}$$

where $y_t^* = [y_t, \sqrt{\lambda}\overline{r}']'$, $x_j'^* = [x_t, \sqrt{\lambda}\overline{R}']'$, $u_t^* = [u_t, \overline{u}_t'^r]'$, $u_t^{**} = [u_t, 0]'$ and $var(vec(u_t^*)) = I_{k+1} \otimes \Sigma_n$. The extended regression model in (13) can be analyzed using the GKY estimator, with related properties. The estimator has the form:

$$\underset{k\times n}{\widehat{\Theta}_t} = \left(\sum_{j=1}^{T} w_{j,t}x_j^*x_j^{*\prime}\right)^{-1}\left(\sum_{j=1}^{T} w_{j,t}x_j^*y_j^*\right), \tag{14}$$

where, for simplicity, we have omitted the dependence of $w_{j,t}$ from the bandwidth H. Separating the contribution of the actual data from that of the constraints, the estimator can equivalently be written as:[7]

$$\underset{k\times n}{\widehat{\Theta}_t} = \left(\sum_{j=1}^{T} w_{j,t}x_jx_j' + \lambda\overline{R}'\overline{R}\right)^{-1}\left(\sum_{j=1}^{T} w_{j,t}x_jy_j' + \lambda\overline{R}'\overline{r}\right) \tag{15}$$

$$= \left(X_{w,t}'X_{w,t} + \lambda\overline{R}'\overline{R}\right)^{-1}\left(X_{ww,t}'Y + \lambda\overline{R}'\overline{r}\right). \tag{16}$$

It is worth making the following observations. First, when $\lambda = 0$ the constrained estimator equals the unconstrained one: $\widehat{\Theta}_{t,GKY} = \left(X_{w,t}'X_{w,t}\right)^{-1}\left(X_{ww,t}'Y\right)$. Second, and vice versa, as $\lambda \to \infty$, $\widehat{\Theta}_t$ converges to the value implied by the constraints, that is $\Theta_t \to \Theta_C = (\overline{R}'\overline{R})^{-1}(\overline{R}'\overline{r})$. Hence, the constant term $\sqrt{\lambda}$ can also be interpreted as the weight of the sample size of the artificial observations (r and R) relative to T, the sample size of the observed data y_t and x_t. It is worth remarking that the value implied by the constraints (Θ_C) is time invariant. This means that the stochastic constraints *anchor* the evolution of Θ_t around a fixed value that is specified ex ante. Third, $\widehat{\Theta}_t$ can be expressed as the weighted sum of its unrestricted and restricted versions, $\widehat{\Theta}_{t,GKY}$ and Θ_C. To see this point, re-write (16) as follows:

$$\underset{k\times n}{\widehat{\Theta}_t} = \left(X_{w,t}'X_{w,t} + \lambda\overline{R}'\overline{R}\right)^{-1}\left[(X_{w,t}'X_{w,t})\widehat{\Theta}_{t,GKY} + (\lambda\overline{R}'\overline{R})\Theta_C\right] \tag{17}$$

$$= S_w^{-1}(X_{w,t}'X_{w,t})\widehat{\Theta}_{t,GKY} + S_w^{-1}(\lambda\overline{R}'\overline{R})\Theta_C, \tag{18}$$

[7]Notice that $\sum_{j=1}^{T} w_{j,t}x_j^*x_j^{*\prime} = \sum_{j=1}^{T} w_{j,t}\begin{bmatrix} x_t & \sqrt{\lambda}\overline{R}' \end{bmatrix}\begin{bmatrix} x_t' \\ \sqrt{\lambda}\overline{R} \end{bmatrix}$ and $\sum_{j=1}^{T} w_{j,t}x_j^*y_j^* = \sum_{j=1}^{T} w_{j,t}\begin{bmatrix} x_t & \sqrt{\lambda}\overline{R}' \end{bmatrix}\begin{bmatrix} y_t' \\ \sqrt{\lambda}\overline{r} \end{bmatrix}$.

where $S_w = \left(X'_{w,t}X_{w,t} + \lambda\overline{R}'\overline{R}\right)$.

The properties of $\widehat{\Theta}_t$ are derived in the following theorem.

Theorem 1 *Let the model be given by (13) where u_t is a martingale difference sequence with finite fourth moments. Let (2) hold and $H = o(T^{1/2})$. Let $X^*_{w,t} = W_{H,t}X^*$ where X^* is obtained by stacking over t the vectors $x_t'^*$, $X^*_{ww,t} = W_{H,t}X^*_{w,t}$, $\Gamma^*_{w,t} = p\lim \frac{1}{H}X^{*\prime}_{w,t}X^*_{w,t}$, $\Gamma^*_{ww,t} = p\lim \frac{1}{H}X^{*\prime}_{ww,t}X^*_{ww,t}$. Then,*

$$\left(\Gamma^{*-1}_{w,t}\Gamma^{**}_{ww,t}\Gamma^{*-1}_{w,t} \otimes \Sigma_n\right)^{-\frac{1}{2}} \sqrt{H} vec\left(\widehat{\Theta}'_t - \Theta'_t - \Theta'^B_t\right) \rightarrow^d N\left(0, I\right), \tag{19}$$

where $\Theta^B_t = p\lim S_w^{-1}\sqrt{\lambda}\overline{R}'\overline{u}^r_t = p\lim S_w^{-1}\lambda\overline{R}'\left(\overline{r} - \overline{R}\Theta_t\right)$ and $\Gamma^{**}_{ww,t}$ is defined in (27).

Proof. Starting from (18), we have:

$$\begin{aligned}
\widehat{\Theta}_t &= (S_w^{-1}X'_{w,t}X_{w,t})\widehat{\Theta}_{t,GKY} + S_w^{-1}(\lambda\overline{R}'\overline{R})\Theta_C & (20)\\
&= (S_w^{-1}X'_{w,t}X_{w,t})\widehat{\Theta}_{t,GKY} + S_w^{-1}(\lambda\overline{R}'\overline{R})\Theta_t + S_w^{-1}\sqrt{\lambda}\overline{R}'\overline{u}^r_t & (21)
\end{aligned}$$

where we have used the fact that

$$\begin{aligned}
\Theta_C &= (\overline{R}'\overline{R})^{-1}(\overline{R}'\overline{r}) & (22)\\
&= (\overline{R}'\overline{R})^{-1}\left(\overline{R}'\left(\overline{R}\Theta_t + \frac{1}{\sqrt{\lambda}}\overline{u}^r_t\right)\right) & (23)\\
&= \Theta_t + (\overline{R}'\overline{R})^{-1}\overline{R}'\frac{1}{\sqrt{\lambda}}\overline{u}^r_t & (24)
\end{aligned}$$

Taking probability limits, recalling that $p\lim \widehat{\Theta}_{t,GKY} = \Theta_t$, and that $S_w = (X'_{w,t}X_{w,t} + \lambda\overline{R}'\overline{R})$ we have that:

$$p\lim (S_w^{-1}X'_{w,t}X_{w,t})\widehat{\Theta}_{t,GKY} + S_w^{-1}(\lambda\overline{R}'\overline{R})\Theta_t + S_w^{-1}\overline{R}'\lambda\overline{u}^r_t \;=\; \Theta_t + \Theta^B_t$$

To determine the normalising factor in (19), let us go back to the representation in (14) and let us take differences from the true parameter matrix Θ_t and from the bias term Θ^B_t. We obtain:

$$\begin{aligned}
\widehat{\Theta}_t - \Theta_t - \Theta^B_t &= \left(\sum_{j=1}^T w_{j,t}x^*_j x^{*\prime}_j\right)^{-1}\left(\sum_{j=1}^T w_{j,t}x^*_j y^*_j\right) - \Theta_t - \Theta^B_t\\
&= \left(\sum_{j=1}^T w_{j,t}x^*_j x^{*\prime}_j\right)^{-1}\left(\sum_{j=1}^T w_{j,t}x^*_j (x^{*'}_j\Theta_j + u^*_j)\right) - \Theta_t - \Theta^B_t\\
&= \left(\sum_{j=1}^T w_{j,t}x^*_j x^{*\prime}_j\right)^{-1}\left(\sum_{j=1}^T w_{j,t}x^*_j x^{*\prime}_j\Theta_j\right) + \left(\sum_{j=1}^T w_{j,t}x^*_j x^{*\prime}_j\right)^{-1}\left(\sum_{j=1}^T w_{j,t}x^*_j u^*_j\right) - \Theta_t - \Theta^B_t
\end{aligned}$$

$$= \left(\sum_{j=1}^{T} w_{j,t} x_j^* x_j^{*\prime}\right)^{-1} \sum_{j=1}^{T} w_{j,t} x_j^* x_j^{*\prime} \left(\Theta_j - \Theta_t\right) +$$
$$+ \left(\sum_{j=1}^{T} w_{j,t} x_j^* x_j^{*\prime}\right)^{-1} \left(\sum_{j=1}^{T} w_{j,t} x_j^* u_j^* - \sum_{j=1}^{T} w_{j,t} x_j^* x_j^{*\prime} \Theta_t^B\right).$$

Now, if the bandwidth is $o(T^{1/2})$, then the term $\left(\sum_{j=1}^{T} w_{j,t} x_j^* x_j^{*\prime}\right)^{-1} \sum_{j=1}^{T} w_{j,t} x_j^* x_j^{*\prime} (\Theta_j - \Theta_t)$ is asymptotically negligible and we can focus on the second element.

First let us simplify the notation and let us write:

- $\underbrace{X_{w,t}^{*\prime} X_{w,t}^*}_{k\times k} \equiv \sum_{j=1}^{T} w_{jt} \underbrace{\underbrace{x_j^*}_{k\times(k+1)} \underbrace{x_j^{*\prime}}_{(k+1)\times k}}_{k\times k}$
- $\underbrace{\underbrace{X_{ww,t}^{*\prime}}_{k\times T(k+1)} \underbrace{U^*}_{T(k+1)\times n}}_{k\times n} \equiv \sum_{j=1}^{T} w_{j,t} \underbrace{\underbrace{x_j^*}_{k\times(k+1)} \underbrace{u_j^*}_{(k+1)\times n}}_{k\times n}$, where the $T(k+1)\times n$ matrix U^* is obtained by stacking over t the matrices $\underbrace{u_t^*}_{(k+1)\times n}$
- $\underbrace{\Lambda^{'}}_{k\times k} \equiv \underbrace{S_w^{-1}}_{k\times k} \sqrt{\lambda} \underbrace{\overline{R}^{\prime}}_{k\times k}$

Multiplying by $\sqrt{H}$ and transposing we obtain:

$$\sqrt{H}\left(\widehat{\Theta}_t^{\prime} - \Theta_t^{\prime} - \Theta_t^{\prime B}\right) = \left(\frac{1}{\sqrt{H}}\left(U^{'*} X_{ww,t}^* - \overline{u}_t^{\prime r}\Lambda\right)\right)\left(\frac{1}{H} X_{w,t}^{*\prime} X_{w,t}^*\right)^{-1}$$

Taking *vec* of both sides yields:

$$\sqrt{H} vec\left(\widehat{\Theta}_t^{\prime} - \Theta_t^{\prime} - \Theta_t^{\prime B}\right) = \left(H\left(X_{w,t}^{*\prime} X_{w,t}^*\right)^{-1} \otimes I_n\right) vec\left(\frac{1}{\sqrt{H}}\left(U^{'*} X_{ww,t}^* - \overline{u}_t^{\prime r}\Lambda\right)\right). \quad (25)$$

Let us analyze more in detail the term $vec\left(U^{'*} X_{ww,t}^* - \overline{u}_t^{\prime r}\Lambda\right)$. First, notice that:

$$vec\left(\frac{1}{\sqrt{H}} U^{'*} X_{ww,t}^* - \overline{u}_t^{\prime r}\Lambda\right) = \frac{1}{\sqrt{H}} \underbrace{\left(X_{ww,t}^{\prime *} \otimes I_n\right)}_{kn\times nT(k+1)} \underbrace{vec\left(U^{'*}\right)}_{nT(k+1)\times 1} - \frac{1}{\sqrt{H}} \underbrace{\left(\Lambda^{'} \otimes I_n\right)}_{kn\times kn} \underbrace{u_t^r}_{kn\times 1} \quad (26)$$

where we have used the fact that $u_t^r = vec(\overline{u}_t^{\prime r})$

We need to derive the asymptotic behaviour of this term. There are four terms to consider.

Term # 1:

$$\frac{1}{H}\left(X_{ww,t}^{\prime *} \otimes I_n\right) var\left(vec\left(U^{'*}\right)\right)\left(X_{ww,t}^* \otimes I_n\right) = \frac{1}{H}\left(X_{ww,t}^{\prime *} \otimes I_n\right)\left(I_{T(k+1)} \otimes \Sigma_n\right)\left(X_{ww,t}^* \otimes I_n\right)$$

$$
\begin{aligned}
&= \frac{1}{H}\left(X^{\prime *}_{ww,t} X^{*}_{ww,t} \otimes \Sigma_n\right) \\
&= \Gamma^{*}_{ww,t} \otimes \Sigma_n
\end{aligned}
$$

Term # 2:

$$
\left(\Lambda' \otimes I_n\right) Var(u_t^r) \left(\Lambda \otimes I_n\right) = \left(\Lambda' \otimes I_n\right) \left(I_k \otimes \Sigma_n\right) \left(\Lambda \otimes I_n\right) = \left(\Lambda'\Lambda \otimes \Sigma_n\right)
$$

Term # 3:

$$
\left(X^{\prime *}_{ww,t} \otimes I_n\right) cov \left(\underbrace{vec(U^{'*})}_{nT(k+1)\times 1}, \underbrace{(u_t^r)'}_{1\times kn} \right) \left(\Lambda \otimes I_n\right)
$$

To analyse this notice that :

$$
vec(U^{'*}) = \begin{bmatrix} u_1 \\ u_1^r \\ u_2 \\ u_2^r \\ ... \\ u_t \\ u_t^r \\ ... \end{bmatrix}
$$

This means that the relevant matrix will contain zeros everywhere but in correspondance of the vector u_t^r appearing in $vec(U^{'*})$, where it will equal $I_k \otimes \Sigma_n$. Compactly, this can be written as:

$$
\underbrace{cov\left(vec(U^{'*}), (vec(\overline{u}_t^r))'\right)}_{nT(k+1)\times kn} = \begin{bmatrix} 0_{[(t-1)(k+1)+1]\times k} \\ I_k \\ 0_{(T-t)(k+1)\times k} \end{bmatrix} \otimes \Sigma_n \equiv \underbrace{\Xi}_{T(k+1)\times k} \otimes \Sigma_n.
$$

Plugging in this term, we have

$$
\begin{aligned}
\left(X^{\prime *}_{ww,t} \otimes I_n\right) cov \left(\underbrace{vec(U^{'*})}_{nT(k+1)\times 1}, \underbrace{(u_t^r)'}_{1\times kn} \right) \left(\Lambda \otimes I_n\right) &= \left(X^{\prime *}_{ww,t} \otimes I_n\right) \left(\Xi \otimes \Sigma_n\right) \left(\Lambda \otimes I_n\right) \\
&= \left(\underbrace{X^{\prime *}_{ww,t} \Xi \Lambda}_{k\times k} \otimes \Sigma_n \right)
\end{aligned}
$$

Term # 4: is simply the transpose of Term # 3.
Collecting terms we have that the main normalizing term is:

$$\begin{aligned}\Gamma^{**}_{ww,t} &= \left(\Gamma^{*}_{ww,t}\otimes\Sigma_n\right)+\left(\Lambda'\Lambda\otimes\Sigma_n\right)-\left(X'^{*}_{ww,t}\Xi\Lambda\otimes\Sigma_n\right)-\left(\Lambda'\Xi X^{*}_{ww,t}\otimes\Sigma_n\right) \\ &= \left(\left(\Gamma^{*}_{ww,t}+\Lambda'\Lambda-X'^{*}_{ww,t}\Xi\Lambda-\Lambda'\Xi X^{*}_{ww,t}\right)\otimes\Sigma_n\right)\end{aligned} \tag{27}$$

From this the proof follows. ■

2.3 Case 2: R does not have a Kronecker structure

Let us now turn to the more general case when R does not have a Kronecker structure. In this case the estimator can be written as:

$$\begin{aligned}\widehat{\beta}_t &= \left[\left(\underbrace{I_n\otimes\sum_{j=1}^{T}w_{j,t}x_jx_j'}_{nk\times nk}\right)+\lambda R'R\right]^{-1}\left[\sum_{j=1}^{T}w_{j,t}\left(I_n\otimes x_j\right)y_j+\lambda R'r\right] \\ &= \left[\left(I_n\otimes\sum_{j=1}^{T}w_{j,t}x_jx_j'\right)+\lambda R'R\right]^{-1}\left[\sum_{j=1}^{T}w_{j,t}vec(x_jy_j')+\lambda R'r\right].\end{aligned} \tag{28}$$

A crucial difference between the estimator in (28) and the one in (16) is that the latter only requires the inversion of k dimensional matrices, which makes it computationally much faster. On the other hand, (28) can handle more general constraints. The properties of $\widehat{\beta}_t$ are derived in the following theorem.

Theorem 2 *Let the model be given by (9) and (10) where u_t is a martingale difference sequence with finite fourth moments. Let (2) hold, $H=o(T^{1/2})$. Let us define $\Phi=\frac{\lambda}{H}R'R$, $\Gamma_{w,t}=plim_{T\to\infty}\frac{1}{H}\sum_{j=1}^{T}w_{j,t}(x_jx_j')$, $\Gamma_{ww,t}=plim_{T\to\infty}\frac{1}{H}\sum_{j=1}^{T}w_{j,t}^2x_jx_j'$ and $\beta_t^B=plim_{T\to\infty}\left[\left(I_n\otimes\sum_{j=1}^{T}w_{j,t}x_jx_j'\right)+\lambda R'R\right]^{-1}\lambda R'r$. Then,*

$$\sqrt{H}\left[\left(I_n\otimes\Gamma_{w,t}+\Phi\right)^{-1}\left(\Sigma_n\otimes\Gamma_{ww,t}+\Phi\right)\left(I_n\otimes\Gamma_{wt}+\Phi\right)^{-1}\right]^{-1/2}\left(\widehat{\beta}_t-\beta_t-\beta_t^B\right)\to^d N\left(0,I\right) \tag{29}$$

Proof. Replacing in (28) y_j and r with the processes implied by the model (9)-(10) we have:

$$\begin{aligned}\widehat{\beta}_t-\beta_t &= \left[\sum_{j=1}^{T}w_{j,t}(I_n\otimes x_jx_j')+\lambda R'R\right]^{-1}\left[\sum_{j=1}^{T}w_{j,t}(I_n\otimes x_jx_j'+\lambda R'R)(\beta_j-\beta_t-\beta_t^B)\right]+ \\ &\left[I_n\otimes\sum_{j=1}^{T}w_{j,t}x_jx_j'+\lambda R'R\right]^{-1}\left[\sum_{j=1}^{T}w_{j,t}(I_n\otimes x_j)u_j+\sqrt{\lambda}R'u_t^r)\right]\end{aligned}$$

where, again, the term that multiplies $(\beta_j - \beta_t)$ is negligible assuming that the bandwidth is $o_p(T^{1/2})$. The analysis of the estimation bias and the convergence to normality follows trivially as in Theorem 1, so we do not repeat it here. ■

2.4 Variance and MSE of the constrained estimator

In this subsection we discuss the effects of penalization on the variance and on the Mean Square Error (MSE) of the penalized estimator, relative to those of the unconstrained one. In particular, in the following theorem we show that the stochastic constraints, whether valid or not, have the unambiguous effect of lowering a quantity related to the sample second moment of the non-parametric estimator.

Theorem 3 *Let the model be given by (9) and (10). Let* $E = I_n \otimes \sum_{j=1}^{T} w_{j,t} x_j x_j'$, $F = \lambda R' R$,*and* $G = \sum_{j=1}^{T} w_{j,t} vec(x_j y_j') = \sum_{j=1} w_{j,t} (I_n \otimes x_j) y_j'$. *Then,*

$$\widehat{\beta}_t = [E + F]^{-1} \left[E \widehat{\beta}_{t,GKY} + \lambda R' r \right].$$

Further, $[E + F]^{-1} \left[E \widehat{\beta}_{t,GKY} \widehat{\beta}_{t,GKY}' E' \right] [E' + F']^{-1} - \widehat{\beta}_{t,GKY} \widehat{\beta}_{t,GKY}'$ *is a positive semi-definite matrix.*

Proof. Let us rewrite the constrained estimator as a linear combination of the unconstrained one and of the constraints. If we define $E = I_n \otimes \sum_{j=1}^{T} w_{j,t} x_j x_j'$, $F = \lambda R' R$,and $G = \sum_{j=1}^{T} w_{j,t} vec(x_j y_j') = \sum_{j=1} w_{j,t} (I_n \otimes x_j) y_j'$,then the unconstrained estimator is $\widehat{\beta}_{t,GKY} = E^{-1} G$. We can therefore write (28) as:

$$\widehat{\beta}_t = [E + F]^{-1} \left[E \widehat{\beta}_{t,GKY} + \lambda R' r \right]$$

Defining $C = \widehat{\beta}_{t,GKY} \widehat{\beta}_{t,GKY}'$ we have that for any vector q and $w = E [E + F]^{-1} q$

$$q' \left(\begin{array}{c} [E + F]^{-1} \left[E \widehat{\beta}_{t,GKY} \widehat{\beta}_{t,GKY}' E' \right] [E' + F']^{-1} - \\ \widehat{\beta}_{t,GKY} \widehat{\beta}_{t,GKY}' \end{array} \right) q = q'(C - [E + F]^{-1} ECE [E + F]^{-1}) q$$

$$\begin{aligned} &= w'(E^{-1} [E + F] C [E + F] E^{-1} - C) w \\ &= w'(\left[I + E^{-1} F \right] C \left[I + F E^{-1} \right] - C) w \\ &= w'(\left[C + E^{-1} F C \right] \left[I + F E^{-1} \right] - C) w \\ &= w'(C + E^{-1} F C + C F E^{-1} + E^{-1} F C F E^{-1} - C) w \\ &= w'(E^{-1} F C + C F E^{-1} + E^{-1} F C F E^{-1}) w \geq 0, \end{aligned}$$

which proves the result.[8] ■

[8] This can also be seen intuitively by noticing that $var\left(\widehat{\beta}_t\right)$ has a lower bound at 0, attained when $\lambda \to \infty$, and an upper bound at $var(\widehat{\beta}_{t,GKY})$, corresponding to $\lambda = 0$. Furthermore $[E + F]^{-1} E$ falls monotonically as λ increases. It follows that $var\left(\widehat{\beta}_t\right) \leq var(\widehat{\beta}_{t,GKY})$ for every positive value of λ

Next, we turn to a comparison of the Mean Squared Error of the constrained and unconstrained non-parametric estimators. We see that the ranking is not clear-cut, unless a certain condition is satisfied.

Following Alkhamisi and Shukur (2008), let us analyze the *canonical*[9] version of model (13). Let Λ and Ψ be the eigenvalues/eigenvectors of $X^{*\prime}_{w,t}X^{*}_{w,t}$, i.e. $X^{*\prime}_{w,t}X^{*}_{w,t} = \Psi\Lambda\Psi'$ and $\Psi\Psi' = I_k$. Defining $\widetilde{y}_t = \sqrt{w_{t,t}}y_t$, $\widetilde{x}_t = \sqrt{w_{t,t}}x_t$, $\widetilde{u}_t = \sqrt{w_{t,t}}y_t$, the *weighted* regression model (in matrix form) is:

$$\begin{aligned} \underset{1\times n}{\widetilde{y}_t'} &= \underset{1\times kk\times n}{\widetilde{x}_t'\,\Theta_t} + \underset{1\times n}{\widetilde{u}_t'}, \\ \underset{k\times n}{\sqrt{\lambda}\overline{r}} &= \underset{k\times k}{\sqrt{\lambda}\overline{R}}\,\underset{k\times n}{\Theta_t} + \underset{k\times n}{\overline{u}^r}. \end{aligned}$$

Using $\Psi\Psi' = I_k$ we can write $\widetilde{z}_t' = \widetilde{x}_t'\Psi$ and $\Xi_t = \Psi'\Theta_t$, and re-state the model in canonical form as:

$$\begin{aligned} \underset{1\times n}{\widetilde{y}_t'} &= \underset{1\times kk\times n}{\widetilde{z}_t'\,\Xi_t} + \underset{1\times n}{\widetilde{u}_t'}, \\ \underset{k\times n}{\sqrt{\lambda}\overline{r}} &= \underset{k\times k}{\sqrt{\lambda}\overline{R}}\,\underset{k\times n}{\Theta_t} + \underset{k\times n}{\overline{u}^r}. \end{aligned}$$

The unconstrained estimator of Ξ_t is then:

$$\begin{aligned} \Xi_t^u &= (Z^{*\prime}_{w,t}Z^{*}_{w,t})^{-1}(Z^{*\prime}_{w,t}Y^{*}_{w,t}) \\ &= \left(\Psi' X^{*\prime}_{w,t}X^{*}_{w,t}\Psi\right)^{-1}(Z^{*\prime}_{w,t}Y^{*}_{w,t}) \\ &= \Lambda^{-1}(Z^{*\prime}_{w,t}Y^{*}_{w,t}), \end{aligned}$$

with $V^u = \left(\frac{1}{H}\Lambda\right)^{-1}\left(\frac{1}{H}Z^{*\prime}_{w,t}W_{H,t}Z^{*\prime}_{ww,t}\right)\left(\frac{1}{H}\Lambda\right)^{-1} \otimes \Sigma_n = \overline{V}^u \otimes \Sigma_n$.The constrained estimator is

$$\Xi_t^c = (\Lambda + \lambda\overline{R}'\overline{R})^{-1}(Z^{*\prime}_{w,t}Y^{*}_{w,t} + \lambda\overline{R}'\overline{r}),$$

with $V^c = \left(\frac{1}{H}\Lambda + \Phi\right)^{-1}\left(\frac{1}{H}Z^{*\prime}_{w,t}W_{H,t}Z^{*\prime}_{ww,t} + \Phi\right)\left(\frac{1}{H}\Lambda + \Phi\right)^{-1} \otimes \Sigma_n = \overline{V}^c \otimes \Sigma_n$, where $\Phi = \frac{\lambda}{H}\overline{R}'\overline{R}$.

To study the conditions under which $\overline{V}^u - \overline{V}^c$ is a semi-positive definite matrix, let us consider equivalently the quadratic form:[10]

$$\begin{aligned} &\beta'(\Lambda+\Phi)\left[\overline{V}^u - \overline{V}^c\right](\Lambda+\Phi)\beta = \\ &= \beta'(\Lambda+\Phi)\Lambda^{-1}\underbrace{Z^{*\prime}_{w,t}W_{H,t}Z^{*\prime}_{ww,t}}_{\equiv A}\Lambda^{-1}(\Lambda+\Phi)\beta - \beta'\underbrace{Z^{*\prime}_{w,t}W_{H,t}Z^{*\prime}_{ww,t}}_{\equiv A} + \Phi\beta \\ &= \beta'\left[\left(I+\Phi\Lambda^{-1}\right)A\left(I+\Lambda^{-1}\Phi\right) - (A+\Phi)\right]\beta \end{aligned}$$

[9] In the canonical form the regressors are orthogonalized, as clarified below.

[10] To simplify the notation, in what follows we have set $H = 1$. This assumption is immaterial, since this term can be factored out.

$$= \beta' \left[A + A\Lambda^{-1}\Phi + \Phi\Lambda^{-1}A + \Phi\Lambda^{-1}A\Lambda^{-1}\Phi - A - \Phi \right] \beta$$
$$= \beta' \left[A\Lambda^{-1}\Phi + \Phi\Lambda^{-1}A + \Phi\Lambda^{-1}A\Lambda^{-1}\Phi - \Phi \right] \beta.$$

Hence, if the quantity $[A\Lambda^{-1}\Phi + \Phi\Lambda^{-1}A + \Phi\Lambda^{-1}A\Lambda^{-1}\Phi - \Phi]$ is positive semi-positive it follows that $\overline{V}^u - \overline{V}^c$ is indeed a semi-positive definite matrix suggesting such a relationship for the mean square errors of the respective estimators. Notice that the required condition is related to the amount of collinearity among the regressors. Specifically, a high degree of collinearity in the time series collected in the matrix $X^*_{w,t}$ will push some eigenvalues of $X^{*\prime}_{w,t}X^*_{w,t}$ close to 0, therefore making Λ^{-1} tend to ∞.

2.5 Time-varying volatilities

When both the error variances and the VAR coefficients change over time, variations in the parameters and in the variances can be confounded, see Cogley and Sargent (2005). An important implication is that if changes in the variances of the errors are neglected then the importance of variation in the VAR coefficients could be overstated. Giraitis, Kapetanios, and Yates (2014) show that the properties of their estimator are unaffected by the presence of stochastic volatilities as long as standard errors are studentized by an appropriate time-varying covariance matrix for the error terms. When performing structural analysis in a VAR context, GKY suggest to model time variation in the variance of the disturbances with a two-step approach. The method consists of fitting first an homoskedastic VAR, then estimating the time-varying volatilities on the residuals obtained in the first stage via the following kernel estimator:

$$\widehat{\Psi}_t = \sum_{j=1}^{T} w_{j,t}(H_\Psi) u_t u_t', \tag{30}$$

where the bandwidth parameter H_Ψ is not necessarily the same as the one used to estimate the coefficients. Orthogonalization of the residuals is then based on the time-varying covariance matrix $\widehat{\Psi}_t$.

Our penalized kernel estimator can be adapted to account for changing volatilities along these lines, using the GLS correction proposed in Theil and Goldberger (1960). In the first step, the VAR coefficients are estimated using (28) and the resulting residuals are used to compute $\widehat{\Psi}_t$ as in (30). In a second step a GLS correction is applied:

$$\beta_t = \left[\sum_{j=1}^{T} w_{j,t} \left(\widehat{\Psi}_j^{-1} \otimes x_j' x_j \right) + \lambda R' R \right]^{-1} \left[\sum_{j=1}^{T} w_{j,t} vec \left(x_j' y_j' \widehat{\Psi}_j^{-1} \right) + \lambda R' r \right] \tag{31}$$

Notice that this GLS correction requires the inversion of potentially large matrices ($nk \times nk$), which slows down computation and limits the size of the VAR. In the empirical applications and in the Monte Carlo analysis discussed in Sections 5 and 4, where we experiment with relatively large systems, we therefore do not apply this correction. However, in Appendix B

we appraise the merits of this GLS correction in the context of a forecast competition that involves 20 variables and quarterly data. In that context, we find that the estimator that does not account for time-varying volatility actually produces more accurate forecasts.

We now turn to discussing how two popular shrinkage methods can be adapted to our setup.

2.6 Ridge type shrinkage

The Ridge regression penalty shrinks all the parameters uniformly towards zero at a given penalty rate λ. A TVP-VAR with Ridge shrinkage can be obtained by setting $R = I_{nk}$ and $r = 0$, which consists of imposing the following stochastic constraints at each t:

$$0 = \sqrt{\lambda}\beta_t + u_t^r. \tag{32}$$

where the properties of u_t^r are as defined in Theorem 1.

The resulting estimator takes the form:

$$\beta_t^{Ridge}(\lambda, H) = \left[I_n \otimes \left(\sum_{j=1}^{T} w_{j,t} x_j' x_j + \lambda I_k \right) \right]^{-1} \left[vec \left(\sum_{j=1}^{T} w_{j,t} x_j' y_j' \right) \right] \tag{33}$$

Notice that, given the Kronecker structure of the constraints (as R is an identity matrix), estimation can proceed equation by equation and the estimator can be written in matrix form as:

$$\Theta_t^{Ridge}(\lambda, H) = \left[X_{w,t}' X_{w,t} + \lambda I_k \right]^{-1} \left[X_{ww,t}' Y \right] \tag{34}$$

2.7 Litterman type shrinkage

Some of the features of the Ridge penalty can be unappealing in the context of VAR models. First, the fact that all the coefficients are shrunk towards zero imposes a structure of serially uncorrelated data, which is at odds with the strong persistence that characterizes most macroeconomic time series. Second, the same penalty is imposed on all the coefficients (including the intercept). Yet, having some flexibility in the penalization of the different parameters could be desirable. A more general set of stochastic constraints, which produce the same effects that the Litterman prior has in a Bayesian framework,[11] is given by setting $\overline{r}$ and $\overline{R}$ as follows:

$$\overline{r} = \begin{pmatrix} \text{diag}(\delta_1\sigma_1, \delta_2\sigma_2, \delta_3\sigma_3, ..., \delta_n\sigma_n) \\ 0_{n(p-1)+1\times n} \end{pmatrix} \qquad \overline{R} = \begin{pmatrix} \overline{\Sigma} & 0 \\ 0 & \sigma_c^2 \end{pmatrix}, \text{ where}$$
$$\overline{\Sigma} = diag(1, 2, 3, ..., p) \otimes diag(\sigma_1, \sigma_2, ..., \sigma_n) \tag{35}$$

[11] Karlsson (2012) distinguishes the Litterman prior from the more general Minnesota prior based on the assumptions on the covariance matrix of the VAR residuals, which is assumed to be diagonal in the Litterman prior, full in the more general Minnesota prior, see Kadiyala and Karlsson (1993, 1997).

Notice that the vector $r = vec(\overline{r})$ towards which the VAR coefficients are driven by the constraints is generally different from zero. In empirical applications, for data in levels, the n values δ_i are typically set to 1, so that the model is pushed to behave like a multivariate random walk plus noise. Moreover, unlike in Ridge regressions, the precision of the constraints is not uniform across parameters but it is higher for more distant lags, as implied by the decay terms $(1, 2, ..., p)$. The scaling factors $\sigma_1^2, ..., \sigma_n^2$ appearing in $\overline{\Sigma}$ can be obtained by univariate regressions and the precision on the intercept σ_c^2 can be set to an arbitrarily small or large value, depending on the application.[12]

Summarizing, by appropriately penalizing the GKY estimator, some discipline on the VAR coefficients can be imposed through stochastic constraints a la Theil and Goldberger (1960). This makes the GKY method, originally designed for small/medium scale VARs, suitable for handling large n dataset. We have seen that the resulting estimator has a well defined asymptotic distribution under rather mild conditions, and is generally more efficient than the unconstrained GKY estimator. Moreover, for popular shrinkage methods the resulting estimator can be cast in matrix form, with notable computational advantages.

The double nature of the estimator (being both nonparametric and penalized) is captured by its dependence on the two constants: H, the bandwidth parameter that determines the weight that each observation has as a function of its distance from t, and λ, a constant that determines the severity of the penalty. In the next section we discuss alternative solutions to the problem of determining these two parameters in empirical applications.

3 Model specification

The problem of setting appropriate values of λ and H can be tackled in two ways. The first is model selection, which typically rests on the optimization of a given criterion. We describe two such criteria. The former adapts to our problem the procedure devised by Banbura, Giannone, and Reichlin (2010), and has an "in sample" fit flavor. The latter favors models with better out of sample performance and is inspired by the method proposed by Kapetanios, Labhard, and Price (2008) for assigning weights to different models in the context of forecast averaging. In the remainder of the paper we will refer to these two criteria as L_{fit} and L_{mse}. The second route consists of pooling the results obtained on the basis of a large range of different specifications. We describe each strategy in turn.

[12]For example, in a Bayesian context, Carriero, Kapetanios, and Marcellino (2009) adopt a very tight prior centered around zero on the intercept in a large VAR, favoring a driftless random walk behavior, to capture the behavior of a panel of exchange rates.

3.1 Model selection criteria

3.1.1 The L_{fit} criterion

The first criterion that we consider adapts to our problem the method by Banbura, Giannone, and Reichlin (2010). The intuition of the method is that, when forecasting with large datasets, some variables are more relevant than others. Over fitting should then be penalized up to the point where a large VAR achieves the same fit as that of a smaller VAR that only includes the key variables of interest. We adapt their criterion to the problem of choosing simultaneously λ and H. Formally, the criterion involves the following steps:

1. Pick a subset of n_1 variables of interest out of the n variables in the VAR.
2. Compute the in sample fit of a benchmark VAR with constant coefficients that only includes these n_1 variables.
3. Select λ and H to minimize the distance between the in sample fit of the large n variate VAR (featuring both time-varying parameters and shrinkage) and the benchmark VAR.

Formally, the loss function to be minimized is the following:

$$L_{fit}(\lambda, H) = \left| \sum_{i=1}^{n_1} \frac{rss_n^i(\lambda, H)}{var(y_{t,i})} - \sum_{i}^{n_1} \frac{rss_{n1}^i}{var(y_{t,i})} \right|$$

where the scaling by $var(y_{t,i})$ is needed to account for the different variance of the variables.

3.1.2 The L_{mse} criterion

As an alternative, λ and H can be selected at each point in time based on the predictive performance of the model in the recent past. The method, which has a cross-validation flavor, is similar in spirit to the one used by Kapetanios, Labhard, and Price (2008) to compute model weights in the context of forecast averaging. The necessary steps are the following:

1. Pick a subset of n_1 variables of interest out of the n variables in the VAR.[13]
2. At each step t in the forecast exercise and for each forecast horizon h consider a relatively short window of recent data $t - L - h, t - L - h + 1, \ldots, t - 1 - h$ and compute the h steps ahead Mean Square Error (MSE) $mse_h^i(\lambda, H)$, for each $i \in n_1$.
3. Pick the values of λ and H that minimize the sum of these n_1 MSEs.

Formally, the loss function to be minimized is:

$$L_{mse}(\lambda, H) = \sum_{i=1}^{n_1} \frac{mse_h^i(\lambda, H)}{var(y_{t,i})}$$

where, again, the mse_i is scaled by the variance of y_i.

[13]Notice that, when using this criterion, we could set $n_1 = n$, that is we could focus on the whole set of variables in the VAR rather than only on a subset.

3.1.3 Practical considerations

In principle, standard optimization algorithms could be used to minimize bot the L_{fit} and the L_{mse} criterion. However, we have often found that the minimum occurs at a kink. A problem of this type could arise because the stochastic constraints shrink the VAR coefficients towards a constant parameter structure but time variation is also affected by the width of the kernel.

Since our estimator is easy to compute a feasible solution is represented by a grid search approach, along the lines of Carriero, Kapetanios, and Marcellino (2009) and Koop and Korobilis (2013). More specifically, in the empirical analysis that follows, we experiment with a wide (38 elements) grid for (the reciprocal) of λ, $\varphi = 1/\lambda$.

$$\varphi_{grid} = 10^{-10}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-2} + .3, 10^{-2} + 2 \times .3, 10^{-2} + 3 \times 03, \ldots, 1 \tag{36}$$

We suggest the use of a wider grid than the one used, for instance, by Koop and Korobilis (2013) because the stochastic constraints that we apply are binding at each point in time (rather than just at the initial condition like in Kalman filter based estimation methods), so that higher values of φ (i.e. lower values of λ) are needed to allow for meaningful time variation in the VAR coefficients.

Regarding the width of the kernel function $w_{j,t}$, we work with a six points grid for the tuning parameter H:

$$H_{grid} = 0.5, 0.6, 0.7, 0.8, 0.9, 1, \tag{37}$$

consistently with the parameterization used in Monte Carlo experiments by Giraitis, Kapetanios, and Price (2013).

3.2 Pooling

An alternative strategy to model selection consists of pooling model estimates obtained with different values of λ and H. This could be particularly valuable in the context of forecasting. From a theoretical standpoint, the rationale for forecast pooling in the presence of structural breaks is offered for example by Pesaran and Pick (2011), who show that averaging forecasts over different estimation windows reduces the forecast bias and mean squared forecast errors, provided that breaks are not too small. In empirical applications pooling is typically found to be effective in improving forecast accuracy both in Bayesian (Koop and Korobilis, 2013) and in frequentist (Kapetanios, Labhard, and Price, 2008) settings.[14] In our context, model pooling could be based on relatively sophisticated weighting schemes, based on the selection criteria described in the previous subsections, or on simpler strategies like equal weights averaging.

[14]Kuzin, Marcellino, and Schumacher (2012) show that forecast pooling also works well in nowcasting GDP.

4 Finite sample properties

To assess the finite sample properties of our estimator we design a Monte Carlo exercise in which we contrast the forecasting performance of the non-parametric estimator with that of a popular parametric alternative.

We consider three alternative DGPs. In the first DGP (DGP-1) we assume that the coefficients follow a random walk plus noise process:

$$Y_t = \Lambda_t Y_{t-1} + \varepsilon_t$$
$$\Lambda_t = \Lambda_{t-1} + \eta_t$$

We make the stochastic process of the coefficients broadly consistent with a Litterman prior by bounding the first autoregressive parameter to lie between 0.85 and 1.[15] In the second one (DGP-2) we let the coefficients break only occasionally rather than at each time t:

$$\Lambda_t = (1 - I(\tau))\Lambda_{t-1} + I(\tau)\Lambda_{t-1} + \eta_t$$

The probability of the coefficients breaking equals a constant τ that we set to .025, implying that, with quarterly data, we would observe on average a discrete break once every ten years. We also relax the bounds on Λ_t and let them fluctuate randomly between 0 and 1.[16] In the third set of simulations (DGP-3) coefficients evolve as a sine functions and are bounded between -1 and 1:

$$\Lambda_t = sin(10\pi t/T) + \eta_t$$

In all DGPs we assume $\eta_t \backsim N(0,1)$ and random walk stochastic volatilities for the measurement equations:

$$\begin{aligned} \varepsilon_{it} &= u_{it}\exp(\lambda_{it}) \\ \lambda_{it} &= \lambda_{it-1} + \nu_{it} \end{aligned}$$

where $u_{it} \backsim N(0,1)$ and $\nu_{it} \backsim N(0,\sigma_\eta)$. We calibrate $\sigma_\eta = 0.01$.[17] For the remaining technical details on the design of the Monte Carlo exercise see Appendix A.1

We assess the performance of our method based on the accuracy of one step ahead forecast errors. As a benchmark, we use the parametric estimator developed by Koop and Korobilis (2013), which we briefly describe in the next sub-section. While the controlled environment

[15] Details on how this is achieved are presented in Appendix A.

[16] With tight boundaries (like the 0.85-1 interval imposed in DGP-1) the difference between coefficients that break only occasionally and coefficients that drift slowly, is negligible.

[17] Notice that this value for σ_η is quite large. Cogley and Sargent (2005) for example, assume that a priori σ_η is distributed as an inverse gamma with a single degree of freedom and scale parameter 0.01^2. Since the scale parameter can be interpreted as the (prior) sum of square residuals, this means that a priori they set the variance of the innovations to the log-volatility to $0.01^2/T$. Assuming $T = 100$, the prior variance is 10^{-6}, as opposed to our choice of 10^{-2}.

provided by the Monte Carlo exercise allows us to evaluate how robust the two methods are to different assumptions on the law of motion of the model parameters, a comparison of the two approaches based on actual data is presented later in section 5.

4.1 A parametric estimator

The model specification adopted by Koop and Korobilis (2013) follows closely the literature on (small) TVP-VARs proposed by Cogley and Sargent (2005) and Primiceri (2005) in that it assumes a random walk evolution of the VAR coefficients. The model can then be cast in State Space, where the VAR equations

$$y_t = Z_t\beta_t + \varepsilon_t \tag{38}$$

serve as measurement equations and the unobserved states, the parameters β_t, evolve as driftless random walks plus noise:

$$\beta_{t+1} = \beta_t + u_{t+1}, \tag{39}$$

with $\varepsilon_t \sim N(0, \Sigma_t)$ and $u_t \sim N(0, Q_t)$. Also ε_t and u_t are independent of one another and serially uncorrelated. Even for medium-sized VARs the estimation algorithms developed by Cogley and Sargent (2005) and Primiceri (2005) become unfeasible due to computational complexity. To overcome these difficulties, following the literature on Adaptive Algorithms, see for example Ljung (1992) and Sargent (1999), Koop and Korobilis (2013) make two simplifying assumptions. The first one involves the matrix Q_t, which is specified as follows:

$$Q_t = \left(\frac{1-\theta}{\theta}\right) P_{t-1/t-1} \tag{40}$$

where $P_{t-1/t-1}$ is the estimated covariance matrix of the unobserved states β_{t-1} conditional on data up to $t-1$ and θ is a forgetting factor $(0 < \theta < 1)$.[18] A similar simplifying assumption on Σ_t ensures that this matrix can be estimated by suitably discounting past squared one step ahead prediction errors:

$$\widehat{\Sigma}_t = \kappa\widehat{\Sigma}_{t-1} + (1-\kappa)v_t v_t' \tag{41}$$

where $v_t = y_t - Z_t\beta_{t/t-1}$. These assumptions make the system matrices Q_t and Σ_t (which are an *input* to the Kalman filter at time t) a function of the t-1 *output* of the Kalman filter itself. This recursive structure implies that, given an initial condition and the two constants θ and κ, an estimate of the coefficients β_t can be obtained through a single Kalman filter pass. Although it is laid out in a Bayesian spirit, the restrictions imposed on the Kalman filter recursions reduce the estimation procedure to a discounted least squares algorithm.

Before moving to the results of the Monte Carlo exercise let us make some remarks on

[18] Equation (40) basically states that the amount of time variation of the model parameters at time t is a small fraction of the uncertainty on the unobserved state β_t, so that large uncertainty on the value of the state at time t translates into stronger parameter time variation

the relative merits of the parametric approach compared to our non-parametric estimator. First, the use of a parametric model, and the simplifications imposed on the model structure to make the estimation feasible, do not come without costs. One potential pitfall is that the model assumes a very specific evolution for the model parameters. The driftless random walk assumption, widely used in econometrics and macroeconomics, does not have any other grounding than parsimony and computational convenience. If the true data generating process (DGP) is, however, very different from the one posited, the model is misspecified and this could result in poor performance. The second issue is that the curse of dimensionality is only partially solved. For 20 variables and 4 lags (a standard application in the large VAR literature with quarterly data) the stacked vector β_t contains 1620 elements. Larger model sizes (arising from a higher number of series in the system or by a higher number of lags, like the 13 lags conventionally used with monthly data in levels) are intractable in this setup. Finally, since the only source of time variation in the model is the prediction error, it can be shown that this forgetting factor algorithm boils down to an exponential smoothing estimator.[19] This means that the effect of the prior on the initial condition β_1 will die out relatively quickly. Also, the longer the sample size, the lower the effect of the prior on the parameter estimates. In contrast, the stochastic constraints that we use to penalize our estimator are effective at each point in time.

4.2 Monte Carlo results

For all the DGPs we fit our non-parametric estimator with Litterman-type stochastic constraints and average forecasts using equal weights across all the possible values of H and λ specified in the grids described in the previous Section. The parametric method also needs a prior on the initial value of the parameters, β_1 to discipline the estimation towards values that are a priori plausible. To keep the comparison with our method as fair as possible we also impose on the initial condition of the parametric model a Litterman type prior. The remaining details of the model specification of the parametric model are quite lengthy and are documented in the Appendix A.2.

The results of the Monte Carlo exercise are shown in Table 4. The methods are compared in terms of 1 step ahead RMSE (relative to that of the parametric estimator and averaged across the n variables using either equal or inverse RMSE weights) for VARs of different sizes ($n = 7$ and $n = 15$) and for different sample sizes (100, 150 and 200). Throughout the exercise forecasts from our proposed estimator are obtained by equal weights averaging across different values for H and λ. Forecasts are computed on the second half of the sample, i.e. when T=100, forecasts are computed recursively for t=51 to t=100, when T=150 forecasts are computed recursively for t=76 to t=150 and so forth.

In the case of DGP-1 the performance of the two estimation methods is broadly comparable,

[19] See Delle Monache and Petrella (2014), Section 2.

with the parametric estimator improving slightly (by at most 2%) on the nonparametric one only for VARs of larger sizes. Notice that in this context one would expect the parametric estimator to have an edge, given the tight correspondence between the assumptions made by the model and the actual DGP. The gain attained by this method proves, however, negligible. When we move to DGP-2 and DGP-3 the relative performance of the nonparametric estimator improves steadily, with gains of the order of 15% in the case of DGP-3 and n=15. Although these DGPs are probably less representative of the typical relationship across macroeconomic time series, they do unveil some fragility of the Kalman filter based method, whose performance rapidly deteriorates when the behavior of the coefficients moves further and further away from the random walk setting. The nonparametric estimator, on the other hand, not only proves robust to heteroschedastic errors but also to a wide range of different specifications of the coefficients.

Summing up, the results of the Monte Carlo analysis are quite supportive of the nonparametric estimator coupled with stochastic constraints. While this does not constitute conclusive evidence in favor of our non-parametric approach, we believe that its good theoretical and finite sample properties, combined with its computational efficiency, make it a very competitive benchmark for modeling and forecasting with large VARs, taking into consideration the possibility of time variation.

5 Forecast Evaluation

After evaluating the finite sample properties of our estimator by means of simulation experiments, we now explore its performance in the context of an extensive forecast exercise based on U.S. data. We first discuss the set-up of the exercise, next we present the results, evaluate the role of forecast pooling and of model size in the TVP context, and finally consider a comparison with the Koop and Korobilis (2013) approach.

5.1 Set-up of the exercise

Throughout the exercise we use Litterman type constraints, like Banbura, Giannone, and Reichlin (2010). The information set is composed of 78 time series spanning around five decades, from January 1959 to July 2013. Table 1 reports the list of the series used in the exercise together with the value of $\overline{r}$ used for each variable. Following the convention in the Bayesian literature we set to 1 the elements of $\overline{r}$ corresponding to variables that display a trend and to 0 those corresponding to variables that have a stationary behavior (typically surveys). We examine the performance of VARs of two sizes. A medium sized VAR that includes only the 20 indicators that are highlighted in red in Table 1 and a large VAR that makes use of all the available information.[20]

[20]Koop and Korobilis (2013) also look at the performance of trivariate VARs with TVP. We do not pursue this route as over fitting is not an issue in small systems and in those cases the use of the unconstrained GKY

We experiment with different model specifications obtained by intersecting various options for setting λ and H as summarized in Table 2. The table is organized in two panels. The top panel refers to model specifications that make use of the L_{fit} criterion, the bottom panel, on the other hand, to specifications based on the L_{mse} criterion. Starting from the top panel, the first set of models (M1 in Table 2) is obtained by fixing H at a given point in the grid and, conditional on this value of H, setting λ optimally at each t at the value that minimizes the L_{fit} function. The second set of models (M2) are obtained as variants of M1 by choosing the λ that minimizes L_{fit} in the pre-sample and then keeping it fixed for the rest of the exercise. In the third set of models (M3) the function L_{fit} is optimized at each t both with respect to λ and H. The fourth case (M4) is obtained as variant of M3 by choosing λ optimally in the pre-sample and then keeping it fixed for the rest of the exercise. The remaining models (M5 to M8) are obtained by replacing the L_{fit} with the L_{mse} criterion. These different model specifications allow us to assess the importance of the various elements that characterize the proposed estimator.

The subset n_1 of variables of interest on which we focus the forecast evaluation is set to $n_1 = 3$, and we monitor the performance of three indicators of particular interest for monetary policy, i.e. the Fed Fund Rates (FEDFUNDS), the number of non farm payroll employees (PAYEMS) and CPI inflation (CPIAUCSL). We fix the lag length to 13 and retain 10 years of data (120 observations) as the first estimation sample. We then produce 1 to 24 months ahead pseudo real time forecasts with the first estimation sample ending in January 1970 and the last one ending in July 2011, for a total number of 499 forecasts. Finally, in the case of the L_{mse} criterion we need to choose L, that is the width of the short window of data on which to measure the predictive performance of the model. We set L =36 (corresponding to three years of data). As a benchmark we adopt the large Bayesian VAR (BVAR) with a Litterman prior and constant coefficients, which can be obtained as a restricted version of our estimator by shutting down the time variation in the VAR coefficients.

5.2 Results

As a first piece of evidence, in Figure 1 (for the 20 variables VAR) and Figure 2 (for the 78 variables VAR) we show the behavior of the penalty parameter λ in the specifications where both λ and H are optimized over time with the L_{fit} criterion (specification M7 in Table 2).[21] As mentioned, high values of λ imply that the constraints hold more tightly, so that the VAR coefficients are less informed by the data. Starting from Figure 1, three distinct phases can be identified. In the first one λ starts from relatively low values and increases smoothly over time. In the 80s and throughout the Great Moderation it stays relatively constant around this value,

estimator is appropriate.

[21] Results obtained using the L_{mse} are qualitatively similar, but for some data points the penalty parameter λ goes to infinity (i.e. the model is driven towards a multivariate random walk) making the visual result less clear.

to start falling again in the mid 1990s, with a steeper slope at the beginning of the 2000s. These results are broadly in line with those stressed in the literature on the predictability of macro times series before and after the Great Moderation. For example D'Agostino, Giannone, and Surico (2006) find that the predictive content for inflation and economic activity of common factors extracted from large panels weakened significantly during the Great Moderation, while in periods of higher volatility cross-sectional information proved more relevant for forecasting. Given the direct relationship between the relevance of cross sectional information and λ, the results in Figure 1 send a similar message, as the contribution of cross-sectional information is progressively penalized by higher values of λ in the 1980s. When the dimension of the VAR increases (Figure 2) the optimal value of λ is higher, confirming the theoretical results in De Mol, Giannone, and Reichlin (2008) on the inverse relationship between the optimal level of shrinkage and cross-sectional dimensions in large panels. An inverse U shaped evolution of λ can be detected also in this case.

To verify that time variation in the coefficients is indeed useful for forecasting, we compare the performance of the 20 variables TVP-VAR with that of its constant coefficient counterpart. The results of this exercise are shown for the various model specifications in Table 3 where we report relative Root Mean Square Forecast Errors (RMSE). Values below 1, which imply that the introduction of time variation through the kernel estimator induces an improvement in prediction accuracy, are highlighted in gray. We assess the statistical significance in forecast accuracy through a Diebold-Mariano test (Diebold and Mariano, 1995) and *underline* the cases in which the null hypothesis can be rejected at the 10% significance. A bird-eye view of the table reveals that in many instances time variation increases forecast accuracy, as the majority of the cells (around 70% of the cases) report values below 1. However, the average improvement appears to be small as in most of the cases the gain is of the order of 5%. As a consequence, most of the differences in forecast accuracy are not significant, according to the Diebold Mariano test. Looking more in detail, three results emerge. First, time variation matters at long horizons for inflation and interest rates, while for employment the improvement is more consistent across different horizons. Second, the specifications that work best are those in which H is fixed at around 0.7 and λ is optimized in real time according to the L_{mse} criterion (M6 in the Table). In this case the TVP-VAR improves on the constant coefficients benchmark by more than 10% at long-horizons. Third, specifications in which both λ and H are optimized in real time (M3 and M8) do not perform well and, in fact, are often outperformed by the benchmark.

5.3 The role of forecast pooling

The substantial heterogeneity observed in the forecasting results across model specifications suggests that the performance of TVP-VAR could be further improved through forecast combination. Since combination schemes based on equal weights are usually found to perform

remarkably well, we proceed by pooling forecasts through simple averaging.[22]

The results obtained by forecast pooling are summarized in Figure 3. The plots, which show the RMSEs of the combined TVP-VARs relative to the fixed coefficients benchmark, are organized in three panels corresponding to the three different target variables, CPI, Fed Fund Rates and employment. The six bars in each panel correspond to different forecast horizons, from 1 to 24 months ahead. Bars in gray identify the forecast horizons for which a Diebold-Mariano test does not reject the null hypothesis of equal forecast accuracy, while those in red denote the cases for which forecast accuracy is significantly different at the 10% confidence level.[23]

The forecasts obtained by pooling predictions from the different time-varying model specifications prove to be more accurate than those obtained from the benchmark at basically all horizons. Furthermore, according to the Diebold Mariano test, the improvement is statistically significant at the 10% confidence level, as evident from the large prevalence of red bars. There is also a tendency of the relative RMSEs to fall as the forecast horizon increases, as it was already apparent in the results displayed in Table 3, suggesting that time variation in the VAR coefficients is relatively more important for forecasting at longer than at shorter horizons. In Figure 4 we report the cumulative sum of squared forecast error differentials, computed as

$$CSSED_t = \sum_{j=1}^{t} (e^2_{j,BVAR} - e^2_{j,TVP-VAR}). \tag{42}$$

This statistics is very useful in revealing the parts of the forecast sample where the TVP-VAR accrues its gains. Positive and increasing values indicate that the TVP model outperforms the benchmark, while negative and decreasing values suggest the opposite. At relatively shorter horizons (top panels) the model with time-varying coefficients performs better than the one with constant parameters around economic downturns, as indicated by the jumps of the CSSED in periods classified by the NBER as recessions (gray shaded areas). At longer horizons (bottom panels), the gain is relatively uniform across the sample for interest rates and employment, while it is relatively concentrated in the 70s-80s for inflation.

5.4 The role of model size

To answer the question of whether enlarging the information set eliminates the need for time variation in the coefficients, we compare the performance of the 20 variables TVP-VAR with that of a fixed coefficients BVAR with 78 variables. The relative RMSEs reported in Figure 5 show that at shorter horizons (1 to 6 months ahead) the performance of the two models is overall comparable, although the time-varying model is more accurate in tracking

[22]More sophisticated weighting schemes, based on the selection criteria described in Table 2, deliver very similar results. The analysis is available upon request.

[23]The test is two sided so that bars in red and higher than 1 indicate that the forecast of the benchmark model is significantly more accurate.

interest rates. However, when we move to longer horizons, the performance of the TVP-VAR improves considerably. Looking at Figure 6 we find again that the CSSED tends to jump around recession periods. Hence, the importance of TVP is not (mainly) due to omitted variables.

The next issue that we want to explore is whether, in the context of a TVP-VAR, it pays off to go larger than around 20 variables, provided that the set of variables of interest is small. We tackle this question by comparing the performance of the 20 variables TVP-VAR with that of a 78 variables TVP-VAR. We find that, on the whole forecast sample, a medium-sized information set is sufficient to capture the relevant dynamics. The predictive accuracy of the 20 variables VAR is, in fact, typically higher than that of the larger model, especially for interest rates (see Figure 7). The evolution of the CSSED, shown in Figure 8, reveals that the accuracy gains of the 20 variables VAR are actually concentrated in the first part of the sample, and that from the 90s onwards, the performance of the two model sizes is very similar. This is an interesting finding that extends to a time-varying coefficients context the results obtained by Banbura, Giannone, and Reichlin (2010) in the case of constant coefficient VARs and those by Boivin and Ng (2006) in constant coefficient factor models.

5.5 Comparison with Koop and Korobilis (2013)

A comparison of the empirical performance of the nonparametric and parametric TVP-VAR needs to take into account the computational limitations to which the latter is subject. This means that a forecast competition based on monthly VARs with 13 lags, like those employed in the previous subsections, is unfeasible. We therefore proceed by taking quarterly transformations of the variables and specify a 20 variables VAR with 4 lags. The forecast exercise is similar to the one performed on monthly data, that is we produce 1 to 8 quarters ahead forecasts of the three key variables in our dataset, CPI, the Fed Fund Rates and payroll employment, with an out sample period ranging from 1970:q1 to 2013:q2 (167 data points).

Figure 9 presents the RMSEs of the kernel based estimator relative to those of the parametric one. Again, we use red bars to highlight the cases where a Diebold-Mariano test rejects the null hypothesis of equal forecast accuracy. Visual inspection of the graph reveals that the nonparametric estimator generates significantly better predictions for inflation and employment, while the parametric estimator is more accurate in forecasting short term interest rates. As for the remaining variables, the only case in which the Diebold-Mariano test rejects in favor of the parametric estimator is for the 10 year rate and for M1 at very short horizons, while for the remaining 10 indicators the evidence is either in favor of the nonparametric approach (red bars lower than one) or inconclusive (gray bars).

Summarizing, the outcome of this extensive forecasting exercise provides further broad support for our method. Moreover, the fact that the estimator can accommodate a large information set has allowed us to address issues that could not be investigated with existing methods, such as the relationship between the size of the information set and parameters'

time variation and the relevance of the model size in the context of models with time-varying parameters.

6 Structural analysis

Our non-parametric estimator can be useful also in the context of structural analysis when time variation in the parameters is considered to be an issue. As an illustration, we use the proposed method to estimate the time-varying responses of industrial production indexes to an unexpected increase in the price of oil. The changing response of key macroeconomic variables to unexpected oil price increases has been greatly debated in the past decade. In particular, using structural VARs and different identification assumptions a number of studies have found that oil price increases are associated with smaller losses in U.S. output in more recent years. While some of these studies have used sample-split approaches, like Edelstein and Kilian (2009), Blanchard and Gali (2007) and Blanchard and Riggi (2013), others have relied on Bayesian VARs with drifting coefficients and volatilities, see Baumeister and Peersman (2013) and Hahn and Mestre (2011). The latter approach, however, severely constraints the size of the system to be estimated so that only a small number of variables can be jointly modeled. Partly as a consequence of this constraint, available evidence on the break in the oil/output nexus mainly refers to aggregate GDP. Sectoral aspects, however, are equally relevant as the recessionary effect of oil price shocks is partly due to a costly reallocation of labor and capital away from energy intensive sectors (Davis and Haltiwanger, 2001). Where a more granular perspective is taken, like in Edelstein and Kilian (2009), special attention is paid to the role of the automotive sector, which is considered the main transmission channel of energy price shocks. Indeed as energy price increases reduce purchases of cars, and given that the dollar value of these purchases is large relatively to the energy they use, even small energy price shocks can cause large effects, an intuition formalized by Hamilton (1988). Given the importance of this sector, one would expect it to be the main responsible for the changing relationship between oil and GDP.

In this section we revisit this issue by extending the analysis conducted in Edelstein and Kilian (2009), based on a bivariate VAR and on a sample-split approach, to a large TVP-VAR setting in which energy prices are modeled jointly with industrial output in different sectors. In particular, we augment our baseline 20 variables VAR with 8 industrial production series split by market destination.[24] The additional series are Business Equipment, Consumer Goods and its two sub-components Durable (half of which is accounted for by Automotive products) and Nondurable (Food, Clothing, Chemical and Paper products), Final Product goods (Construction and Business Supplies and Defense and Space equipment), Material goods and its two

[24]Since we are not concerned with forecasting, in the structural analysis presented in this Section we follow Giraitis, Kapetanios, and Yates (2012) and use a two sided Gaussian kernel with smoothing parameter $H = 0.5$. Furthermore, the penalty parameter λ is chosen over the full sample through the L_{fit} criterion and changing volatilities are accounted for through the GLS correction in equation (31).

sub-components Durable (Consumer and Equipment parts) and Nondurable (Textile, Paper and Chemical). A list of the series, together with their weight on the overall index, is reported in Table 5.

The identification of energy price shocks follows Edelstein and Kilian (2009), i.e. we assume that energy price shocks are exogenous relative to contemporaneous movements in the other variables in the system, which implies ordering the price of oil first in a recursive structural VAR.[25] Kilian and Vega (2011) provide a test of this assumption by regressing daily changes in the price of oil to daily news on a wide range of macroeconomic data and find no evidence of feedback from macroeconomic news to energy prices, concluding that energy prices are indeed predetermined with respect to the U.S. macroeconomy. A shortcoming of this approach is that it does not allow us to separate the source of variation behind oil price shocks, i.e. whether they are driven by supply rather than by demand.[26] In other words, our identified energy price shocks will be a linear combination of demand and supply shocks. However, given that we are interested in identifying the sectors that are central to the propagation of energy price shocks, rather than in determining the determinants of energy price fluctuations, the recursive identification assumption is appropriate.

Figure 10 shows that an innovation to the real price of oil generates a protracted fall in overall industrial output. Furthermore, in line with the literature, the recessionary impact of an exogenous oil price disturbance is generally more severe in the Seventies than in later decades. Notice, however that the difference across the two sub-samples is entirely accounted for by the very early Seventies, a finding that cannot be uncovered with the simple sample-split strategy considered in Edelstein and Kilian (2009), Blanchard and Gali (2007) and Blanchard and Riggi (2013) and that validates the use of time-varying coefficients models.

The results for the individual sectors are reported in Figure 11. A number of interesting results emerge. First, in most sectors the effect of an unexpected increase in the real price of oil is generally negative in the first part of the sample, and the fall in production is much more pronounced in energy intensive segments, like Business Equipment, Durable Consumption and Material Goods. Second, most sectors display an attenuation of the recessionary impact of energy price shocks. In some of them unexpected increases in the real price of oil end up being associated with an expansion in production, consistently with the findings in Kilian (2009) that attribute energy price surprises in the 2000s to increased demand for commodities rather than to supply disruptions.[27] Again, most of the changes over time occur in more energy intensive sectors.

To assess the relative importance of each sector in explaining the changing pass-through of energy price shocks to overall industrial activity we proceed by weighing the IRFs in different

[25] A similar identification assumption is maintained by Blanchard and Gali (2007) and Blanchard and Riggi (2013).

[26] The debate on the relative role of supply and demand factors in determining oil prices dates back to Kilian (2009).

[27] Blanchard and Gali (2007) also find that oil price innovations are associated with an increase in output after the 80s in France and in Germany, see Figure 7.6 therein.

sectors by their shares in overall industrial output (reported in Table 5). The resulting weighted IRFs are reported in Figure 12 where, for the sake of clarity, we only focus on the responses twelve months after the initial shock. When the relative weight of the various sectors is taken into account, the relevance for overall business cycle fluctuations of developments in the motor vehicles sector, which accounts for half of Durable Consumption, appears less relevant than that of other sectors. Instead, the response of overall industrial output to oil price shocks and its evolution over time are largely determined by that of the Durable Material sector, which includes intermediate goods for a wide range of final products. This outcome suggests that the increased efficiency in the energy use of automobiles has played a minor role in shaping the oil/output relationship in the U.S. over the past forty years. In turn, greater energy efficiency at the higher stages of the supply chain, as well as a larger role for demand shocks, are likely to be the driving forces behind changes in the relationship between oil prices and U.S. aggregate output.

7 Conclusions

In this paper we propose an estimator for large dimensional VAR models with flexible parameter structure, capable of accommodating breaks in the relationships among economic time series. Our procedure is based on the mixed estimator by Theil and Goldberger (1960), which imposes stochastic constraints on the model coefficients, and on the nonparametric VAR estimator proposed by Giraitis, Kapetanios, and Yates (2014). The use of stochastic constraints mimics in a classical context the role of the prior in Bayesian models, allowing to bypass the over-fitting problem that arises in large dimensional models.

We derive the asymptotic distribution of the estimator and evaluate the determinants of its efficiency. We also discuss various aspects of the practical implementation of the estimator, based on two alternative (fit and forecasting) criteria, and assess its finite sample performance in Monte Carlo experiments.

We then use the non-parametric estimator in a forecasting exercise where we model up to 78 U.S. macroeconomic time series. We find that the introduction of time variation in the VAR model parameters yields an improvement in prediction accuracy over models with a constant parameter structure, in particular when forecast combination is used to pool forecasts obtained with models with different degrees of time variation and penalty parameters. We also shed light on an issue that is central to the forecasting literature, namely how the size of the information set interacts with time variation in the model parameters. Specifically, we find that the relevance of time variation is not related to omitted variable problem and that, as in the constant parameter case, a medium-sized TVP-VAR is at least as good as a large TVP-VAR.

In a forecasting context, our non-parametric estimator compares well with the alternative parametric approach by Koop and Korobilis (2013), when using either actual or simulated data,

and can handle a larger number of variables.

Finally, to illustrate the use of our method in structural analysis, we analyze the changing effects of oil price shocks on economic activity, a question that has spurred a large number of studies in the applied macro literature in recent years. We find that the declining role of oil prices in shaping U.S. business cycle fluctuations stems from changes related to Business Equipment and Materials sector, rather than from the automobiles sector as argued by part of the literature.

Overall, we believe that our findings illustrate how the econometric tool that we have proposed opens the door to a number of interesting analyses on forecasting and on the nonlinear transmission of shocks, which have been so far constrained by computational issues.

References

Aastveit, K. A., A. Carriero, T. E. Clark, and M. Marcellino (2014): "Have Standard VARs Remained Stable since the Crisis?," Working Paper 1411, Federal Reserve Bank of Cleveland.

Alkhamisi, M., and G. Shukur (2008): "Developing Ridge Parameters for SUR Model," *Communications In Statistics-Theory And Methods*, 37(4), 544–564.

Banbura, M., D. Giannone, and L. Reichlin (2010): "Large Bayesian vector auto regressions," *Journal of Applied Econometrics*, 25(1), 71–92.

Baumeister, C., and G. Peersman (2013): "Time-Varying Effects of Oil Supply Shocks on the US Economy," *American Economic Journal: Macroeconomics*, 5(4), 1–28.

Benati, L., and H. Mumtaz (2007): "U.S. evolving macroeconomic dynamics: a structural investigation," Working Paper Series 0746, European Central Bank.

Benati, L., and P. Surico (2008): "Evolving U.S. Monetary Policy and The Decline of Inflation Predictability," *Journal of the European Economic Association*, 6(2-3), 634–646.

Blanchard, O. J., and J. Gali (2007): "The Macroeconomic Effects of Oil Price Shocks: Why are the 2000s so different from the 1970s?," in *International Dimensions of Monetary Policy*, NBER Chapters, pp. 373–421. National Bureau of Economic Research, Inc.

Blanchard, O. J., and M. Riggi (2013): "Why are the 2000s so different from the 1970s? A structural interpretation of changes in the macroeconomic effects of oil prices," *Journal of the European Economic Association*, 11(5), 1032–1052.

Boivin, J., and S. Ng (2006): "Are more data always better for factor analysis?," *Journal of Econometrics*, 132(1), 169–194.

Canova, F., and L. Gambetti (2010): "Do Expectations Matter? The Great Moderation Revisited," *American Economic Journal: Macroeconomics*, 2(3), 183–205.

Carriero, A., T. Clark, and M. Marcellino (2016): "Large Vector Autoregressions with asymmetric priors and time-varying volatilities," manuscript, Queen Mary University of London.

Carriero, A., G. Kapetanios, and M. Marcellino (2009): "Forecasting exchange rates with a large Bayesian VAR," *International Journal of Forecasting*, 25(2), 400–417.

Cogley, T., and T. J. Sargent (2005): "Drift and Volatilities: Monetary Policies and Outcomes in the Post WWII U.S," *Review of Economic Dynamics*, 8(2), 262–302.

D'Agostino, A., D. Giannone, and P. Surico (2006): "(Un)Predictability and macroeconomic stability," Working Paper Series 0605, European Central Bank.

Davis, S. J., and J. Haltiwanger (2001): "Sectoral job creation and destruction responses to oil price changes," *Journal of Monetary Economics*, 48(3), 465–512.

De Mol, C., D. Giannone, and L. Reichlin (2008): "Forecasting using a large number of predictors: Is Bayesian shrinkage a valid alternative to principal components?," *Journal of Econometrics*, 146(2), 318–328.

Delle Monache, D., and I. Petrella (2014): "Adaptive Models and Heavy Tails," Birkbeck Working Papers in Economics and Finance 1409, Birkbeck, Department of Economics, Mathematics & Statistics.

Diebold, F. X., and R. S. Mariano (1995): "Comparing Predictive Accuracy," *Journal of Business & Economic Statistics*, 13(3), 253–63.

Edelstein, P., and L. Kilian (2009): "How sensitive are consumer expenditures to retail energy prices?," *Journal of Monetary Economics*, 56(6), 766–779.

Eickmeier, S., W. Lemke, and M. Marcellino (2015): *Journal of the Royal Statistical Society, Series A, forthcoming.*

Giraitis, L., G. Kapetanios, and S. Price (2013): "Adaptive forecasting in the presence of recent and ongoing structural change," *Journal of Econometrics*, 177(2), 153–170.

Giraitis, L., G. Kapetanios, K. Theodoridis, and T. Yates (2014): "Estimating time-varying DSGE models using minimum distance methods," Bank of England working papers 507, Bank of England.

Giraitis, L., G. Kapetanios, and T. Yates (2012): "Inference on multivariate stochastic time varying coefficient and variance models," mimeo.

Giraitis, L., G. Kapetanios, and T. Yates (2014): "Inference on stochastic time-varying coefficient models," *Journal of Econometrics*, 179(1), 46–65.

Hahn, E., and R. Mestre (2011): "The role of oil prices in the euro area economy since the 1970s," Working Paper Series 1356, European Central Bank.

Hamilton, J. D. (1988): "A Neoclassical Model of Unemployment and the Business Cycle.," *Journal of Political Economy*, 96(2), 593–617.

——— (2009): "Causes and Consequences of the Oil Shock of 2007-08," *Brookings Papers on Economic Activity*, 40(1 (Spring), 215–283.

Hastie, T., R. Tibshirani, and J. Friedman (2003): *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, no. 9780387952840 in Statistics. Springer.

Hooker, M. A. (1999): "Oil and the macroeconomy revisited," Discussion paper.

Kapetanios, G., V. Labhard, and S. Price (2008): "Forecasting Using Bayesian and Information-Theoretic Model Averaging: An Application to U.K. Inflation," *Journal of Business & Economic Statistics*, 26, 33–41.

Kilian, L. (2009): "Not All Oil Price Shocks Are Alike: Disentangling Demand and Supply Shocks in the Crude Oil Market," *American Economic Review*, 99(3), 1053–69.

Kilian, L., and C. Vega (2011): "Do Energy Prices Respond to U.S. Macroeconomic News? A Test of the Hypothesis of Predetermined Energy Prices," *The Review of Economics and Statistics*, 93(2), 660–671.

Koop, G., and D. Korobilis (2013): "Large time-varying parameter VARs," *Journal of Econometrics*, 177(2), 185–198.

Koop, G. M. (2013): "Forecasting with Medium and Large Bayesian VARS," *Journal of Applied Econometrics*, 28(2), 177–203.

Kuzin, V., M. Marcellino, and C. Schumacher (2012): "Pooling versus model selection for nowcasting with many predictors: An application to German GDP," *Journal of Applied Econometrics*, XX(XX), XX.

Ljung, L. (1992): "Applications to Adaptive Algorithms,," in *Stochastic Approximations and Op timization of Random Systems, edited by L. Ljung, Georg Pflug, and Harro Walk*, pp. 95–113. Birkhauser.

Mumtaz, H., and P. Surico (2012): "Evolving International Inflation Dynamics: World And Country-Specific Factors," *Journal of the European Economic Association*, 10(4), 716–734.

Pesaran, M. H., and A. Pick (2011): "Forecast Combination Across Estimation Windows," *Journal of Business & Economic Statistics*, 29(2), 307–318.

Pesaran, M. H., and A. Timmermann (2007): "Selection of estimation window in the presence of breaks," *Journal of Econometrics*, 137(1), 134–161.

Primiceri, G. (2005): "Time Varying Vector Autoregressions and Monetary Policy," *The Review of Economic Studies*, 1(19), 465–474.

Raftery, P., M. Karny, and P. Ettler (2010): "ï£¡Online prediction under model uncertainty via dynamic model averaging: Application to a cold rolling mill," *Technometrics*, 52, 52–66.

Rossi, B., A. Inoue, and L. Jin (2014): "Optimal Window Selection in the Presence of Possible Instabilities," Discussion Paper 10168, CEPR.

Sargent, T. J. (1999): *The Conquest of American Inflation*, Cambridge Books. Princeton University Press.

Stock, J. H., and M. W. Watson (2012): "Disentangling the Channels of the 2007-09 Recession," *Brookings Papers on Economic Activity*, 44(1 (Spring), 81–156.

Theil, H., and A. Goldberger (1960): "On pure and mixed statistical estimation in econometrics," *International Economic Review*, 2, 65–78.

No.	Acronym (FRED database)	Description	SA	Logs	Prior mean
1	AAA	Interest rates on AAA bonds	Not Seasonally Adjusted	0	1
2	AHEMAN	Average Hourly Earnings Of Production And Nonsupervisory Employees: Manufacturing	Not Seasonally Adjusted	1	1
3	AWHMAN	Average Weekly Hours of Production and Nonsupervisory Employees:	Seasonally Adjusted	1	0
4	AWOTMAN	Average Weekly Overtime Hours of Production and Nonsupervisory	Seasonally Adjusted	1	0
5	BAA	Interest rates on BAA bonds	Not Seasonally Adjusted	1	1
6	CE16OV	Civilian Employment	Seasonally Adjusted	1	1
7	CPIAPPSL	Consumer Price Index for All Urban Consumers: Apparel	Seasonally Adjusted	1	1
8	CPIAUCSL	Consumer Price Index for All Urban Consumers: All Items	Seasonally Adjusted	1	1
9	CPILFESL	Consumer Price Index for All Urban Consumers: All Items Less Food &	Seasonally Adjusted	1	1
10	CPIMEDSL	Consumer Price Index for All Urban Consumers: Medical Care	Seasonally Adjusted	1	1
11	CPITRNSL	Consumer Price Index for All Urban Consumers: Transportation	Seasonally Adjusted	1	1
12	CPIULFSL	Consumer Price Index for All Urban Consumers: All Items Less Food	Seasonally Adjusted	1	1
13	DMANEMP	All Employees: Durable goods	Seasonally Adjusted	1	0
14	DSPIC96	Real Disposable Personal Income	Seasonally Adjusted	1	1
15	DPCERA3M086SBEA	Real personal consumption expenditures (chain-type quantity index)	Seasonally Adjusted	1	1
16	FEDFUNDS	Effective Federal Funds Rate	Not Seasonally Adjusted	0	1
17	GS1	1-Year Treasury Constant Maturity Rate	Not Seasonally Adjusted	0	1
18	GS10	10-Year Treasury Constant Maturity Rate	Not Seasonally Adjusted	0	1
19	GS5	5-Year Treasury Constant Maturity Rate	Not Seasonally Adjusted	0	1
20	HOUST	Housing Starts: Total: New Privately Owned Housing Units Started	Seasonally Adjusted Annual Rate	1	0
21	HOUSTMW	Housing Starts in Midwest Census Region	Seasonally Adjusted Annual Rate	1	0
22	HOUSTNE	Housing Starts in Northeast Census Region	Seasonally Adjusted Annual Rate	1	0
23	HOUSTS	Housing Starts in South Census Region	Seasonally Adjusted Annual Rate	1	0
24	HOUSTW	Housing Starts in West Census Region	Seasonally Adjusted Annual Rate	1	0
25	INDPRO	Industrial Production Index	Seasonally Adjusted	1	1
26	IPBUSEQ	Industrial Production: Business Equipment	Seasonally Adjusted	1	1
27	IPCONGD	Industrial Production: Consumer Goods	Seasonally Adjusted	1	1
28	IPDCONGD	Industrial Production: Durable Consumer Goods	Seasonally Adjusted	1	1
29	IPDMAT	Industrial Production: Durable Materials	Seasonally Adjusted	1	1
30	IPFINAL	Industrial Production: Final Products (Market Group)	Seasonally Adjusted	1	1
31	IPMAT	Industrial Production: Materials	Seasonally Adjusted	1	1
32	IPNCONGD	Industrial Production: Nondurable Consumer Goods	Seasonally Adjusted	1	1
33	IPNMAT	Industrial Production: nondurable Materials	Seasonally Adjusted	1	1
34	LOANS	Loans and Leases in Bank Credit, All Commercial Banks	Seasonally Adjusted	1	1
35	M1SL	M1 Money Stock	Seasonally Adjusted	1	1
36	M2SL	M2 Money Stock	Seasonally Adjusted	1	1
37	MANEMP	All Employees: Manufacturing	Seasonally Adjusted	1	0
38	NAPM	ISM Manufacturing: PMI Composite Index	Seasonally Adjusted	0	0
39	NAPMEI		Seasonally Adjusted	0	0
40	NAPMII	[]	Not Seasonally Adjusted	0	0
41	NAPMNOI	ISM Manufacturing: New Orders Index	Seasonally Adjusted	0	0
42	NAPMPI	[]	Seasonally Adjusted	0	0
43	NAPMSDI	[]	Seasonally Adjusted	0	0
44	NDMANEMP	All Employees: Nondurable goods	Seasonally Adjusted	1	0
45	OILPRICE	[]	Not Seasonally Adjusted	1	1
46	PAYEMS	All Employees: Total nonfarm	Seasonally Adjusted	1	1
47	PCEPI	Personal Consumption Expenditures: Chain-type Price Index	Seasonally Adjusted	1	1
48	PERMIT	New Private Housing Units Authorized by Building Permits	Seasonally Adjusted Annual Rate	1	0
49	PERMITMW	New Private Housing Units Authorized by Building Permits in the	Seasonally Adjusted Annual Rate	1	0
50	PERMITNE	New Private Housing Units Authorized by Building Permits in the	Seasonally Adjusted Annual Rate	1	0
51	PERMITS	New Private Housing Units Authorized by Building Permits in the South	Seasonally Adjusted Annual Rate	1	0
52	PERMITW	New Private Housing Units Authorized by Building Permits in the West	Seasonally Adjusted Annual Rate	1	0
53	PI	Personal Income	Seasonally Adjusted Annual Rate	1	1
54	PPIACO	Producer Price Index: All Commodities	Not Seasonally Adjusted	1	1
55	PPICRM	Producer Price Index: Crude Materials for Further Processing	Seasonally Adjusted	1	1
56	PPIFCG	Producer Price Index: Finished Consumer Goods	Seasonally Adjusted	1	1
57	PPIFGS	Producer Price Index: Finished Goods	Seasonally Adjusted	1	1
58	PPIITM	Producer Price Index: Intermediate Materials: Supplies & Components	Seasonally Adjusted	1	1
59	SandP	S&P 500 Stock Price Index		1	1
60	SRVPRD	All Employees: Service-Providing Industries	Seasonally Adjusted	1	1
61	TB3MS	3-Month Treasury Bill: Secondary Market Rate	Not Seasonally Adjusted	0	1
62	TB6MS	6-Month Treasury Bill: Secondary Market Rate	Not Seasonally Adjusted	0	1
63	UEMP15OV	Number of Civilians Unemployed for 15 Weeks & Over	Seasonally Adjusted	1	0
64	UEMP15T26	Number of Civilians Unemployed for 15 to 26 Weeks	Seasonally Adjusted	1	0
65	UEMP27OV	Number of Civilians Unemployed for 27 Weeks and Over	Seasonally Adjusted	1	0
66	UEMP5TO14	Number of Civilians Unemployed for 5 to 14 Weeks	Seasonally Adjusted	1	0
67	UEMPLT5	Number of Civilians Unemployed - Less Than 5 Weeks	Seasonally Adjusted	1	0
68	UEMPMEAN	Average (Mean) Duration of Unemployment	Seasonally Adjusted	1	0
69	UNRATE	Civilian Unemployment Rate	Seasonally Adjusted	0	0
70	USCONS	All Employees: Construction	Seasonally Adjusted	1	1
71	USFIRE	All Employees: Financial Activities	Seasonally Adjusted	1	1
72	USGOOD	All Employees: Goods-Producing Industries	Seasonally Adjusted	1	0
73	USGOVT	All Employees: Government	Seasonally Adjusted	1	1
74	USMINE	All Employees: Mining and logging	Seasonally Adjusted	1	0
75	USPRIV	All Employees: Total Private Industries	Seasonally Adjusted	1	1
76	USTPU	All Employees: Trade, Transportation & Utilities	Seasonally Adjusted	1	1
77	USTRADE	All Employees: Retail Trade	Seasonally Adjusted	1	1
78	USWTRADE	All Employees: Wholesale Trade	Seasonally Adjusted	1	1

Table 1: Data description

		λ	H
L_{fit}	M1	Optimized at each t	0.5 0.6 0.7 0.8 0.9 1
	M2	Fixed at pre-sample optimal level	0.5 0.6 0.7 0.8 0.9 1
	M3	Optimized at each t	Optimized at each t
	M4	Fixed at pre-sample optimal level	Optimized at each t
L_{mse}	M6	Optimized at each t	0.5 0.6 0.7 0.8 0.9 1
	M7	Fixed at pre-sample optimal level	0.5 0.6 0.7 0.8 0.9 1
	M8	Optimized at each t	Optimized at each t
	M9	Fixed at pre-sample optimal level	Optimized at each t

Table 2: Specifications for the TVP-VARs. The L_{fit} criterion is computed as $L_{fit}(\lambda, H) = \left|\sum_{i=1}^{n_1} \frac{rss_n^i(\lambda,H)}{var(y_{t,i})} - \sum_i^{n_1} \frac{rss_{n1}^i}{var(y_{t,i})}\right|$ where n_1 is a number of reference variables, rss_{n1} is the residual sum of squares obtained with an $n1$ variate VAR, and $rss_{n1}(\lambda, H)$ is the residual sum of squares obtained with the TVP-VAR. The L_{mse} criterion is computed as $L_{mse}(\lambda, H) = \sum_{i=1}^{n_1} \frac{mse_h^i(\lambda,H)}{var(y_{t,i})}$, where mse_h is the mean square prediction error h steps ahead obtained with the TVP-VAR.

Selection method		λ	H	CPI				Fed Funds Rates				Employment			
				Forecast horizon				Forecast horizon				Forecast horizon			
				1	6	12	24	1	6	12	24	1	6	12	24
L_{fit}	M1	Optimized	0.5	1.17	1.10	1.03	0.92	1.07	0.85	0.84	0.94	1.40	1.33	1.13	0.90
			0.6	1.12	1.01	1.02	0.99	1.02	0.90	0.93	0.96	1.28	1.25	1.12	0.90
			0.7	1.03	0.97	0.99	1.03	0.97	0.97	0.99	0.98	0.96	0.93	0.95	0.96
			0.8	1.01	0.98	0.99	0.97	0.99	0.97	0.96	0.95	0.97	0.94	0.95	0.95
			0.9	1.02	0.98	0.98	0.96	1.00	0.96	0.95	0.94	0.98	0.96	0.97	0.96
			1	1.02	0.98	0.98	0.95	1.00	0.96	0.95	0.93	0.98	0.97	0.98	0.96
	M2	Fixed	0.5	1.03	1.04	1.08	1.33	0.95	0.97	1.03	1.27	0.97	0.94	0.98	0.98
			0.6	1.00	0.97	1.03	1.13	0.93	1.02	1.07	1.12	0.97	0.94	0.98	1.04
			0.7	1.00	0.98	1.01	1.05	0.95	0.99	1.01	0.99	0.97	0.95	0.98	1.01
			0.8	1.01	0.99	0.99	0.99	0.97	0.98	0.97	0.96	0.97	0.96	0.97	0.98
			0.9	1.02	0.99	0.98	0.96	0.98	0.97	0.95	0.94	0.98	0.97	0.98	0.98
			1	1.02	0.99	0.98	0.95	0.98	0.96	0.95	0.94	0.98	0.98	0.99	0.98
	M3	Optimized	OPT	1.02	0.99	1.01	1.02	0.99	1.01	1.00	0.95	0.96	0.93	0.94	0.92
	M4	Fixed	OPT	1.01	0.98	1.01	1.03	0.96	1.01	1.01	0.96	0.98	0.97	1.00	1.02
L_{mse}	M6	Optimized	0.5	1.06	1.09	1.10	1.62	1.01	0.98	0.98	1.49	1.16	1.09	1.07	1.06
			0.6	1.08	1.00	0.89	0.84	0.99	0.96	0.90	0.98	1.12	1.05	0.98	0.93
			0.7	1.12	1.03	0.95	0.88	1.01	0.91	0.87	0.89	1.16	1.10	0.99	0.88
			0.8	1.15	1.06	0.98	0.88	1.01	0.88	0.84	0.85	1.18	1.13	1.00	0.84
			0.9	1.14	1.07	0.98	0.88	1.02	0.87	0.84	0.86	1.18	1.15	1.03	0.88
			1	1.14	1.05	0.98	0.88	1.02	0.87	0.84	0.86	1.19	1.16	1.04	0.90
	M7	Fixed	0.5	1.01	1.08	1.26	2.13	0.93	1.02	1.16	2.03	0.99	0.97	1.02	1.15
			0.6	1.01	0.97	1.02	1.11	0.93	1.00	1.05	1.11	0.97	0.93	0.96	1.01
			0.7	1.01	0.97	0.97	0.99	0.95	0.97	0.97	0.96	0.96	0.93	0.94	0.92
			0.8	1.03	0.98	0.96	0.93	0.98	0.94	0.93	0.92	0.98	0.95	0.94	0.89
			0.9	1.03	0.98	0.96	0.92	1.00	0.94	0.92	0.91	1.00	0.98	0.97	0.92
			1	1.04	0.98	0.95	0.92	1.00	0.93	0.92	0.91	1.00	0.99	0.98	0.92
	M8	Optimized	OPT	1.10	1.10	1.11	1.61	1.01	0.97	0.93	1.42	1.15	1.06	1.00	1.02
	M9	Fixed	OPT	0.99	1.04	1.16	2.08	0.95	1.00	1.02	1.99	0.99	0.92	0.95	1.10

Table 3: Root Mean Square Forecast Errors: TVP-VARs versus constant coefficients BVAR (20 variables). The tables show the RMSE obtained by models with time varying parameters described in Table 2 relative to those obtained with the benchmark large BVAR with constant parameters. Values below 1 (shaded in grey in the table) imply that the model outperforms the benchmark. Values underlined indicate the cases in which the Diebold Mariano test rejects the null hypothesis of equal forecast accuracy at the 10% confidence level. The RMSE are computed on 499 out-of-sample forecast errors, from January 1970 to July 2013.

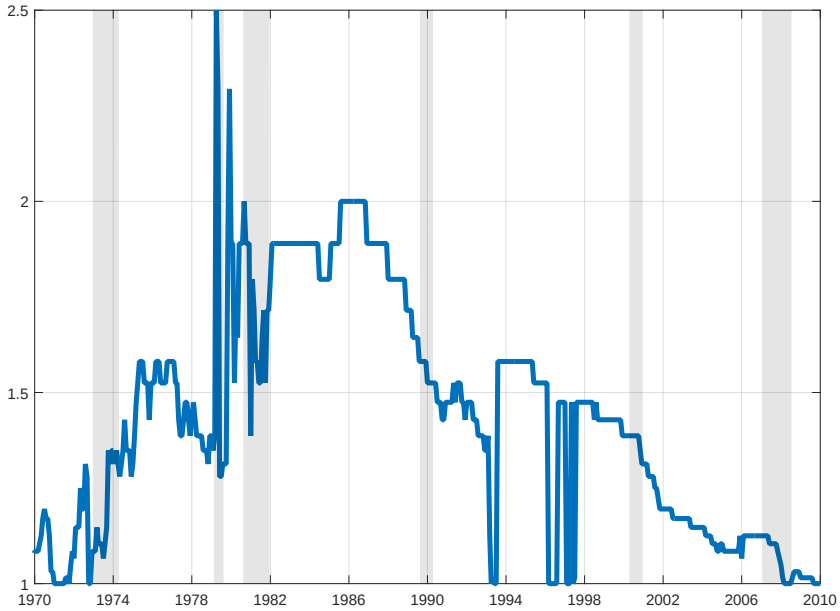


Figure 1: Optimal λ - 20 variables TVP-VAR. The figure shows the evolution of the value of λ optimized using the L_{fit} criterion in the TVP-VAR with 20 variables. Shaded areas indicate NBER-dated recessions.

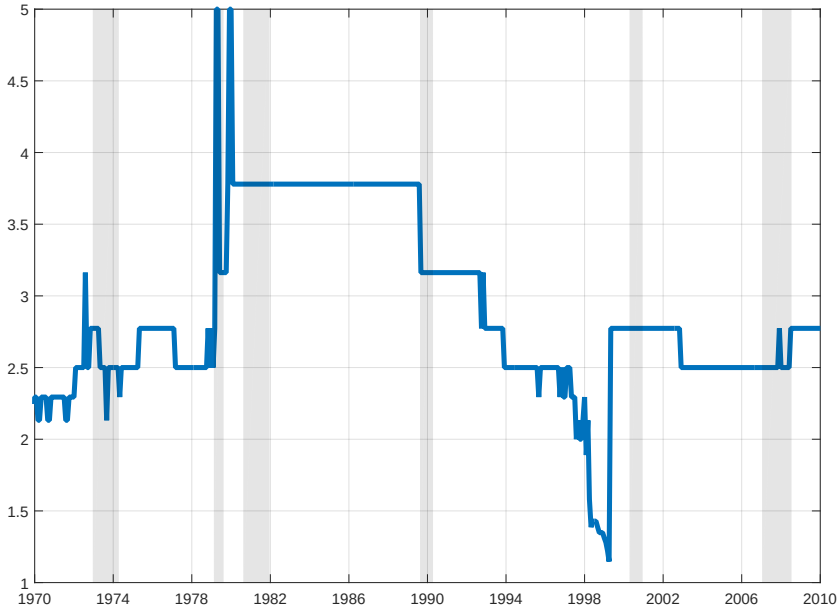


Figure 2: Optimal λ - 78 variables TVP-VAR. The figure shows the evolution of the value of λ optimized using the L_{fit} criterion in the TVP-VAR with 78 variables. Shaded areas indicate NBER-dated recessions.

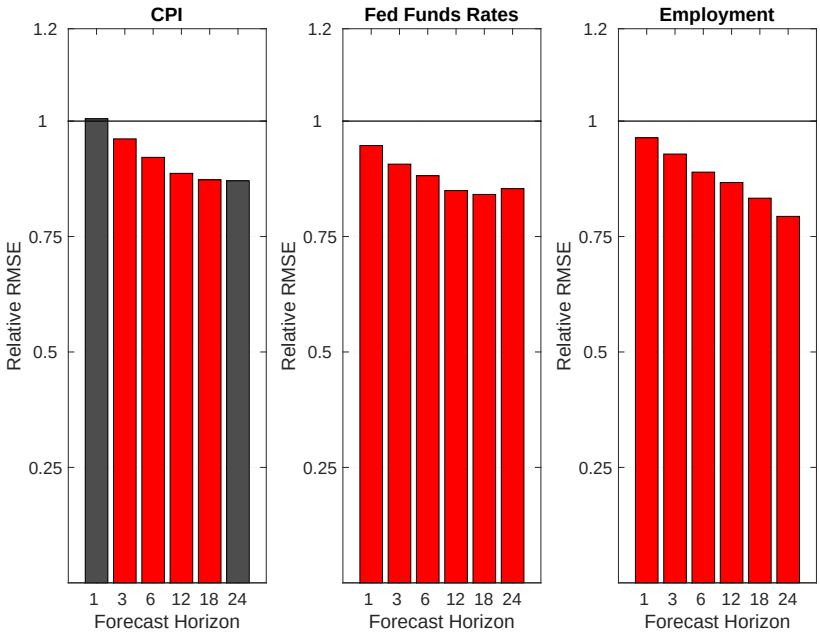


Figure 3: Root Mean Square Forecast Errors: combined TVP-VARs versus constant coefficients BVAR (20 variables VARs). The bar plots show the RMSE obtained by equal weights forecast combination of models with time varying parameters relative to that obtained with the BVAR with constant coefficients. Values below 1 imply that the TVP model outperforms the benchmark. Bars in grey indicate the forecast horizons for which a Diebold-Mariano test does not reject the null hypothesis of equal forecast accuracy, those in red denote the cases for which forecast accuracy is significantly different at the 10% confidence level.

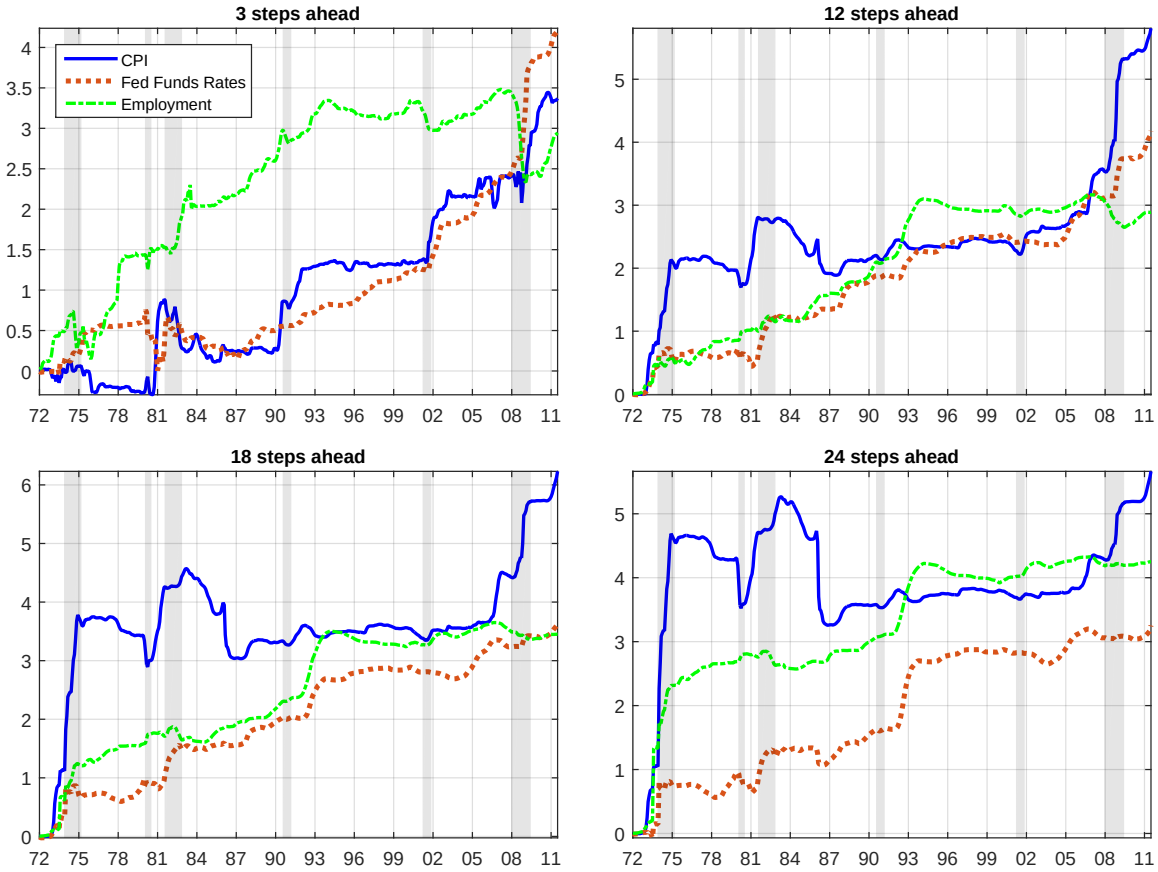


Figure 4: Cumulative sum of squared forecast error differentials: combined TVP-VARs versus constant coefficients BVAR (20 variables VARs). The figure shows the Cumulative Sum of Squared Forecast Errors Differentials between the equal weights forecast combination of models with time varying parameters and the BVAR with constant coefficients. Positive and increasing values indicate that the TVP model outperforms the benchmark, while negative and decreasing values suggest the opposite. Shaded areas indicated NBER-dated recessions.

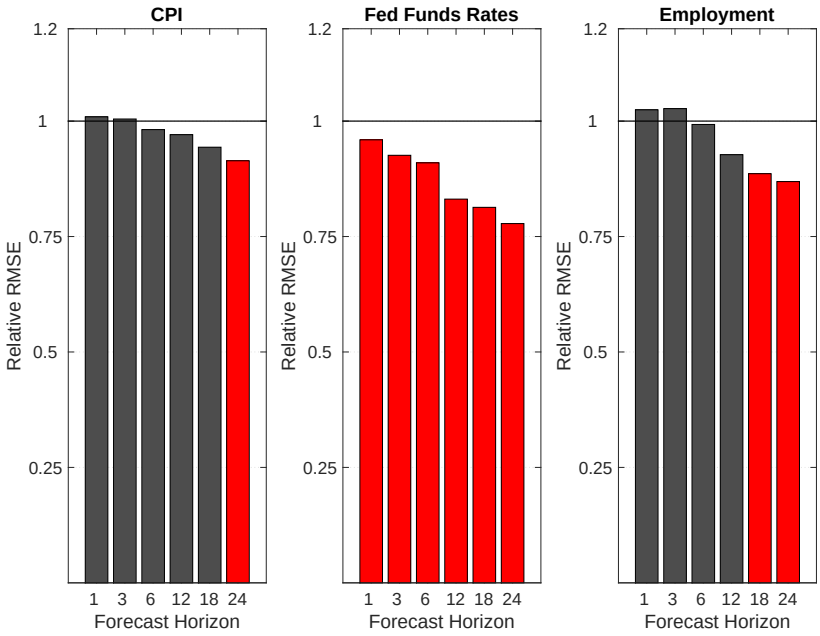


Figure 5: Root Mean Square Forecast Errors: 20 variables combined TVP-VARs versus 78 variables constant coefficients BVAR. The bar plots show the RMSE obtained by equal weights forecast combination of 20 variables VARs with time varying parameters relative to that obtained with a 78 variables BVAR with constant coefficients. Values below 1 imply that the TVP model outperforms the benchmark. Bars in grey indicate the forecast horizons for which a Diebold-Mariano test does not reject the null hypothesis of equal forecast accuracy, those in red denote the cases for which forecast accuracy is significantly different at the 10% confidence level.

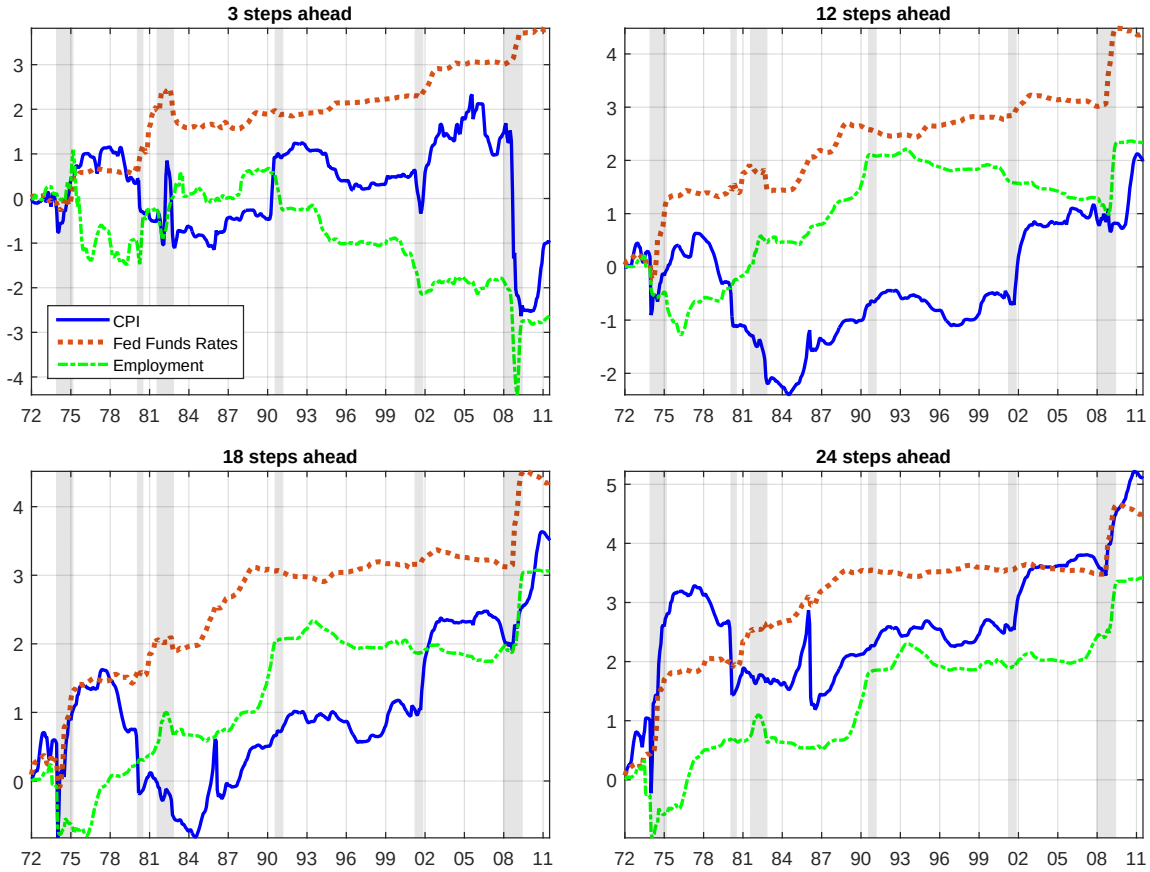


Figure 6: Cumulative sum of squared forecast error differentials: 20 variables combined TVP-VARs versus 78 variables constant coefficients BVAR. The figure shows the Cumulative Sum of Squared Forecast Errors Differentials between the equal weights forecast combination of 20 variables VARs with time varying parameters and a 78 variables BVAR with constant coefficients. Positive and increasing values indicate that the TVP model outperforms the benchmark, while negative and decreasing values suggest the opposite. Shaded areas indicated NBER-dated recessions.

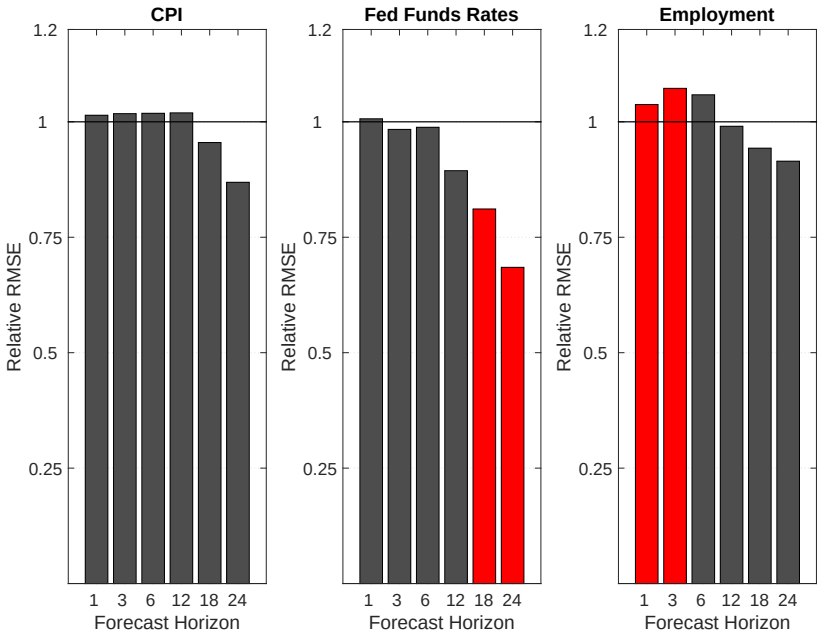


Figure 7: Root Mean Square Forecast Errors: 20 variables combined TVP-VARs versus 78 variables combined TVP-VARs. The bar plots show the RMSE obtained by equal weights forecast combination of 20 variables VARs with time varying parameters relative to that obtained by equal weights forecast combination of 78 variables VARs with time varying parameters. Values below 1 imply that the TVP model with 20 variables outperforms the benchmark. Bars in grey indicate the forecast horizons for which a Diebold-Mariano test does not reject the null hypothesis of equal forecast accuracy, those in red denote the cases for which forecast accuracy is significantly different at the 10% confidence level.

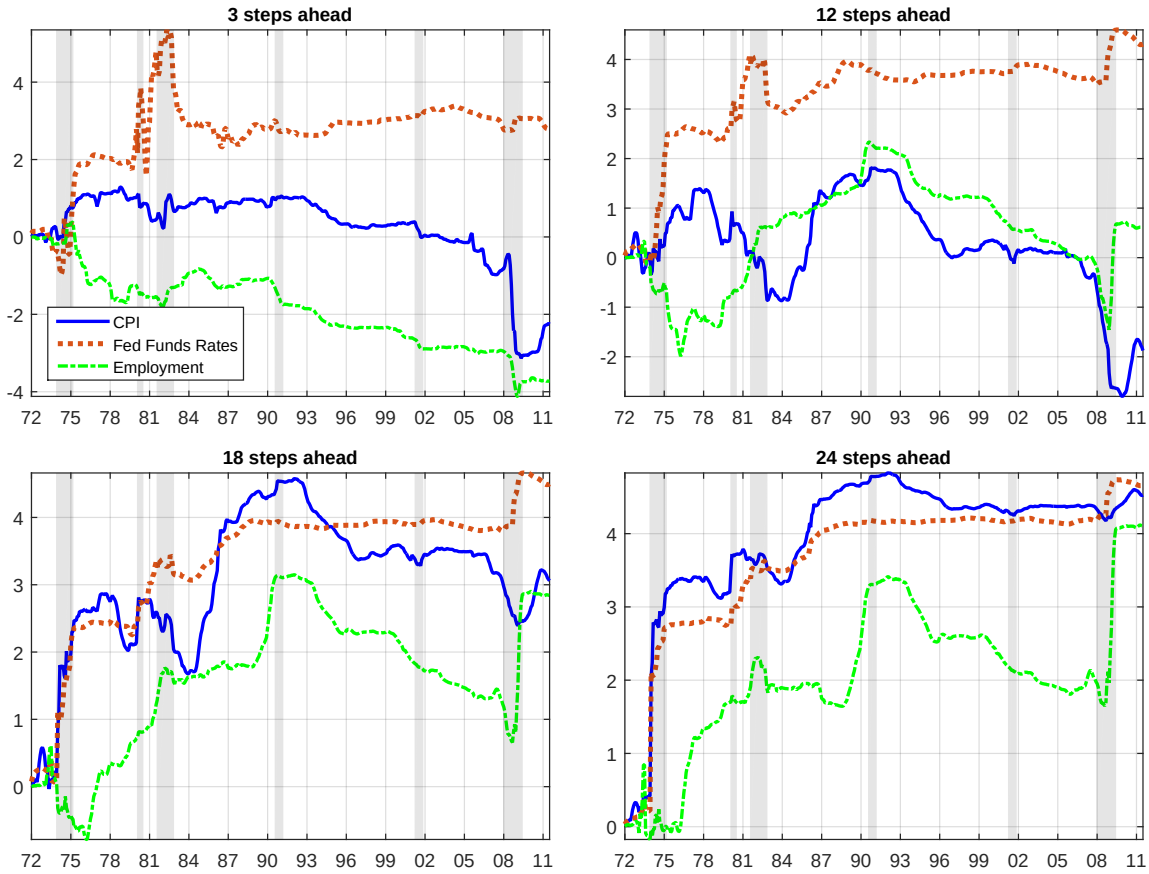


Figure 8: Cumulative sum of squared forecast error differentials: 20 variables combined TVP-VARs versus 78 variables combined TVP-VARs. The figure shows the Cumulative Sum of Squared Forecast Errors Differentials between the equal weights forecast combination of 20 variables VARs with time varying parameters and equal weights forecast combination of 78 variables VARs with time varying parameters. Positive and increasing values indicate that the TVP model with 20 variables outperforms the benchmark, while negative and decreasing values suggest the opposite. Shaded areas indicated NBER-dated recessions.

Table 4: 1 step ahead, relative RMSEs

T	Parametric	Non parametric	
		Inv. RMSE	Equal weights
DGP-1 (Random walk coefficients)			
		n=7	
100	1	1.004	1.005
150	1	0.999	1.000
200	1	0.997	0.997
		n=15	
100	1	1.021	1.024
150	1	1.012	1.013
200	1	1.006	1.007
DGP-2 (Occasionally breaking coefficients)			
		n=7	
100	1	0.96	0.96
150	1	0.96	0.96
200	1	0.96	0.96
		n=15	
100	1	0.96	0.96
150	1	0.95	0.95
200	1	0.94	0.94
DGP-3 (Sine function coefficients)			
		n=7	
100	1	0.95	0.96
150	1	0.96	0.98
200	1	0.97	0.99
		n=15	
100	1	0.87	0.88
150	1	0.86	0.87
200	1	0.85	0.86

Note to Table 4. The table shows the ratio between the one step ahead RMSE attained by, respectively, the nonparametric and the parametric model, averaged across the n variables. Forecasts are computed on the second half of the sample, i.e. when T=100, forecasts are computed recursively for t=51 to t=100, when T=150 forecasts are computed recursively for t=76 to t=150 and when T=200 forecasts are computed recursively for t=101 to t=200.

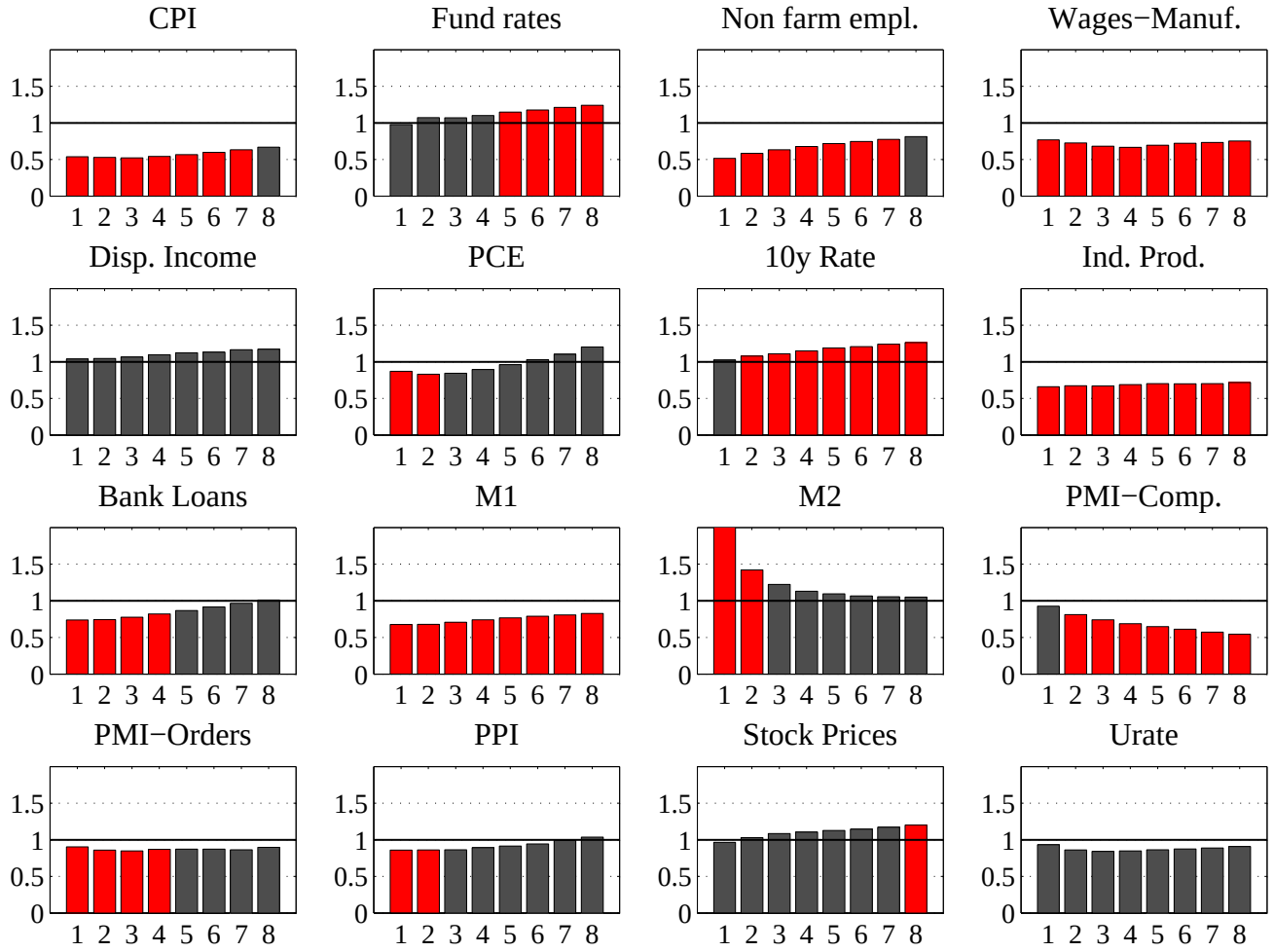


Figure 9: Forecast accuracy, nonparametric and parametric estimators

Note to Figure 9. The bar plots show the ratio between the RMSE attained by, respectively, the nonparametric and the parametric model. Values below 1 imply that the nonparametric model outperforms the parametric one. Bars in grey indicate that the Diebold-Mariano test does not reject the null hypothesis of equal forecast accuracy, while those in red denote the cases for which forecast accuracy is significantly different at the 10% confidence level.

Market group	Acronym	Weight
Industrial Production Index	INDPRO	100
Industrial Production: Business Equipment	IPBUSEQ	9.18
Industrial Production: Consumer Goods	IPCONGD	27.2
Industrial Production: Durable Consumer Goods	*IPDCONGD*	*5.59*
Industrial Production: Nondurable Consumer Goods	*IPNCONGD*	*21.62*
Industrial Production: Final Products (Market Group)	IPFINAL	16.58
Industrial Production: Materials	IPMAT	47.03
Industrial Production: Durable Materials	*IPDMAT*	*17.34*
Industrial Production: Nondurable Materials	*IPNMAT*	*11.44*

Table 5: Industrial production indexes by market group

Note to Table 5. The shares of market groups refer to 2011 Value added in nominal terms. Nondurable consumer goods includes Consumer Energy products, which account for 5.7% of total IP. We have excluded from the analysis Industrial Production of Energy Materials, which is part of the Materials (IPMAT) group and accounts for 18.3% of overall output. Source http://www.federalreserve.gov/releases/g17/g17tab1.txt

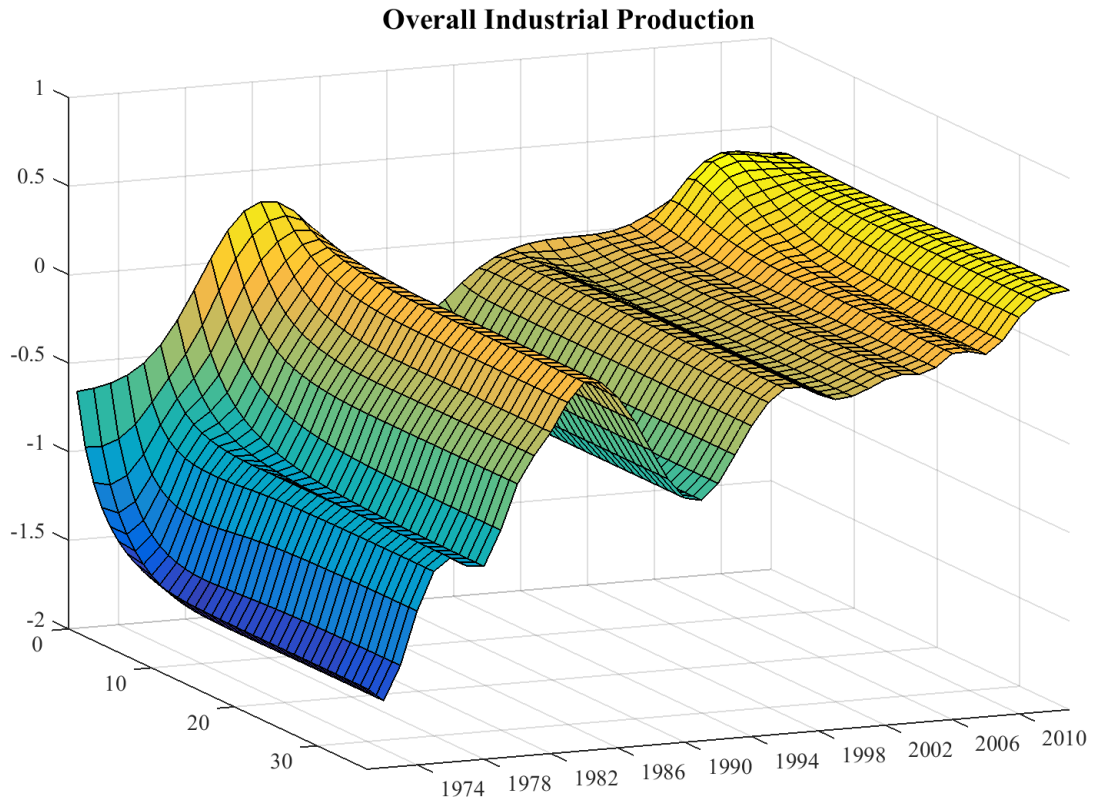


Figure 10: Response of Industrial production (overall index) to a 1% shock to the real price of oil

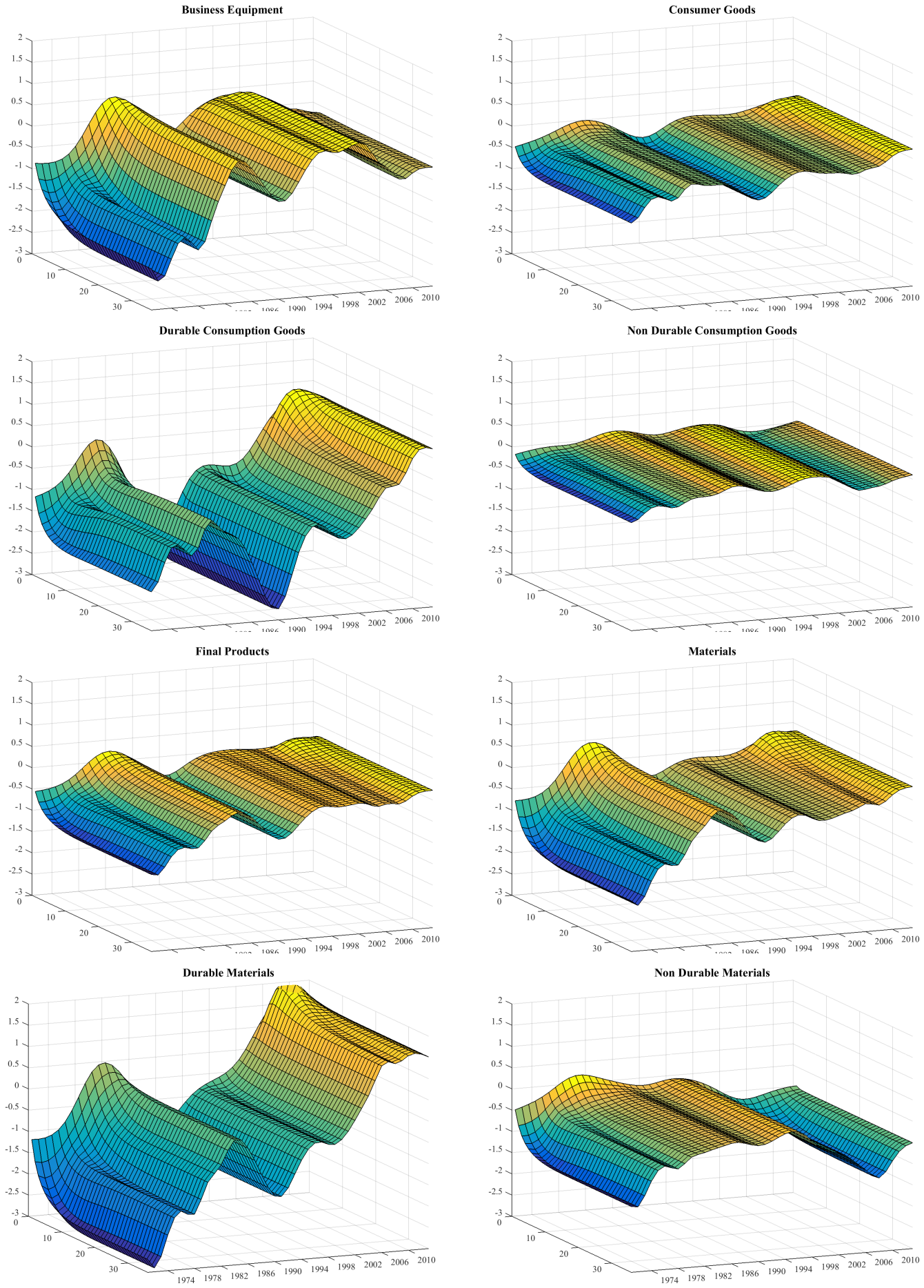


Figure 11: Response of Industrial production (sectors) to a 1% shock to the real price of oil

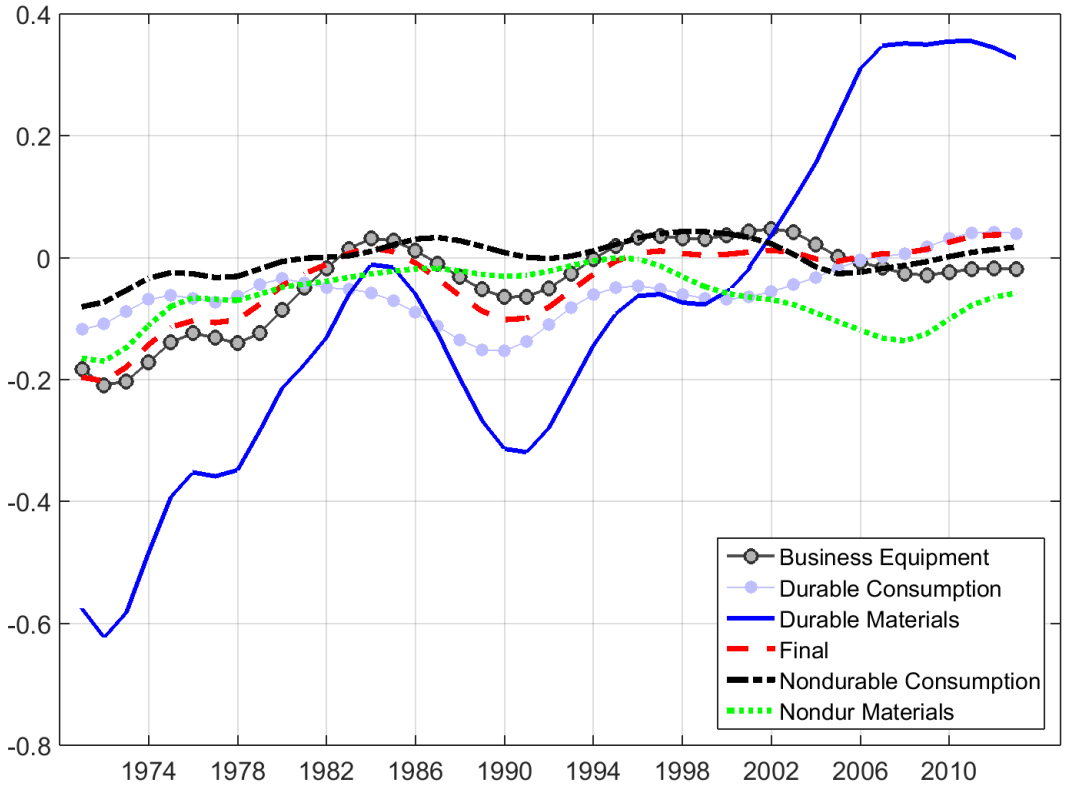


Figure 12: Contribution of selected sectors to the response of overall Industrial production (12 months out) to a 1% shock to the real price of oil

A Comparison with the parametric estimator: technical details

A.1 Monte Carlo exercise

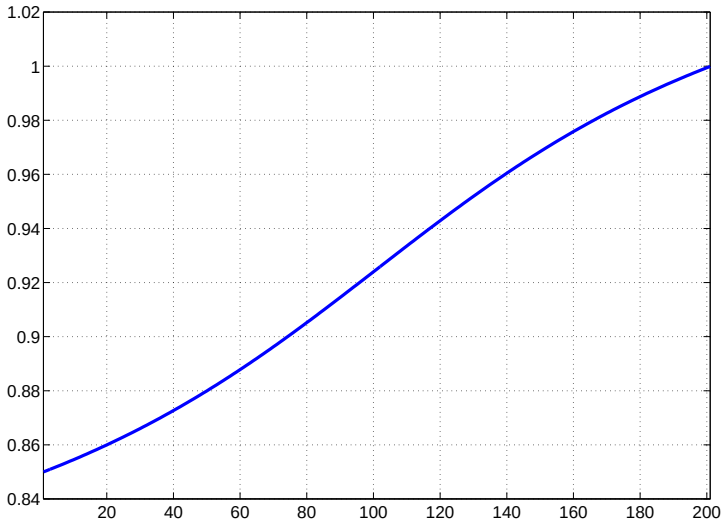
In the Monte Carlo exercise the coefficients Λ_t are obtained through the following algorithm.

1. First, simulate n^2+n coefficients according to the chosen DGP, where n is the size of the VAR.
2. The coefficients are then bounded by the largest one, so as to be lower or equal to 1 in absolute value.
3. At each t the first n^2 coefficients are orthogonalized through the Grahm-Shmidt procedure and used to form an $n \times n$ orthonormal matrix. Call this matrix P_t.
4. The last n coefficients are used to form a diagonal matrix of eigenvalues L_t. At each point in time the elements of L_t (call them l_t) are transformed using the following function

$$\widetilde{l}_t = 0.5(1+\theta_L+eps)+0.5(1-\theta_L-eps)\arctan(l_t)/\arctan(1)-eps$$

   where l_t are the input eigenvalues (that by construction lie between -1 and 1), θ_L is the desired lower bound and eps is a small constant that ensures that the upper bound is $1-eps$, that is strictly below 1. The resulting function is relatively smooth, as it can be seen in Figure A.13 where $\widetilde{l}_t$ is plotted against the possible values of l_t (on the x axis there are 201 equally spaced points between -1 and 1) and $\theta_L = 0.85$.
5. Construct $\Lambda_t = P_t \widetilde{L}_t P_t'$

Figure A.13: Constrained simulated coefficients



A.2 Dynamic model selection

The estimation method of the parametric model suggested by Koop and Korobilis depends on a number of constants, the so called forgetting factor, θ, the prior tightness for the initial conditions λ and the constant κ. To select θ and λ we follow their dynamic model selection (DMS) algorithm. First the forgetting factor θ is made time-varying as follows:

$$\theta_t = \theta_{\min} + (1 - \theta_{\min}) L^{f_t} \tag{43}$$

where $f_t = -NINT(\widehat{\varepsilon}'_{t-1}\widehat{\varepsilon}_{t-1})$, $NINT$ is the rounding to the nearest integer function, $\widehat{\varepsilon}_{t-1}$ are the one step ahead forecast errors, $\theta = 0.96$, $L = 1.1$ (values calibrated to obtain a forgetting factor between 0.96 and 1). As for the prior tightness we use a grid of J values. Each point in this grid defines a new model. Weights for each model j (defined $\pi_{t/t-1,j}$) are obtained by Koop and Korobilis as a function of the predictive density at time $t-1$ through the following recursions:

$$\pi_{t/t-1,j} = \frac{\pi^{\alpha}_{t-1/t-1,j}}{\sum_{l=1}^{J} \pi^{\alpha}_{t-1/t-1,l}} \tag{44}$$

$$\pi_{t/t,j} = \frac{\pi_{t/t-1,j} p_j(y_t|y^{t-1})}{\sum_{l=1}^{J} \pi_{t/t-1,l} p_l(y_t|y^{t-1})} \tag{45}$$

where $p_j(y_t|y^{t-1})$ is the predictive likelihood. Since this is a function of the prediction errors and of the prediction errors variance, which are part of the output of the Kalman filter, the model weights can be computed at no cost along with the model parameters estimation. Note that here a new forgetting factor appears, α, which discounts past predictive likelihoods and is set to 0.99. The constant κ is set to 0.96 throughout the exercise. At each point in time, forecast are obtained on the basis of the model with the highest weight $\pi_{t/t-1,j}$.

B The role of time-varying volatilities for forecasting

To check whether time-varying volatilities are important for the forecasting performance of our model we have run a forecast competition between the estimator with and without the correction for heteroskedasticity shown in equation (30). Since the estimation of the model in (30) can not proceed equation by equation, we run the exercise for a quarterly VAR with 20 variables and 4 lags as in section 5.5 and we experiment with two possible values for H (0.96 and 0.98) and a grid of values for the reciprocal of the penalty parameter: $\varphi = 1/\lambda$.

The results of the exercise are reported in tables B.1 and B.2. Values of the RMSEs lower than 1 (highlighted in bold) indicate that the estimator with constant volatilities outperforms the one with time varying volatility. We underlie the cases in which differences in forecast accuracy are significant at the 10% confidence level, according to the Diebold Mariano test. The results turn out to be sharply in favour of the specification with constant covariances.

For low values of the shrinkage parameters the performance of the homoskedastic VAR is either in line or slightly worse than the one obtained with time-varying volatilities. When the constraints are slightly relaxed, however, both the absolute and relative performance of the baseline specification improve significantly. For values of φ higher than 10^{-2} the gains in predictive accuracy yielded by the homoschedastic VAR increase steadily with the forecast horizon and become as high as 20% at longer horizons.

	Relative RMSE								Absolute RMSE
Horizon	1	2	3	4	5	6	7	8	
				$\lambda = 10^{-10}$					
CPI	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.99**	**0.99**	0.054
FFRATE	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.99**	0.209
EMPL	1.00	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	0.240
				$\lambda = 10^{-4}$					
CPI	1.29	1.23	1.21	1.18	1.16	1.14	1.11	1.09	0.054
FFRATE	1.05	1.00	1.01	1.01	1.02	1.02	1.02	1.02	0.210
EMPL	1.42	1.27	1.20	1.16	1.13	1.12	1.11	1.10	0.241
				$\lambda = 10^{-2}$					
CPI	1.06	1.07	1.15	1.16	1.14	1.14	1.12	1.08	0.046
FFRATE	**0.87**	**0.87**	**0.93**	**0.90**	**0.90**	**0.94**	**0.94**	**0.93**	0.203
EMPL	1.06	1.05	1.05	1.05	1.06	1.07	1.08	1.08	0.208
				$\lambda = 2 \times 10^{-2}$					
CPI	**0.99**	**0.99**	1.05	1.06	1.04	1.04	1.02	**1.00**	0.044
FFRATE	**0.85**	**0.85**	**0.90**	**0.86**	**0.86**	**0.89**	**0.89**	**0.89**	0.201
EMPL	**0.96**	**0.96**	**0.97**	**0.98**	**0.99**	1.01	1.02	1.03	0.198
				$\lambda = 3 \times 10^{-2}$					
CPI	**0.96**	**0.96**	1.00	1.00	**0.99**	**0.98**	**0.97**	**0.95**	0.042
FFRATE	**0.83**	**0.84**	**0.88**	**0.84**	**0.83**	**0.86**	**0.87**	**0.86**	0.201
EMPL	**0.91**	**0.92**	**0.93**	**0.94**	**0.95**	**0.97**	**0.99**	**1.00**	0.193
				$\lambda = 4 \times 10^{-2}$					
CPI	**0.95**	**0.94**	**0.97**	**0.97**	**0.95**	**0.94**	**0.93**	**0.91**	0.041
FFRATE	**0.83**	**0.84**	**0.87**	**0.83**	**0.82**	**0.85**	**0.85**	**0.84**	0.201
EMPL	**0.89**	**0.90**	**0.91**	**0.91**	**0.93**	**0.95**	**0.97**	**0.98**	0.190
				$\lambda = 5 \times 10^{-2}$					
CPI	**0.94**	**0.93**	**0.96**	**0.95**	**0.93**	**0.92**	**0.91**	**0.89**	0.041
FFRATE	**0.83**	**0.84**	**0.86**	**0.82**	**0.81**	**0.83**	**0.83**	**0.82**	0.201
EMPL	**0.88**	**0.89**	**0.89**	**0.90**	**0.91**	**0.93**	**0.96**	**0.97**	0.188

Table B.1: Relative RMSE - Constant versus time-varying covariances - $H = 0.96$

Note to Table B.1. Each cell under the *Relative RMSE* heading reports the ratio between the RMSE obtained assuming a constant covariance matrix of the VAR errors and that obtained using a time varying covariance matrix. In bold we highlight numbers below 1, indicating that constant-covariance approach provides more accurate forecasts. We underline the cases in which a Diebold Mariano test rejects the null hypothesis of equal forecast accuracy at the 10% confidence level. In the *Absolute RMSE* column we report the average RMSE (scaled by the variance of the target variable) over the 8 forecast horizons obtained with the constant-covariance method. For both methods we use exponentially weighted moving average kernels with a discount factor $H = 0.96$. The out of sample period runs from the first quarter of 1970 to the second quarter of 2013 (167 data points).

	Relative RMSE								Absolute RMSE
Horizon	1	2	3	4	5	6	7	8	
				$\lambda = 10^{-10}$					
CPI	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.997**	0.059
FFRATE	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.998**	0.203
EMPL	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.998**	0.241
				$\lambda = 10^{-4}$					
CPI	1.43	1.38	1.37	1.34	1.31	1.28	1.25	1.22	0.059
FFRATE	1.06	1.00	1.02	1.02	1.02	1.02	1.01	1.01	0.203
EMPL	1.59	1.40	1.30	1.23	1.19	1.17	1.15	1.13	0.240
				$\lambda = 10^{-2}$					
CPI	1.01	1.04	1.10	1.11	1.09	1.09	1.07	1.05	0.046
FFRATE	**0.85**	**0.84**	**0.87**	**0.84**	**0.83**	**0.86**	**0.87**	**0.86**	0.196
EMPL	**0.98**	**0.98**	**0.98**	**0.99**	**0.99**	1.00	1.01	1.01	0.199
				$\lambda = 2 \times 10^{-2}$					
CPI	**0.96**	**0.97**	**1.00**	**0.99**	**0.97**	**0.96**	**0.95**	**0.94**	0.043
FFRATE	**0.82**	**0.82**	**0.84**	**0.80**	**0.79**	**0.82**	**0.82**	**0.81**	0.196
EMPL	**0.90**	**0.90**	**0.90**	**0.91**	**0.92**	**0.94**	**0.96**	**0.97**	0.191
				$\lambda = 3 \times 10^{-2}$					
CPI	**0.94**	**0.94**	**0.96**	**0.94**	**0.92**	**0.90**	**0.89**	**0.88**	0.041
FFRATE	**0.81**	**0.81**	**0.82**	**0.78**	**0.77**	**0.79**	**0.80**	**0.79**	0.197
EMPL	**0.86**	**0.86**	**0.86**	**0.87**	**0.88**	**0.91**	**0.93**	**0.94**	0.188
				$\lambda = 4 \times 10^{-2}$					
CPI	**0.93**	**0.93**	**0.93**	**0.91**	**0.89**	**0.87**	**0.86**	**0.84**	0.041
FFRATE	**0.80**	**0.81**	**0.81**	**0.76**	**0.76**	**0.78**	**0.78**	**0.77**	0.198
EMPL	**0.84**	**0.84**	**0.84**	**0.85**	**0.86**	**0.89**	**0.91**	**0.93**	0.186
				$\lambda = 5 \times 10^{-2}$					
CPI	**0.93**	**0.92**	**0.92**	**0.89**	**0.87**	**0.85**	**0.84**	**0.82**	0.040
FFRATE	**0.80**	**0.80**	**0.80**	**0.75**	**0.75**	**0.77**	**0.77**	**0.76**	0.199
EMPL	**0.82**	**0.83**	**0.83**	**0.83**	**0.84**	**0.87**	**0.90**	**0.91**	0.186

Table B.2: Relative RMSE - Constant versus time-varying covariances - $H = 0.98$

Note to Table B.2. Each cell under the *Relative RMSE* heading reports the ratio between the RMSE obtained assuming a constant covariance matrix of the VAR errors and that obtained using a time varying covariance matrix. In bold we highlight numbers below 1, indicating that constant-covariance approach provides more accurate forecasts. We underline the cases in which a Diebold Mariano test rejects the null hypothesis of equal forecast accuracy at the 10% confidence level. In the *Absolute RMSE* column we report the average RMSE (scaled by the variance of the target variable) over the 8 forecast horizons obtained with the constant-covariance method. For both methods we use exponentially weighted moving average kernels with a discount factor $H = 0.98$. The out of sample period runs from the first quarter of 1970 to the second quarter of 2013 (167 data points).