ESSAYS IN NONLINEAR ECONOMETRICS

Jacob Warren

A DISSERTATION

in

Economics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2017

Supervisor of Dissertation

Francis X. Diebold
Professor of Economics

Co-Supervisor of Dissertation

Frank Schorfheide
Professor of Economics

Graduate Group Chairperson

Jesús Fernández-Villaverde
Professor of Economics

Dissertation Committee

Francis X. Diebold, Professor of Economics
Frank Schorfheide, Professor of Economics
Xu Cheng, Associate Professor of Economics

ESSAYS IN NONLINEAR ECONOMETRICS

ACKNOWLEDGEMENT

Completing a PhD is a long and arduous task, and there are many people I would like to thank, without whom this dissertation would not have been possible. First of all, I am indebted to both my advisors, Francis X. Diebold and Frank Schorfheide, and my third committee member, Xu Cheng. Their guidance began in the first year econometrics sequence where they inspired a love of time-series and Bayesian econometrics, both of which were integral to my research. Beyond coursework, they have helped develop and nurture both my interest in econometrics in general and my particular projects by constantly providing advice and time. They are both outstanding economists and statisticians, and I have benefited greatly from working with them.

I would also like to thank my other econometrics instructor Francis J. DiTraglia. His coursework on factor models and model selection was immensely interesting and helpful, and he was always available to provide insight into potential research topics.

I am very grateful to all the presenters and participants at the Econometrics Seminar and the Econometrics Lunch throughout my time here, specifically Pooyan Amir-Ahmadi. His presentation and research created a natural springboard for thinking about my TVP chapter, and he played an integral role in setting an early path for the project. I also greatly benefitted from all the participants and visiting professors who made numerous helpful comments on all my research projects.

Next, I must express my gratitude to the econometrics students, both past and present. The students in the cohorts ahead of me, Minchul Shin, Lorenzo Braccini, Molin Zhong and Laura Liu, showed me how to do research, and how to deal with the failures of research. Whenever there was a piece of bad news, they were there putting a positive spin on the situation from their experience. The students in cohorts below me, Minsu Chang and Paul Sangrey, consistently pushed me to develop ideas and think about things from alternative perspectives. Thanks also to Ross Askanazi, without whom I would not have completed

this dissertation. Between in-depth discussions of research ideas, constant brainstorming sessions on new ideas, and co-authored papers, he has been invaluable. Lastly, I am grateful to my office mates, Ross, Paul and Hanna Wang. They are all great economic minds, and made every day enjoyable.

I am also much obliged to all my friends outside of school, especially my Philly Phamily. When things were difficult, they provided a welcome respite from work.

Lastly, I am thankful to my family, both my parents and in-laws. Their love and support was integral through every step of this process. And most importantly, I would like to express my gratitude towards my wife Dana. Her many sacrifices and constant encouragement enabled me to pursue this degree. I truly could not have asked for a better partner in life.

ABSTRACT

ESSAYS IN NONLINEAR ECONOMETRICS

Jacob Warren

Francis X. Diebold

Frank Schorfheide

In this dissertation, I study standard models, but investigate the necessity of (possibly large) deviations from basic assumptions. In Chapter 1, my co-author Ross Askanazi and I revisit the use of factor models in finance. Historical literature on the subject decomposes volatility into a factor component (systemic risk) and a remainder (idiosyncratic risk). Recent work has suggested that a market shock to volatility may increase both systemic risk *and* idiosyncratic risk — specifically, that idiosyncratic volatility of US equities data has a factor structure, with the factor highly correlated with, and possibly precisely the market volatility. In this paper we attempt to characterize the underlying factor and find that it can be decomposed into a statistical (PCA) and structural (market volatility) factor. We also show that this feature is more common than expected, appearing in diverse sets of financial data. Lastly, we find that this dual-factor approach is slightly dominated in forecasting environments by a single statistical factor. In Chapter 2 I revisit the classical Vector Autoregression (VAR) model, but allow parameters to time-vary. Time-Varying parameter models have become more popular in recent years, especially as they are adapted to accommodate larger datasets. However, all recent developments use standard priors, specifically the Inverse-Wishart class of priors over the parameter error covariance matrix. In this paper, I show that Inverse-Wishart priors have a number of negative properties, and that those properties are salient in a TVP context since there is little information from the likelihood. Fully aware of these deficiencies, the Bayesian Random Effects literature has developed a series of uninformative priors to correct these weaknesses. In this paper, I adapt one of those priors into an *informative* and easily understandable prior for covariances. I show that the new

prior effects posterior inference and displays improved frequentist properties. I apply my prior to the canonical Primiceri (2005) dataset and find that their results were sensitive to the choice of prior. I further apply the prior to two forecasting exercises and find that while it improves forecasts for the Primiceri data, it does not for an alternative (larger) dataset.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1

Factor Analysis For Volatility [1]

1.1 Introduction

As economists we find that large complex dynamics can usually be modeled as resulting from a small number of fundamental shocks. Factor models approach this formally:

$$y_t = \beta F_t + e_t, \quad E(F_t e_t) = 0,$$

where $dim(F_t) = k << dim(y_t) = N$ and $t = 1, \dots T$. One popular application of factor models (especially in finance) is for covariance matrix estimation. The factor model presents a useful decomposition, assuming factors and errors are orthogonal:

$$\Sigma^y = \beta \Sigma^F \beta' + \Sigma^e.$$

Here Σ^e is sparse, if not diagonal, and Σ^F is of small dimension, so $\beta \Sigma^F \beta'$ is of low rank. This "low rank plus sparse" decomposition via factor models has facilitated tractable dynamic volatility: For Σ^y to be time-varying, at least one of β, Σ^F, or Σ^e must be time-varying.

Over the years, there have been many variations to induce time-varying volatility in Σ^y. Most commonly (Diebold and Nerlove (1989), Jacquier et al. (1994)), Σ^F is endowed with stochastic volatility, while other elements remain constant. More recently though (Kim et al.

[1] This chapter is co-authored with Ross Askanazi

(1998), Pitt and Shephard (1999), Aguilar and West (2000)), the diagonal elements of Σ^e were also allowed to time-vary, adding an additional layer of complexity.

However, recent empirical work has indicated that despite the factor model inducing orthogonal structure on the level equation, it ignores higher order dependence between the factor and idiosyncratic component. Specifically, Herskovic et al. (2014) find that idiosyncratic variances tend to (strongly) comove, and Barigozzi and Hallin (2014) further show that the comovement extends to the volatility of the level factor (Σ_t^F) as well. Kalnina and Tewou (2015) and Duarte et al. (2014) are in the same vein. More specifically, let $\boldsymbol{\sigma_t^e} = diag(\Sigma_t^e)$, then those papers suggest:

$$\log(\boldsymbol{\sigma_t^e}) = AV_t + \varepsilon_t, \quad E(V_t\varepsilon_t) = 0, \quad dim(V_t) << N,$$

where V_t is a factor for idiosyncratic volatility.

Our paper immediately builds off those recent contributions by using high-frequency based Realized Volatilties on two datasets of US Equities. In general, our findings support prior research: the panel of idiosyncratic volatilities has clear and strong factor structure, and the first principal component of the panel is highly correlated with market volatility.

The above literature is split on the nature of the factor for idiosyncratic volatility. While all agree that idiosyncratic volatility is dynamic and has factor structure, there is no consensus as to what precisely is the factor. Some use the market volatility as the factor, while others take a more statistical approach and merely use the first principal component. We attempt to provide clarity on that issue by accomplishing three main goals: First, we provide a framework for estimating the factor structure in idiosyncratic volatility using realized measures. Second, we attempt to answer (via a series of graphical tools and statistical tests) how exactly the factor for idiosyncratic volatility is related to market volatility. More specifically, we are interested in whether they are precisely the same, or if one supersedes the other. Third, we demonstrate that the structure is a general feature of volatility, and

not just limited to equities.

To accomplish the third goal, we extend this work to a panel of exchange rate volatilities in addition to the equities datasets. The same tractable dynamic volatility modeling has been used in forecasting exchange rate volatility (Diebold and Nerlove (1989)), and we explore the same questions of the nature of exchange rate idiosyncratic volatility. In contrast to equities, the correlation between the factor for idiosyncratic volatility and market volatility falls dramatically.

Despite that large difference, all datasets support the same general framework — namely that both the market volatility *and* an additional principal components factor is necessary for explaining cross-sectional variation. While on the one hand this presents a robust statistical fact, it is also troubling from an economic modeling perspective. Indeed, the question of *why* these statistical facts occur become all the more pronounced. Is there an economic theory that can support the phenomenon for both FX returns and equities? Or perhaps, is the framework a product of network effects, time-varying volatilities and financial markets? While we do not attempt to answer these questions in this paper, they provide the foundation for this and future work in the area.

The outline for the remainder of the paper is as follows: In Section 1.2, we outline the framework for estimating dynamic idiosyncratic volatility. In Section 1.3, we present the US equities data, and in subsections explore the outcomes of our model selection framework. In Section 1.4, we conduct the same set of exercises for foreign exchange rates. Section 1.5 explores robustness to the most obvious counterpoint to the proposed framework — namely that features of idiosyncratic volatility can simply be the result of conditional mean mis-specification. Finally, Section 1.6 explores the implications of our findings for out-of-sample forecasting and Section 1.7 concludes. Post-conclusion, we provide simulation evidence that our battery of statistical tests perform and behave appropriately in our environment. This can be found in Section 1.8.1.

1.2 Modeling Procedure

1.2.1 Continuous Time Setup

For equities, we start with a continuous time price process that mimics the setup in Barndorff-Nielsen and Shephard (2004). Let $S(t)$ be the price process of a security (or possibly a vector of securities), and $X(t) = \log(S(t))$ be a semi-martingale, so

$$X(t) = \alpha(t) + m(t),$$

where $\alpha(t)$ is the drift term and $m(t)$ is a local martingale. For any sequence of partitions, $t_0 = 0 < t_1 < t_2 \cdots < t_M = t$, with $sup_j\{t_{j+1} - t_j\} \to 0$ for $M \to \infty$, we define the quadratic variation on day t as:

$$[X](t) = plim_{M\to\infty} \sum_{j=0}^{M-1} \{X(t_{j+1}) - X(t_j)\}\{X(t_{j+1}) - X(t_j)\}'.$$

In practice we only have a finite partition, so we construct the Realized Volatility as an estimator of the quadratic variation:

$$\widehat{[X](t)} = RV_t = \sum_{j=0}^{M-1} \{X(t_{j+1}) - X(t_j)\}\{X(t_{j+1}) - X(t_j)\}'.$$

This is the standard definition of Realized Volatility, which has been well described and analyzed over the recent years (see, among others, Andersen et al. (2007)).

We further utilize two information sets, as in Sheppard and Xu (2014): a high frequency information set $\mathcal{F}_t^{HF}$ and a low frequency information set $\mathcal{F}_t^{LF}$. The high frequency information set contains all the information of the low frequency information set, plus the intraday data necessary to construct the realized measure at date t (so that $\mathcal{F}_t^{LF} \subset \mathcal{F}_t^{HF}$). We will subscript the high frequency information set by t_j, $j = 1, ..., M_t$ for each date t.

Our primary objects of interest are as follows: We have returns r_t, factor loadings β_t, a level factor f_t, and idiosyncratic shocks v_t for the level equation. We posit the existence of a single factor structure at high frequency, so that the volatility of the factor is a scalar σ_t^f. The covariance of the idiosyncratic shocks is Ω_{v_t}.

$$
\begin{aligned}
r_{t_j} &= \beta_t f_{t_j} + v_{t_j} \qquad t = 1, \dots T \quad j = 1, \dots M_t, \\
f_{t_j} | \mathcal{F}_t^{HF} &\sim iidN(0, \sigma_t^f), \\
v_{t_j} | \mathcal{F}_t^{HF} &\sim iidN(0, \Omega_{v_t}).
\end{aligned}
$$

Since the market factor and idiosyncratic error are continuous-time return sequences that are observed at distinct time partitions, we can compute their respective Realized Volatilities (assuming β is fixed and known intraday):

$$
\begin{aligned}
RV_{f_t} &= \sum_{j=0}^{M-1} \{f_{t_{j+1}} - f_{t_j}\}\{f_{t_{j+1}} - f_{t_j}\}', \\
RV_{v_t} &= \sum_{j=0}^{M-1} \{v_{t_{j+1}} - v_{t_j}\}\{v_{t_{j+1}} - v_{t_j}\}'.
\end{aligned}
$$

This factor structure at high frequencies time aggregates to a factor structure at the low (daily) frequency,

$$
\begin{aligned}
r_t^{LF} &= \beta_t f_t^{LF} + v_t^{LF}, \\
f_t^{LF} | \mathcal{F}_t^{LF} &\sim N(0, \sigma_t^f), \\
v_t^{LF} | \mathcal{F}_t^{LF} &\sim iidN(0, \Omega_{v_t}).
\end{aligned}
$$

From this point forward the LF superscript will be suppressed for brevity. We will at times use the notation $X_t^{HF} = [X_{t_0}, X_{t_1}, \dots X_{t_M}]'$ to represent the vector of high-frequency intraday observations of asset X.

Factor Loadings

It remains to specify dynamics on the factor loadings as well. There is considerable debate on whether factor loadings actually have time-variation, and if so, at what frequency they should vary. There is also a debate about whether this time-variation has any broader implications for risk or returns. Braun et al. (1995) use bivariate EGARCH models to measure estimate conditional covariances of returns, but find only weak evidence of time-varying conditional (monthly) betas. Using an international panel, Ferson and Harvey (1993) find that nation-specific betas do time-vary with international risk factors, but that movements in the betas contribute only a small fraction to predicted variation in expected returns. Bali and Engle (2010) find substantial time-variation in betas with the market, and Bali et al. (2013) shows that the time-variation is meaningful for trading. Supporting this, Jagannathan and Wang (1996) allow betas to time-vary in a CAPM model, which is better able to explain cross-sectional returns. Lewellen and Nagel (2006) agree that betas time-vary, but disagree about their ability to explain cross-sectional returns. Sheppard and Xu (2014) combine realized measures with GARCH dynamics (HEAVY-GARCH) on factor models (including loadings) to great success. Most applicable to our setup, Andersen et al. (2006) compute Realized Betas, and find that they have much shorter memory than Realized Volatilities.

The debate about whether (and how much) betas vary over time is specifically important to our setup. Take for example, a toy model with time-varying betas:

$$y_t = \beta_t F_t + e_t,$$

but the econometrician instead estimates a model with constant betas:

$$y_t = \beta F_t + \bar{e}_t.$$

Then observe that the error term will include the time-variation in betas:

$$\bar{e}_t = (\beta_t - \beta)F_t + e_t.$$

This has large implications for the observed idiosyncratic covariance matrix from the misspecified regression:

$$\Sigma_{\bar{e}} = (\beta_t - \beta)\Sigma_F(\beta_t - \beta)' + \Sigma_e.$$

Thus, one could observe factor structure in the residual variances (and the factor would be highly correlated with factor volatility) simply due to misspecified dynamics in the factor loadings.

We therefore allow betas to time-vary at the daily level, but leave them fixed intraday. Mimicking the approach of Andersen et al. (2006), we use a Realized Beta setup:

$$R\beta_{i,t} = \frac{Cov(r_{it}^{HF}, f_t^{HF})}{Var(f_t^{HF})}$$

We allow dynamics on the factor loadings to follow independent autoregressions:

$$\Phi_{\beta_i}(L)\beta_{i,t} = \eta_{i,t}^{\beta}, \qquad \eta_{i,t}^{\beta} \sim iidN(0, \sigma_i^{\beta}) \tag{1.2.1}$$

Dynamics on the factor loadings are given as independent univariate autoregressions, because having to estimate a vector autoregression of factor loadings defeats the purpose of employing

a factor structure in the first place, since there are N series of loadings.

It is important to note: while variation in Realized Betas has important implications for the cross-sectional and time-variation of asset Realized Volatility, modeling it greatly increases the number of parameters of the model (there are N times k times T Realized Betas). Therefore, for the purposes of *forecasting* asset Realized Volatility, it is not clear that allowing for variation in Realized Betas will improve outcomes. In fact we find that it is not — allowing for this variation increases mean squared forecast error. In our forecasting exercise, we therefore hold factor loadings constant, with the understanding that this may inflate the measured time-variation in idiosyncratic volatility. On balance, however, we find that this approach and a conservative interpretation of idiosyncratic volatility dynamics is more appropriate for forecasting.

Factor Structure and PCA

In all empirical exercises, we use an observed factor for F_t in the level equation. This allows us to both ignore estimation error in F_t, and provides us with observed high frequency data for F_t, yielding realized measures of σ_{F_t} and $\boldsymbol{\sigma_{e_t}}$.

In order to extract a statistical factor, V_t, we use principal components to extract a static factor for the idiosyncratic volatilities. Recall that for a panel $\log(\boldsymbol{\sigma_e})$, principal components extracts factors via the minimization problem

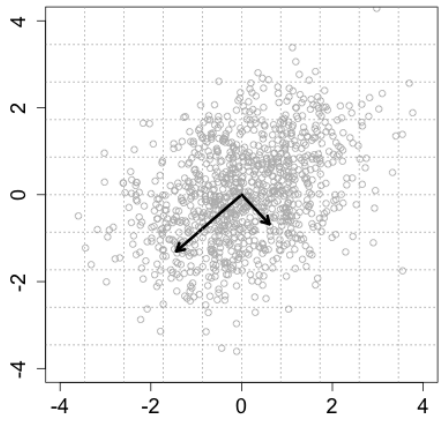
$$V(k) = \min_{\Lambda, F^k} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (\log(\sigma_{e_{i,t}}) - \lambda_i^k V_t^k)^2,$$
$$\text{subject to } \Lambda^{k'} \Lambda^k / N = I_k \text{ or } V^{k'} V^k / T = I_k.$$

Here k is the number of factors, V are the factors, and λ are the factor loadings. Since both λ and V are latent and it is their combined component that we are centrally interested

in, we can quickly see that the two are not separately identified. This is why we need the orthonormalization identifying constraint above. We can think of this minimization problem as extracting the directions of greatest variation. This can be visualized like so:

Figure 1: Principal Components Analysis

A cloud of data. The black vectors represent the directions of greatest variation extracted by PCA. The length of each vector represents the variance in that direction.



Much like standard in-sample mean squared error (MSE) analysis, we can see from the above formula that $V(k)$ is strictly decreasing in k. Therefore, optimizing $V(k)$ is a poor choice for selecting the number of factors. As with MSE, this loss function can be augmented with a penalty function for the number of factors to create a consistent information criterion. We use the Bai and Ng (2002) Information Criterion to select the number of factors, which is given by:

$$PC(k) = V(k) + kg(N, T)$$

Where N and T are the dimensions of the panel of interest, and $g(\cdot)$ need only satisfy

$$g(N,T) \xrightarrow[N,T\to\infty]{} 0$$

$$min(\sqrt{N},\sqrt{T})g(N,T) \xrightarrow[N,T\to\infty]{} \infty$$

We use PCA as one of the options for generating a factor for volatility. Similar to most equity log- realized volatilities, the extracted PCA factor is approximately Gaussian and has long-memory. Since this factor is a linear combination of log volatilities, these features are to be expected.

1.3 Equities Data

The date ranges for the data analysis runs from January 2007 to November 2014. All low frequency (daily) returns were downloaded from the Center for Research in Security Prices (CRSP), while all high frequency data was downloaded from the Ticker and Quote (TAQ) dataset. We use high frequency data to construct realized measures from intraday returns, but use the low-frequency (daily) returns to make average Realized Volatility the same as the variance of returns.

We use two datasets: for low dimensional analysis, we use the DOW 10 and the SPY (a highly liquid ETF tracking the S&P 500) as an observed market factor. All companies in the DOW 10 are observed over the entire 2007-2014 trading period.

For high dimensional analysis, we use the S&P 100. Since companies enter and leave the index over the sample period, we keep the stocks in the index as of November 2014 that are traded across the entire 7 years. That leaves us with 90 assets. As with the DOW 10, we use the SPY as an observed market factor for this dataset. Lists of the DOW 10 and the stocks used in the S&P 100 (with sector designations) are presented in the Appendix.

1.3.1 Construction of Realized Measures

Continuing the discussion above, the Quadratic Variation of a log-price process is defined as

$$QV_t = \operatorname*{plim}_{M\to\infty} \sum_{j=0}^{M-1} \{X(t_{j+1}) - X(t_j)\}\{X(t_{j+1}) - X(t_j)\}'$$

The natural estimator of true quadratic variation truncates the number of intraday observations at some finite number. This estimator was introduced by Andersen et al. (2001) and Andersen et al. (2003) and it was shown to converge to QV_t as the number of observations goes to infinity by Barndorff-Nielsen and Shephard (2004).

Unfortunately that estimator is not robust to measurement error or jumps in the price process, so many variations have been introduced in the subsequent years. In the presence of classical measurement error, the standard Realized Variance estimator is biased, and that bias depends on sample size. So as the sampling frequency increases, the estimator becomes worse and worse. To solve this issue, Ait-Sahalia et al. (2005a) propose a complex bias-corrected estimator, but also suggest that a subsampling approach can be nearly as good. Subsampling requires multiple intraday grids for the price process, where each sampling grid (say, 5 minutes) can be further subsampled at a higher frequency (say, 1 minute). Formally, let $G^{(i)}$ be the partition of intraday returns at the i$^{\text{th}}$ minute, $G^{(i)} = \{t_i, t_{i+5}, t_{i+10}, \dots t_{i+5(M-1)}$, and associated estimate of Realized Volatility: $\widehat{[X,X]}_t^{(i)} = \sum_{j\in G^{(i)}} \{X(t_{j+1}) - X(t_j)\}\{X(t_{j+1}) - X(t_j)\}'$. Then the estimate for daily Realized Volatility is $\widehat{RV}_t = \frac{1}{5}\sum_{i=1}^{5} \widehat{[X,X]}_t^{(i)}$. Liu et al. (2015) thoroughly investigate over 400 different estimators and find that 5 minute intervals (perhaps with 1-minute subsampling) is very hard to beat in terms of forecasting. Following their lead and the theoretical contributions of Ait-Sahalia et al. (2005a), that is the estimator we use. In our application, X is a vector of returns, which delivers a full Realized Covariance Matrix: $\widehat{RCov}_t$.

To create our sampling time grid, we use the first observed return within minute j as X_{t_j}

and fill in missing values with a return of 0. We also exclude the first and last 30 minutes of each trading day to avoid open and close effects.

Computing the daily Realized Betas in practice is a matter of simply taking components from the full Realized Covariance matrix described above:

$$\widehat{R\beta}_t = \frac{\widehat{RCov}(r_{it}^{HF}, f_t^{HF})}{\widehat{RV}(f_t^{HF})}.$$

Our method for computing realized measures is obviously not the only method of constructing a Realized Volatility — given the number of modeling choices including sampling rate, subsampling rate, functional form of the estimator (RV versus, say, a realized kernel), there are hundreds of volatility estimators. Briefly, the realized kernel estimator of Barndorff-Nielsen et al. (2011) is an advanced method for these purposes, and has been further improved upon by Hautsch and Kyj (2009) and Hautsch et al. (2011) in an effort to construct more efficient estimators in high dimensions. Hautsch et al. (2011) finds that regularizing the kernel density estimator has significant implications for portfolio management. However, an additional branch of literature suggests that the marginal gains of more advanced estimators relative to the complexity required to calculate them is unclear. Once again, we refer to Liu et al. (2015), who show that complexity usually does not significantly increase accuracy.

1.3.2 Data Transformations

As a potential issue, we recognize that despite the theoretical and practical support for the Ait-Sahalia et al. (2005a) estimator, it does leave out significant trading information since it ignores possible overnight changes in returns. Since the low-frequency data is constructed using close-to-close returns, this lack of overnight information results in a nontrivial discrepancy between the high frequency realized measures and the low frequency realized measure,

which is

$$\frac{1}{T}\sum_{t=1}^{T} r_t r_t^{'}.$$

We employ a simple scaling that matches the moments of realized measures of different frequencies, proposed in Sheppard and Xu (2014). Given

$$\begin{aligned}
\overline{\Sigma} &= \frac{1}{T}\sum_{t=1}^{T} r_t r_t^{'}, \\
\overline{M} &= \frac{1}{T}\sum_{t=1}^{T} \widehat{RCov}_t, \\
\overline{\Gamma} &= \overline{\Sigma}^{1/2}\overline{M}^{-1/2}.
\end{aligned}$$

Then define the scaled realized covariance:

$$\widetilde{RC}_t = \overline{\Gamma}\widehat{RCov}_t\overline{\Gamma}.$$

This yields

$$\frac{1}{T}\sum_{t=1}^{T} \widetilde{RC}_t = \frac{1}{T}\sum_{t=1}^{T} r_t r_t^{'}.$$

As long as T is sufficiently larger than N, this transformation will be numerically stable. We apply the transformation to the entire Realized Covariance matrix, and then use the transformed values to construct Realized Betas. This means that although the moments for the full realized covariance match the low-frequency counterparts, the moments for realized betas do not. In practice we find that this overnight transformation does not impact the qualitative results, but in combination with improved intraday realized measures it is important to consider.

1.3.3 Estimation Procedure

Whether market volatility is precisely the factor for idiosyncratic volatility presents three possible DGPs, which in turn should influence theory and mechanisms explaining the phenomenon. There are three distinct cases for how the two can be related, and they lead to three separate models of interest that we must estimate:

1. The factor(s) for idiosyncratic volatility are precisely the volatilities of the market factor. This is the case employed in Kalnina and Tewou (2015). We call this FVOL MKT.
2. The factor(s) for idiosyncratic volatility are orthogonal to the volatilities of the market factors. We call this FVOL2.
3. The factor(s) for idiosyncratic volatility are separate from, though highly correlated with, the volatilities of the market factor. This case remains largely unexplored, though is related to work in Chen and Petkova (2012). We call this FVOL PCA.

While Duarte et al. (2014), Herskovic et al. (2014), Barigozzi and Hallin (2014), and Christoffersen et al. (2014) all utilize a statistical factor as their factor for idiosyncratic volatility, they do not comment on the relationship between Market Volatility and their statistical factor. It is therefore difficult to discern whether they support FVOL2 or FVOL PCA.

Case 1 would correspond to the following model:

$$
\begin{aligned}
r_t &= \beta F_t + e_t \\
\log(\sigma_{F_t}) &= \mu_F + \beta_F \log(\sigma_{F_{t-1}}) + u_t^F \\
\log(\sigma_{e_t^i}) &= \mu_i + \beta_i^e \log(\sigma_{F_t}) + u_t^i
\end{aligned}
$$

Case 2 would correspond to:

$$r_t = \beta F_t + e_t$$
$$\log(\sigma_{F_t}) = \mu_F + \beta_F \log(\sigma_{F_{t-1}}) + u_t^F$$
$$\log(\sigma_{e_t^i}) = \mu_i + \beta_i^e \log(\sigma_{F_t}) + \gamma_i V_t + u_t^i$$

Where V_t is an additional factor for volatility. The third case is if idiosyncratic volatility is orthogonal to market volatility, $\beta_i = 0$. Beginning with high frequency returns r_{t_j}, we proceed as follows.

- At each date t, we run the intraday regression

$$r_{t_j} = \beta_t f_{t_j} + v_{t_j}, \qquad j = 1, \dots M_t$$

- We construct the daily estimate of realized volatility for f_t and v_t according to Section 1.2.1. In practice, we compute the entire $RCov$ for $[r_t, f_t]$, which is an $(N+1)\times(N+1)$ matrix.

- We conduct the data transformations, namely the scaling transformation to adjust for overnight returns, according to Section 1.3.2.

- Decompose the adjusted $RCov$ into market volatility, σ_t^f and idiosyncratic volatility, $diag(\Omega_{v_t})$.

- Finally, collect all elements of $diag(\Omega_{v_t})$ into a $T \times N$ panel.

- Analyze the panel according to the applicable model

    1. FVOL MKT - Single factor on idiosyncratic volatility, where the factor is market volatility.
    2. FVOL2 - Two factor model on idiosyncratic volatility, where the first factor is

market volatility, and the second factor is extracted via PCA from the residuals.

3. FVOL PCA - Single factor on idiosyncratic volatility, where the factor is extracted via PCA on the idiosyncratic volatility panel.

1.3.4 Equities: Factor structure in Idiosyncratic Volatility

For both datasets, we start by verifying that idiosyncratic volatility is indeed dynamic and exhibits factor structure. We verify that it is dynamic by running univariate autoregressions with lag length chosen by AIC, all of which reject the null hypothesis of constant volatility with white noise. We verify factor structure by visual inspection of the panel and scree plots, which can be found in Figure 2.

1.3.5 Relationship between factor for volatility and factor volatility

Based on the figures, it is clear that there exists factor structure in idiosyncratic volatility. This is consistent with prior research in the field, as in Herskovic et al. (2014), Barigozzi and Hallin (2014), Kalnina and Tewou (2015) and Duarte et al. (2014). However, what is not clear from the above literature is the relationship between the factor for idiosyncratic volatility (i.e. the first principal component of the panel) and the volatility of the market factor. Kalnina and Tewou (2015) assume that they are the same, while Herskovic et al. (2014) and Barigozzi and Hallin (2014) do not.

In the following two sections, we argue that while the factor for idiosyncratic volatility and market volatility are highly correlated, they are not the same. We argue these facts based on graphical analysis and a battery of statistical tests from the panel data literature.

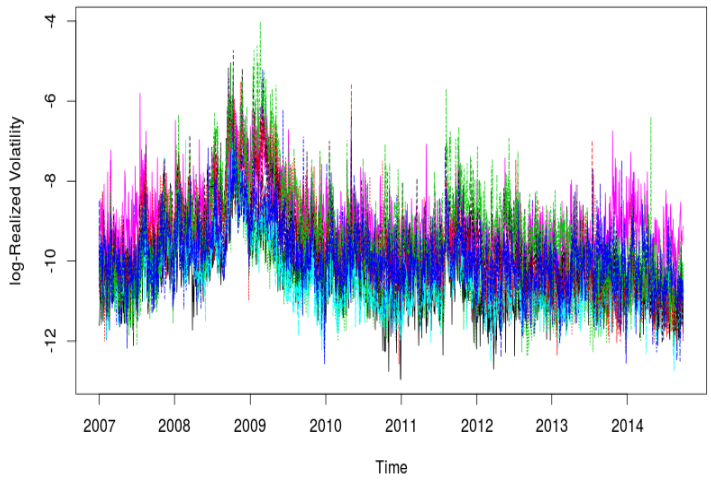
Graphical Analysis

We start by presenting the volatility of the market factor (SPY) overlaid on the plots of idiosyncratic volatility. The plots are in Figure 3. Taken together, the plots suggest that the market volatility explains amount of cross-sectional variation in the panel of idiosyncratic volatility. For the DOW 10, market volatility explains, on average 50% of cross sectional

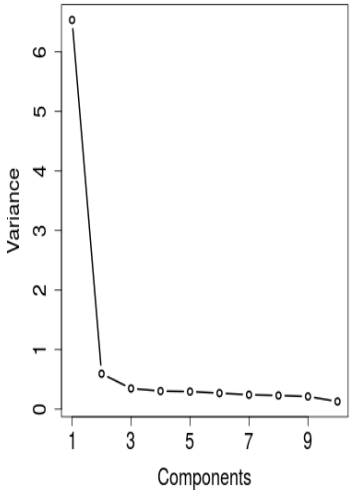
Figure 2: Factor Structure in Equities Idiosyncratic Volatility

Figures 2a and 2c plot the log-realized volatilities of the DOW 10 and S&P 100 datasets from 2007 - 2015. Figures 2b and 2d plots corresponding scree plots (variances of the first 10 principal components)

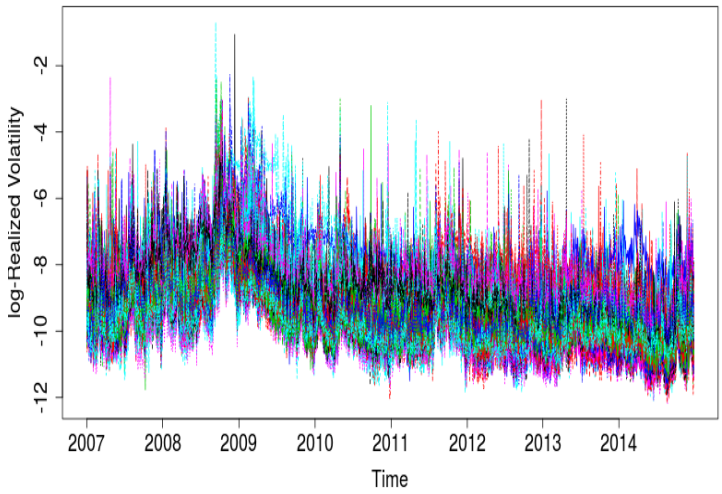
(a) DOW 10 Idiosyncratic Volatility



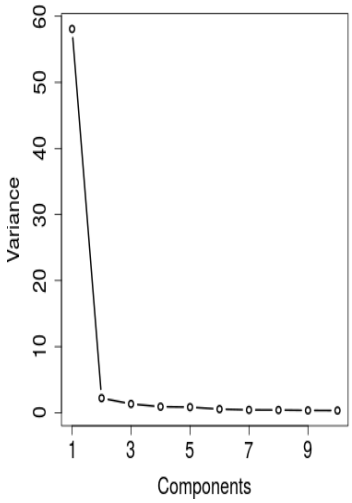
(b) DOW 10 Scree Plot



(c) S&P 100 Idiosyncratic Volatility



(d) S&P 100 Scree Plot



variation, while for the S&P100, it explains 55%. The distribution of explained variation across assets is in Figure 4. The explained variation is rather high for both panels, especially considering the naive modeling strategy would presume market volatility is unrelated to idiosyncratic volatility. These images heuristically support the methods in Kalnina and Tewou (2015).

However, we also entertain the idea, as in Duarte et al. (2014), Herskovic et al. (2014), and

Figure 3: Market Volatility and Idiosyncratic Volatility

The log-volatilities of the panel plotted against the SPY index volatility (in black) from 2007-2015.

(a) DOW 10 (b) S&P 100

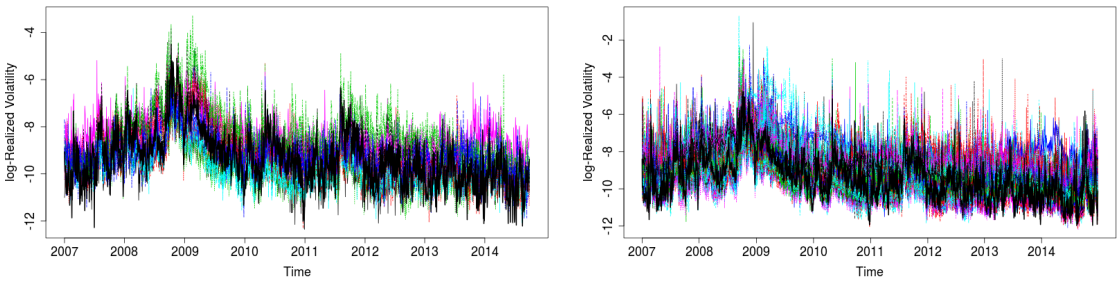
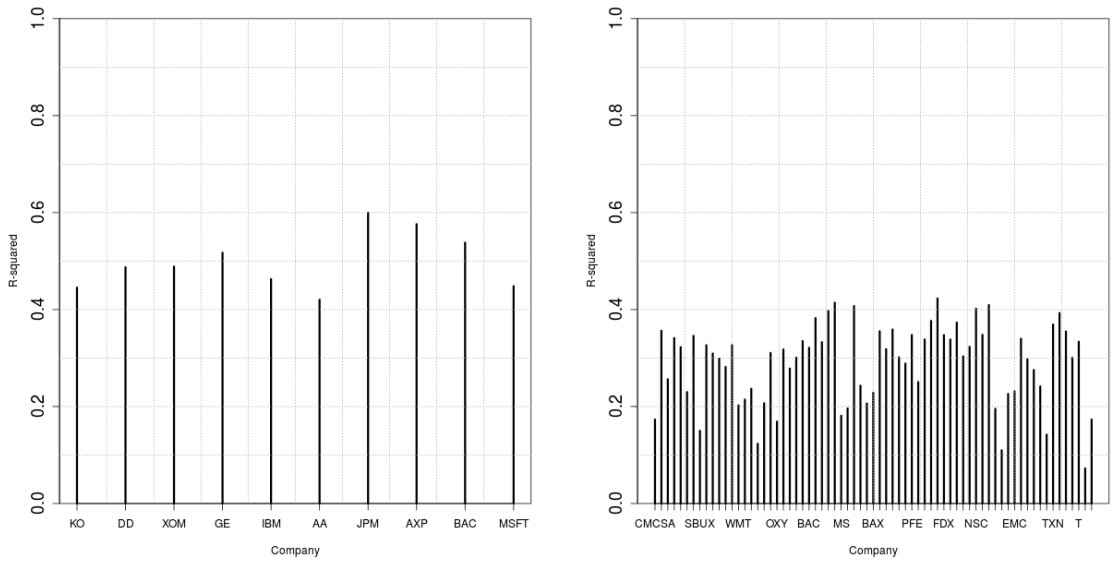


Figure 4: Market Volatility: Explained Variation

The panel of R^2 for each asset volatility in the panel regressed against SPY volatility. Approximately the same fraction of variation is explained by the SPY for each asset.

(a) DOW 10 (b) S&P 100

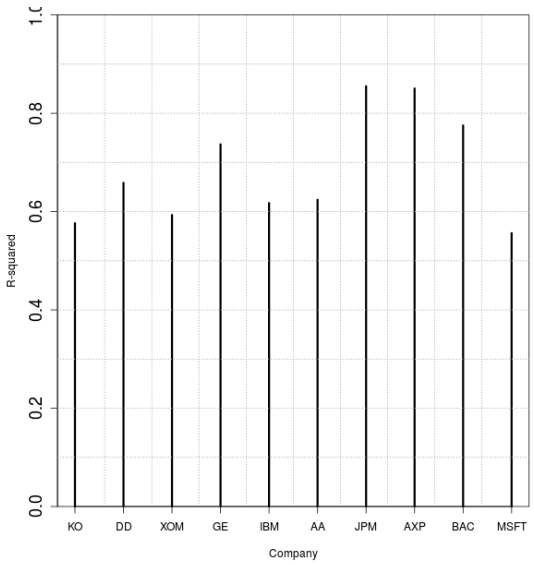


others, that the factor for idiosyncratic volatility is a separate, PCA factor, that is possibly unrelated to market volatility. To support this, we present the distribution of explained variation, but this time with the first principal component of the panel of idiosyncratic volatilities replacing that of market volatility. These are in Figure 5. The average cross sectional R^2 in the DOW 10 panel is 68%, while that in the S&P 100 panel is 76%. Unsurprisingly the first PC explains substantially more cross sectional variation than does market volatility. This supports Model 3.

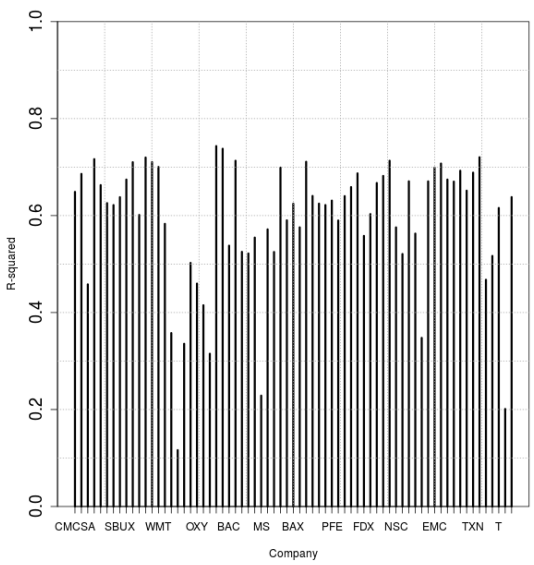
Figure 5: First PC of Idiosyncratic Volatility

The panel of R^2 for each asset volatility in the panel regressed against first principal component of idiosyncratic volatility.

(a) DOW 10



(b) S&P 100

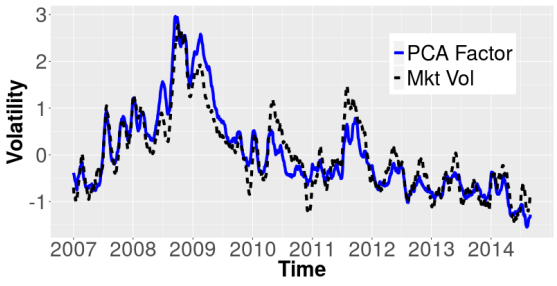


Lastly, we also show that while the first PC explains more cross sectional variation than market volatility, the two are nonetheless highly correlated. In Figure 6 we plot the 22-day rolling average of the PCA factor and the Market log-Volatility. For both equities datasets, the correlation between the two (unsmoothed) is 0.85.

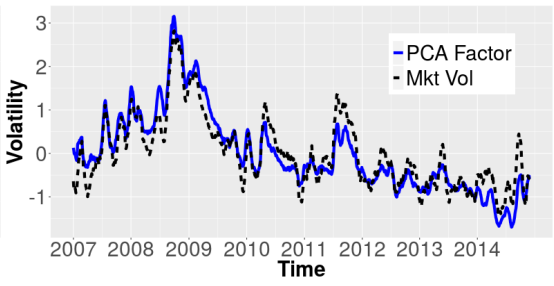
Figure 6: Market Volatility and First PC of Idiosyncratic Volatility

Each plot displays the 22-day rolling mean of the Market Volatility (black, dashed line) and the First PC of Idiosyncratic Volatility (blue, solid line) for that panel. Both volatilities have been centered and scaled to have mean 0 and variance 1.

(a) DOW 10



(b) S&P 100



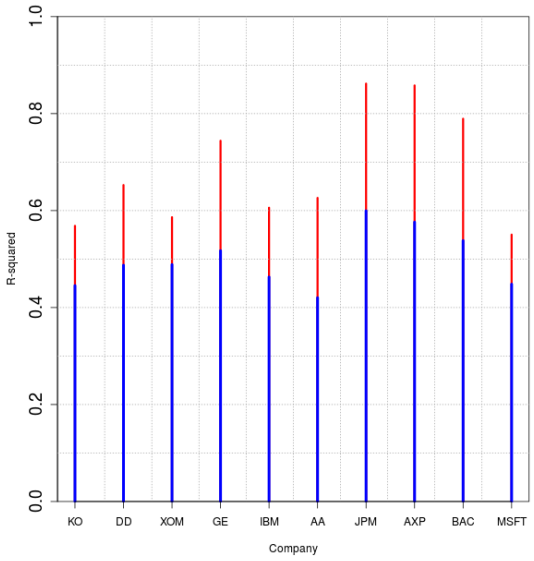
Despite this high correlation, we also consider whether both market volatility and the PC

factor are important for explaining cross sectional variation in the panel. This is the Model 2 paradigm. We therefore once again plot the distribution of explained variations in Figure 7 with two factors — the first is the market volatility and the second is a PCA factor on residuals after regressing out market volatility. In this case, the average cross sectional R^2 for DOW 10 is 68%, while that in the S&P 100 panel is 76%.

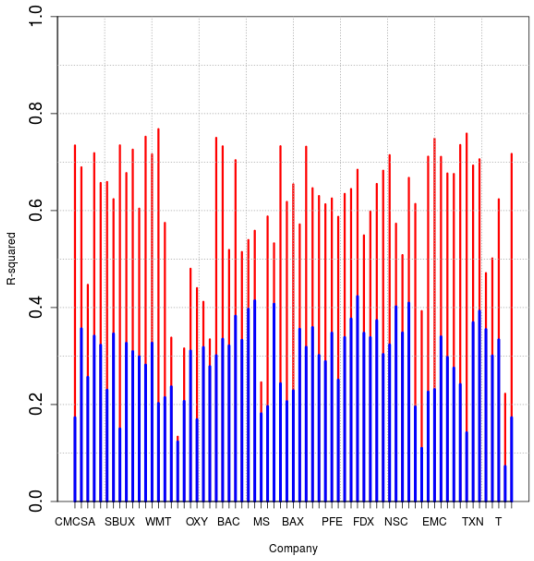
Figure 7: Market Volatility and PCA factor of Idiosyncratic Volatility

In blue, the panel of R^2 for each asset volatility in the panel regressed against SPY volatility. In red is the increased in R^2 from also regressing against first PCA.

(a) DOW 10



(b) S&P 100



One should note that the average explained variation for the two-factor paradigm is exactly the same as that for the principal components factor. Based on that observations, one might think that the market volatility plus a PCA factor merely spans the same space as the first PCA factor. Supporting this claim would be the fact that the canonical correlation between the first PCA factor and the two-factor model is almost exactly 1. Despite that, the two are not the same, insofar as it relates to explaining the panel of idiosyncratic volatility. Indeed, some assets are better explained by the two factor paradigm, and others are better explained by the principal components factor. Thus, while a linear combination of the two factors can nearly exactly generate the first PC factor, that linear combination is not optimal for explaining the panel.

Overall, graphical analysis supports the idea that both Model 2 or Model 3 are highly plausible. Despite the high correlation between Market Volatility and the first PC of Idiosyncratic Volatility, the PCA factor is able to explain a much larger share of overall variation.

Statistical Tests

We propose a series of statistical tests for whether the factor for idiosyncratic volatility is the same, related or different from market volatility. We propose two versions of a likelihood ratio test, a test of factor structure from Onatski (2009), and a test for relating an observed factor to a PCA factor that is due to Bai and Ng (2006).

Using a normality assumption, we can use a likelihood ratio test for $\beta_i^e = 0 \; \forall i$ in order to differentiate between cases 2 and 3. However, there are two LR tests necessary, since the construction of V_t via Principal Components will be different depending on whether the market volatility has been regressed out or not. As shown above, before regressing out the market volatility, the factor for idiosyncratic volatility is highly correlated with market volatility. As such, one would expect that if V_t is extracted from the entire panel of idiosyncratic volatility, then V_t might mainly include redundant information with σ_{f_t}. As such, we wish to test whether σ_{f_t} includes new information both before and after V_t has been extracted. In test LR-1 we construct V_t based on the residuals from first regressing out σ_{f_t}. In test LR-2, we construct V_t on the full panel, before regressing out σ_{f_t}. We expect, and find, that the test statistics for LR-1 are always substantially larger than those for LR-2. The LR test has asymptotic distribution as χ^2_k, where k is the number of restrictions imposed. In all cases, k is the size of the cross sectional dimension.

In addition to a likelihood ratio test, we consider tests motivated by Bai and Ng (2006) and Onatski (2009). The former consists of using an observed factor G_t and PCA factor F_t, where the null hypothesis is that they are statistically the same. To deal with non-identification of the factor under rotation, the test statistics are constructed via canonical correlations as follows. Suppose we regress G_t against F_t, yielding $\widehat{G}_t$. Then construct

$$\widehat{\tau}_t = \frac{\left(\widehat{G}_t - G_t\right)}{\widehat{var}(\widehat{G}_t)^{(1/2)}}$$

In other words, this is the t-statistic for the null that G_t is spanned by F_t. Let Φ_α^τ be the α percentage point of the standard normal distribution. Then the statistics are

$$A = \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}\left(|\widehat{\tau}_t| > \Phi_\alpha^\tau\right)$$
$$M = max|\widehat{\tau}_t|$$

These exact tests have asymptotic distributions

$$A \rightarrow_p 2\alpha$$
$$M \text{ such that } P(M \leq x) \approx 2\Phi(x) - 1.$$

The rejection region for test M is found via simulation, as the $(1-\alpha)$ quantile of the maximum absolute value of standard normal vectors of length T.

They also propose approximate tests that are more heuristic. Consider regressing G_t against F_t. Then, under the null, the noise-to-signal ratio should be 0 and the R^2 should be one. The heuristic tests say that the R^2 should be "high," and the noise-to-signal ratio should be "low."

Lastly, we consider the test from Onatski (2009), which examines the number of factors in a panel. The exact test statistic is

$$R = \max_{k_0 < i \leq k_1} \frac{\gamma_i - \gamma_{i+1}}{\gamma_{i+1} - \gamma_{i+2}}, \quad 0 \leq k_0 < k_1 \leq N - 2,$$

where γ_i is the i$^{\text{th}}$ largest eigenvalue of the smoothed periodogram estimate of the spectral density matrix of data at a prespecified frequency. This test is valid for testing against a null of 0 factors. Therefore, after regressing out the market volatility, we test for the presence of factor structure, where the null hypothesis is no factor structure, and the alternative is anywhere from 1-3 factors. The test statistic has asymptotic Tracy-Widom distribution, whose critical values are tabulated in Onatski (2009).

For clarity, consider Table 1, where we provide the behavior of each test under the null hypotheses for Models 1-3 respectively.

Table 1: Statistical Tests Explained - Expected Outcomes

Test	Model 1	Model 2	Model 3
LR -1	Power, $\beta \neq 0$	Power, Reject	Power to correlated regressor, Reject
LR -2	Under-reject (correlated regressors)	Power, Reject	Under-reject (correlated regressors)
Onatski	Correct size	Power, Reject	Power, Reject more than 5%
A	0.1	N/A	N/A
M	~ 4	N/A	N/A
NS	~ 0	Moderately low	Moderately low
R^2	~ 1	Moderately high	Moderately high

The tests are statistically conclusive, and provide statistically significant estimates (except LR-2 for the DOW 10). All Bai and Ng (2006) easily reject the null that market volatility is the same as the PCA factor. The Onatski (2009) test supports the existence of at least one more factor after regressing out market volatility. The LR-1 test resoundingly rejects the null for both datasets, which supports the graphical evidence that the market volatility is a driver of the overall panel. The LR-2 null hypothesis is rejected for the S&P 100, but not for the DOW 10. This suggests that for the DOW 10 dataset the market volatility might

be extraneous once the first PC is extracted, but that for the S&P 100 dataset, the market volatility still holds meaningful information for the cross-section even after extracting the first PC. All results can be found in Table 2.

Table 2: Statistical Tests for Equities

Table with statistical tests for the two equities datasets (DOW 10 and S&P 100). LR-1 and LR-2 tests display likelihood ratio statistics for the null hypothesis that the coefficients on market volatility should be 0. LR-1 performs the test on the panel of idiosyncratic volatilities, while LR-2 performs the test on panel residuals after extracting the first Principal Component. Onatski is the test for factor structure described in Onatski (2009) where the null hypothesis is that there is no factor structure after regressing out the market volatility. A and M are exact tests from Bai and Ng (2006), while NS and R^2 are approximate tests from the same paper. Note that A has no critical values, but the test statistic should converge to 2α for α confidence level. ** denotes significant at 5%, *** denotes significant at 1%.

Test	DOW 10	S&P 100
LR - 1	36461^{***}	471729^{***}
LR - 2	13.72	135^{***}
Onatski	12.13^{***}	12.36^{***}
A	0.50^{***}	0.84^{***}
M	14^{***}	75^{***}
NS	0.38	0.37
R^2	0.72	0.73
$CI(R^2)$	(0.70, 0.79)	(0.71, 0.75)

The statistical tests therefore strongly support Model 2. Both market volatility *and* a principal component factor are needed to explain the panel. The two are not the same, and neither makes the other extraneous.

1.4 Foreign-Exchange Rates

Next we move on to our analysis of Foreign Exchange rate returns. We consider a panel of 15 exchange rates from major economies (a full list can be found in the Appendix).

Our data consists of daily FX returns downloaded from FRED, confined to the post-Euro era, so our sample runs from January 1999 to October 2015. Since the returns are daily, we aggregate to monthly realized volatility. For the market factor, we use an equal-weighted average of all the returns. This index is 99.9% correlated with the first principal component of returns. The order of our estimation procedure is exactly analogous to our equities data

analysis, save that the frequencies are all shifted to be lower — high frequency exchange rate returns are now daily returns.

1.4.1 Comparison with Equities Results

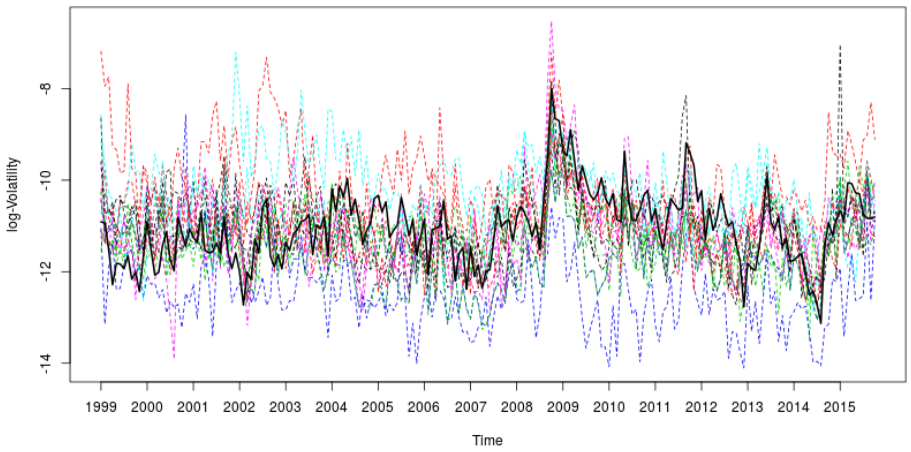
Our motivation for considering FX data is that we are interested if the structure in idiosyncratic volatility is confined merely to equities or also applies to other financial datasets. It turns out that many of the general features present in equities is also present in FX, though there are some important differences. As before, we start with a graphical analysis of the data and then continue on to the statistical tests.

GraphicalAnalysis

The graphical analysis begins in Figure 8, where we present the panel of idiosyncratic volatility together with the market volatility. As in the case of equities, there are clear dynamics in idiosyncratic volatility, and they display factor structure. Moreover, the market volatility has dynamics consistent with the rest of the panel. In contrast to equities, the factor structure seems weaker here, as individual exchange rates frequently deviate from the rest of the panel.

Figure 8: FX Factor and Idiosyncratic Volatility

Panel of log exchange rate volatilities from 1999-2016 (AL, BZ, CA, DN, JP, KO, MX, NZ, NO, SI, SF, SZ, UK, EU). In black is the equal-weighted average of all returns (approximately the first PCA).

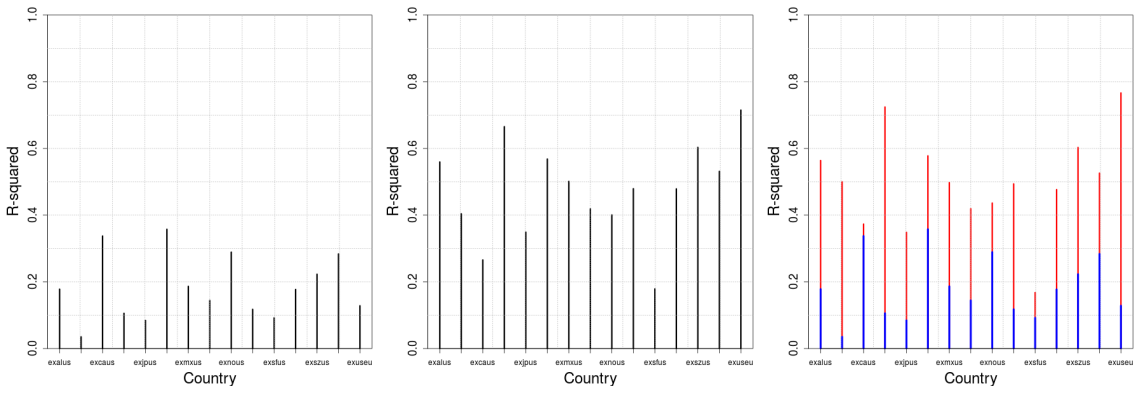


The weaker factor structure is further supported by Figure 9. Whereas in equities the average R^2 were 50% and 55% for DOW 10 and S&P 100, respectively, the market volatility only explains, on average, 18% of cross sectional variation. Additionally, the first PC explains only 47% of cross sectional variation, compared to 68% and 76% for the DOW 10 and S&P 100. Similar to equities, when we take two factors, the structure is familiar, though again, the levels are lower. Market volatility and a PC factor explain 50% of the cross sectional variation (as compared to 68% and 76% for DOW 10 and S&P 100).

Figure 9: Explained Variation in FX Idiosyncratic Volatility

Panel of R^2 for each exchange rate when regressed against market volatility (the equal weighted average), the first PCA, and both. Despite high correlation of market volatility and first PCA, the first PCA has on average greater explanatory power. However, for a few assets, the gains from adding market volatility to the first PCA are also nontrivial (see BZ and CA).

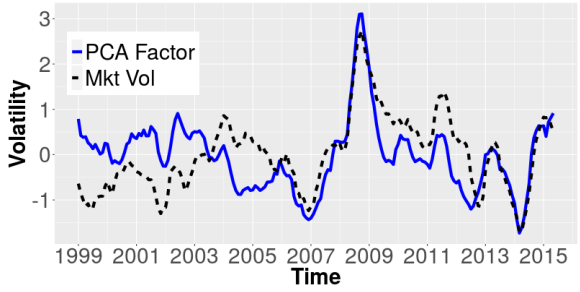
(a) Market Vol (b) First PC (c) Market Vol + PC



Thus, in the case of FX returns, there are three major differences. First, the factor is much weaker. No matter which factor you use, the amount of cross-sectional variation is substantially lower. Second, the discrepancy between average explained variation from market volatility and the first PC is much larger. The first PC explains almost 30 percentage points more than cross-sectional variation of FX returns. Lastly, the two are much more dissimilar than their counterparts in the equities datasets. Indeed, the correlation between market volatility and the first PC of idiosyncratic volatility is 0.57, which is much lower than the 0.85 for both equities datasets. The 6-month rolling window of the PCA factor and the market volatility are plotted in Figure 10.

Figure 10: PCA factor vs Market Volatility

6 month rolling window of volatility. Blue (solid) line displays the first PC of idiosyncratic volatility, while the black (dashed) line displays the market volatility. Both volatilities have been centered and scaled to have mean 0 and variance 1.



Thus, from graphical analysis, we immediately gain insight into similarities and differences between FX returns and equity returns. In the case of FX, market return does not do a good job explaining cross sectional variation, whereas the first PC does much better. Indeed, they are weakly correlated at 57%. Nonetheless, when the two are paired together, most cross sectional variation is explained. Once again, the average R^2 from the two factor model is exactly the same as that of only the first PC, but similar to equities, the distribution of R^2s is not the same.

Statistical Tests

We run the same battery of statistical tests on the FX data as we did equities. Due to the graphical analysis above, we expect to easily reject the null that the PC factor is the same as market volatility (Bai and Ng (2006) tests) and that there is no factor structure once the market is taken into account (Onatski (2009) test). Somewhat surprisingly though, both LR tests also reject the null that market volatility should not be included at all. All results are in Table 3.

In conclusion, both datasets support the notion that there is factor structure in idiosyncratic volatility and that the panel of idiosyncratic volatility is best explained via two factors; one is the market factor and one is a PC factor.

Table 3: Statistical Tests for FX panel

Table with statistical tests for the FX rate dataset. LR-1 and LR-2 tests display likelihood ratio statistics for the null hypothesis that the coefficients on market volatility should be 0. LR-1 performs the test on the panel of idiosyncratic volatilities, while LR-2 performs the test on panel residuals after extracting the first Principal Component. Onatski is the test for factor structure described in Onatski (2009) where the null hypothesis is that there is no factor structure after regressing out the market volatility. A and M are exact tests from Bai and Ng (2006), while NS and R^2 are approximate tests from the same paper. Note that A has no critical values, but the test statistic should converge to 2α for α confidence level. ** denotes significant at 5%, *** denotes significant at 1%.

Test	Forex
LR - 1	1221^{***}
LR - 2	95^{***}
Onatski	17.46^{***}
A	0.76^{***}
M	16^{***}
NS	1.86
R^2	0.35
$CI(R^2)$	(0.24, 0.45)

1.5 Conditional Mean Misspecification

As explained earlier, one misspecification that could generate the factor structure is time-variation in the factor loadings. Another is conditional mean misspecification. As a preliminary exercise, observe that if the true DGP is:

$$y_t = \beta_1 f_t + \beta_2 X_t + e_t$$

$$f_t \sim N(0, \sigma^2_{f,t}) \qquad X_t \sim (0, \sigma^2_X)$$

Yet the estimated model is:

$$y_t = \bar{\beta}_1 f_t + \bar{e}_t$$

Then:

$$\mathbb{E}[\bar{\beta}_1] = \beta_1 + \beta_2 \frac{Cov(f_t, X_t)}{\mathbb{V}[f_t]} = \beta_1 \quad \text{if } Cov(f_t, X_t) = 0$$

$$\bar{e}_t = \beta_2(X_t) + e_t$$

$$\mathbb{V}_t[\bar{e}_t] = 2\sigma_X^2 \beta_2 \beta_2^{'} + \mathbb{V}_t[e_t]$$

Even if $\mathbb{V}_t[e_t] = c$, $\mathbb{V}_t[\bar{e}_t]$ will be time-varying with factor structure. If X_t is a function of f_t, in particular suppose the conditional mean is a higher-order polynomial of f_t, $\mathbb{V}_t[\bar{e}_t]$ will also comove with market volatility! While this example is obviously contrived, it is important to point out that in the presence of *any* omitted factors from the level equation, there will be factor structure in idiosyncratic volatility. Indeed, Herskovic et al. (2014) were aware of this issue and fit a large factor model (5 principal components) to the level equation, but still found the same structure in idiosyncratic volatility. Since we are specifically interested in how the structure might effect the relationship with factor volatility, we investigate whether there might be omitted factors due to omitted nonlinearities of f_t in the conditional mean. To explore this question, we run our intraday factor regression with four powers of the observed factor.

For the DOW10, the correlation between the market factor and the first PC of idiosyncratic volatility is 0.85, exactly the same as it was when we fit a single factor. For the S&P100, the correlation between market volatility and the first PC of idiosyncratic volatility drops from 0.85 to 0.56. While this drop is fairly large, our statistical tests, especially the LR tests, show that the market volatility is still a vital component of the panel. While the Bai and Ng (2006) test produces a statistic for all four observed factors (the volatilities of the powers of market returns), we only report statistics for the market volatility, as the results are nearly identical for all of them. The results of the statistical tests are in Table 4. A particularly striking result is the difference between the second likelihood ratio test at $N = 10$ (the DOW10) and $N = 100$ (the S&P100). This is likely owing to the blessing of dimensionality

and improved inference of factor structure as N becomes large.

Table 4: Statistical Tests for Higher Powers of Market Return

Table with statistical tests for the two equities datasets (DOW 10 and S&P 100) where the factors are the first four powers of the observed market factor (SPY). LR-1 and LR-2 tests display likelihood ratio statistics for the null hypothesis that the coefficients on market volatility should be 0. LR-1 performs the test on the panel of idiosyncratic volatilities, while LR-2 performs the test on panel residuals after extracting the first Principal Component. Onatski is the test for factor structure described in Onatski (2009) where the null hypothesis is that there is no factor structure after regressing out the market volatility. A and M are exact tests from Bai and Ng (2006), while NS and R^2 are approximate tests from the same paper. Note that A has no critical values, but the test statistic should converge to 2α for α confidence level. While the Bai and Ng (2006) tests generate test statistics for each of the four powers, we only report the results for the first power (market volatility). ** denotes significant at 5%, *** denotes significant at 1%.

Test	DOW 10	S&P 100
LR - 1	33603.22^{***}	395929^{***}
LR - 2	41.1511	15436^{***}
Onatski	8.75^{**}	122^{***}
A	0.87^{***}	0.89^{***}
M	64^{***}	124^{***}
NS	0.38	1.88
R^2	0.72	0.34
$CI(R^2)$	(0.70, 0.75)	(0.32, 0.36)

While the test statistics change, since we are now testing more restrictions (in the case of the LR tests), the overall picture is still the same. The LR tests are all resoundingly rejected, so the volatilities of powers of market returns cannot be excluded from the model.

1.6 Forecasting

In addition to assessing the relationship between the factor for idiosyncratic volatility and market volatility, we also explore what, if any, impact the factor for volatility has on volatility forecasting. In addition to the three models we presented in Section 1.3.5, we include two additional benchmark models:

1. BMK - The benchmark model where only the factor has time-varying volatility (constant idiosyncratic volatility). Jacquier et al. (1994) proposed a Stochastic Volatility version of this model, though they did not estimate it. Diebold and Nerlove (1989) proposed and estimated a similar model, where the factor volatility is an ARCH pro-

cess.

2. AR - In addition to time-varying volatility in the factor, idiosyncratic volatility is also time-varying, but they vary as independent autoregressions. Kim et al. (1998) proposed this multivariate stochastic volatility model, though Pitt and Shephard (1999) and Aguilar and West (2000) independently (and with different MCMC techniques) actually produced estimation procedures.

In order to estimate our three Factor for Idiosyncratic Volatility models, we proceed in one of the following ways:

- For model 1 of 3 (FVOL MKT), we regress each of the log- diagonal vector of Ω_{v_t}, $\sigma_{e_t^i}$, against $\log(\sigma_t^f)$ to estimate β_i.

- For model 2 of 3 (FVOL2), regress $\log(\sigma_{e_t^i})$ against $\log(\sigma_{F_t})$, and conduct PCA on the panel of residuals.

- For model 3 of 3 (FVOL PCA), we conduct PCA directly on the panel of log-idiosyncratic volatlitities. $\sigma_{e_t^i}$. We then regress the residuals against σ_{F_t}.

For all datasets we focus on the forecast errors of the panel of variances. Correlations are modeled via loadings from the level regression, which are the same for all models. All models and datasets forecast poorly at the beginning of the financial crisis in 2008, so we report both average Mean Squared Error (MSE) and Median Absolute Error (MAE), where the mean/median is taken across time for each asset and then averaged across assets. We also plot the cumulative squared one-step ahead forecast errors, both for the whole sample and pre- and post-2008 (FX is also plotted pre-2008).

1.6.1 Equities

For both equities datasets, we use a 200 day rolling window estimation period. In each period we estimate each of the five competing models and forecast ahead 1-12 days. Due to the fact that there are some large outliers (even outside the financial crisis), we record both

Average MSE and MAE. The DOW 10 forecasting results are presented in Table 5, while results for the S&P 100 dataset are in Table 6. One-step-ahead cumulative squared forecast errors for both datasets are plotted in Figure 11. To ensure the results are not solely driven by dynamics in the crisis, we also present (in the Appendix) tables of forecasting results and figures with squared forecast errors using forecasts only after January 2009. The DOW 10 forecasting results are in Table 12, while the S&P 100 results are in Table 13. Squared forecast errors for both datasets are plotted in Figure 13.

Table 5: Average Mean Square Error and Median Absolute Error of DOW 10 Rvariances

All values are relative to BMK forecasts. Bolded value in each row is the minimum, when better than BMK. BMK is benchmark, AR is with univariate autoregressive idiosyncratic volatility, FVOL MKT uses market volatility as a single idiosyncratic vol factor, FVOL PCA uses a single principal component as an idiosyncratic vol factor, FVOL 2 uses both. All models use a 200-day rolling window to estimate parameters, followed by forecasts for 1-12 days ahead.

	Average MSE				Average MAE			
h	AR	FVOL MKT	FVOL PCA	FVOL2	AR	FVOL MKT	FVOL PCA	FVOL2
1	**0.82**	1.01	1.03	1.30	0.83	0.89	**0.82**	0.83
2	**0.86**	0.99	0.98	1.11	0.88	0.90	**0.83**	0.85
3	**0.90**	1.00	1.02	1.08	0.90	0.91	**0.85**	0.87
4	**0.92**	0.99	0.99	1.05	0.90	0.90	**0.85**	0.88
5	**0.93**	1.00	1.05	1.06	0.91	0.91	**0.85**	0.88
6	**0.94**	1.01	1.03	1.05	0.92	0.92	**0.86**	0.90
7	**0.94**	1.02	1.05	1.03	0.94	0.94	**0.87**	0.91
8	**0.95**	1.01	1.02	1.02	0.94	0.94	**0.86**	0.91
9	**0.94**	1.02	1.03	1.00	0.94	0.93	**0.86**	0.91
10	**0.94**	1.01	1.01	1.00	0.94	0.94	**0.87**	0.92
11	**0.95**	1.01	1.02	1.00	0.94	0.94	**0.88**	0.93
12	**0.95**	1.01	1.00	1.00	0.95	0.94	**0.88**	0.94

First focus on the DOW 10 dataset in Table 5. By average MSE, all FVOL models forecast variances about as well, though FVOL 2 does slightly worse than the others at short horizons. In addition, the model of Pitt and Shephard (1999) (AR) does very well, clearly supporting the hypothesis that idiosyncratic variance is at least time-varying. Despite the FVOL models not performing particularly well, their worse performance is mainly centered around the financial crisis, specifically around late 2008. When we look at average MAE instead of MSE, we see that all models provide substantial forecasting improvements as compared to

Table 6: Average Mean Square Error and Median Absolute Error of S&P 100 Rvariances

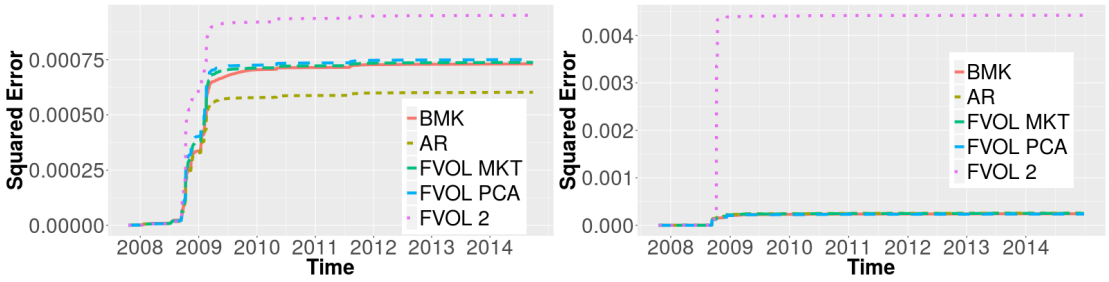
All values are relative to BMK forecasts. Bolded value in each row is the minimum, when better than BMK. BMK is benchmark, AR is with univariate autoregressive idiosyncratic volatility, FVOL MKT uses market volatility as a single idiosyncratic vol factor, FVOL PCA uses a single principal component as an idiosyncratic vol factor, FVOL 2 uses both. All models use a 200-day rolling window to estimate parameters, followed by forecasts for 1-12 days ahead.

	Average MSE				Average MAE			
h	AR	FVOL MKT	FVOL PCA	FVOL2	AR	FVOL MKT	FVOL PCA	FVOL2
1	1.06	1.04	**0.97**	18.18	0.70	0.79	0.70	**0.70**
2	1.17	1.05	**0.98**	27.19	0.76	0.83	**0.74**	0.74
3	1.18	1.07	**0.99**	9.77	0.79	0.84	**0.75**	0.77
4	1.34	1.09	**0.98**	15.11	0.81	0.85	**0.75**	0.77
5	1.16	1.08	**0.98**	9.49	0.83	0.86	**0.76**	0.79
6	1.29	1.04	**0.99**	7.58	0.85	0.87	**0.77**	0.81
7	1.22	1.07	**1.00**	10.56	0.86	0.87	**0.78**	0.82
8	1.18	1.05	**0.99**	4.61	0.87	0.89	**0.79**	0.84
9	1.24	1.11	**0.99**	3.97	0.88	0.89	**0.79**	0.84
10	1.28	1.12	**0.99**	3.23	0.89	0.90	**0.79**	0.84
11	1.26	1.13	**0.99**	2.49	0.89	0.91	**0.79**	0.86
12	1.24	1.16	**0.99**	1.36	0.90	0.91	**0.81**	0.87

Figure 11: Equities Squared One-Step Prediction Errors

Cumulative squared errors over time, 2007-2015, of DOW 10 and S&P100. Each date adds the average squared distance of true volatility to forecasted volatility over the panel. The models perform similarly outside the financial crisis 2008-2010, but there the discrepancies are large.

(a) DOW10 (b) S&P 100



the benchmark model. The Pitt and Shephard (1999) (AR) model still performs about as well, but introducing some sort of factor on idiosyncratic volatility also performs comparably well with much fewer estimated parameters. Specifically, using a PCA factor to forecast idiosyncratic volatility works best at all horizons.

In the larger, S&P 100, sample, the results are qualitatively similar. Once again, all models perform very similarly when compared via average MSE. This time though, the AR model slightly underperforms the benchmark, the PCA factor slightly outperforms the benchmark, and the FVOL 2 model performs substantially worse. Once again though, the forecasting deficiencies are mainly due to the financial crisis, and by using average MAE, all FVOL models see large improvements over the benchmark model. Once again, the AR model performs very well, but this time both FVOL PCA and FVOL2 do even better. The FVOL MKT once again underperforms the other models, but still beats the benchmark.

Taken together, as the panel of volatilities grows in cross-sectional dimension, the improvements of using FVOL models increases. While using both the market volatility (model 1) and the PCA factor are each helpful, the PCA factor is better for forecasting. This reaffirms the traditional “Blessing of Dimensionality" in factor models - that when dimensions grow, there are increasingly large benefits to fitting factor models rather than attempting to model each series individually.

1.6.2 Exchange Rates

We use the same set of competing models to predict FX monthly volatilities, but this time use a rolling window of 50 months. Once again, we report both average MSE and MAE prediction error, as forecast errors are non-gaussian. The table with forecasting performance is in Table 7 while the plot of squared prediction error is in Figure 12. We also include figures of squared prediction error for pre-August 2008 and post January 2009 in the Appendix (Figure 14), and forecasting results only post-2009 (Table 14).

Once again, when compared via MSE, most models do not make much of an improvement over the benchmark, if any at all. The FVOL MKT model performs slightly better at horizon 1, though worse at all other horizons. FVOL PCA performs best at horizon 2 and 3, but overall they both underperform the benchmark.

On the other hand, when compared via average MAE, the factor in idiosyncratic volatility

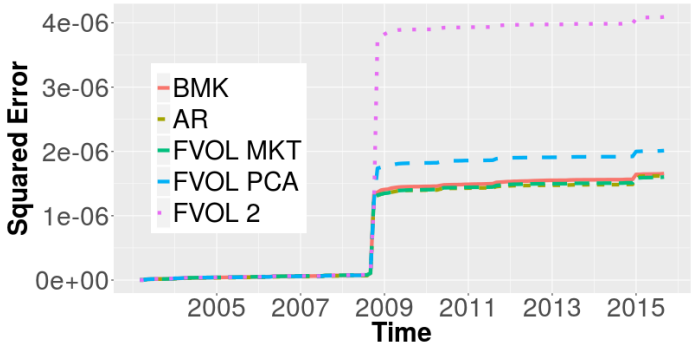
Table 7: Average Mean Square Error and Median Absolute Error of FX rate Rvariances

All values are relative to BMK forecasts. Bolded value in each row is the minimum, when better than BMK. BMK is benchmark, AR is with univariate autoregressive idiosyncratic volatility, FVOL MKT uses market volatility as a single idiosyncratic vol factor, FVOL PCA uses a single principal component as an idiosyncratic vol factor, FVOL 2 uses both. For all models, we use a 50-month rolling window where we estimate the model in every window and then forecast for 1-12 months ahead.

	Average MSE				Average MAE			
h	AR	FVOL MKT	FVOL PCA	FVOL2	AR	FVOL MKT	FVOL PCA	FVOL2
1	0.98	**0.97**	1.22	2.47	0.83	0.93	**0.81**	0.83
2	1.13	1.01	**0.98**	1.18	0.86	0.96	**0.85**	0.88
3	1.16	1.02	**0.98**	1.69	0.91	1.00	**0.88**	0.92
4	1.10	1.03	1.17	2.47	0.94	0.98	**0.87**	0.94
5	1.08	1.03	37.84	2.40	0.96	1.00	**0.92**	0.96
6	1.11	1.06	4.44	15.41	0.99	1.02	**0.95**	1.01
7	1.06	1.11	1.42	25.39	0.97	0.99	**0.95**	1.00
8	1.03	1.03	4.20	1.53	0.98	1.01	**0.94**	1.03
9	1.02	1.05	1.60	1.21	0.97	1.00	**0.95**	0.99
10	1.03	1.04	1.32	1.19	0.97	1.01	**0.95**	1.01
11	1.03	1.02	1.01	1.06	0.95	0.97	**0.88**	0.96
12	1.04	1.02	1.00	1.02	0.98	0.99	**0.92**	0.97

Figure 12: FX squared One-Step prediction errors

Cumulative squared errors over time, 2007-2015, of the panel of exchange rate volatilities. Each date adds the average squared distance of true volatility to forecasted volatility over the panel. The models perform similarly outside the financial crisis 2008-2010, but there the discrepancies are large.



has a large impact on improving forecasts. All FVOL models perform much better (10-20%) than the benchmark, especially at short horizons. Similar to equities, the FVOL PCA model performs best at all horizons.

1.7 Conclusion

We have revisited the standard factor model, and its use in facilitating tractable dynamic volatility. We have shown that Σ_{e_t} is correlated with Σ_{F_t}, but that Σ_{F_t} alone is not sufficient for explaining time-variation in idiosyncratic volatility. This suggests that the classic decomposition is ultimately not an optimal approach to modeling time-varying volatility. Furthermore, one might conclude that if modeling panels of volatilities, and not covariances, is the practitioner's goal, then one should fit factor models to panels of volatilities directly. This result holds across a wide variety of asset classes and time frequencies.

We briefly explored the implications of these results for forecasting, but much remains to be done. In particular, do these heirarchical factor structures help in constructing density forecasts for returns? Are these risk factors for idiosyncratic volatility priced? Our preliminary evidence on both questions suggest negative results, but these results could be sensitive to the time horizon of the sample, the specific equity market, or even the industry.

Second, the presence of this heirarchical structure in both equities and FX data suggests it may be a more general feature of volatility. It remains to be argued why the nature of panels of volatility should lend themselves to such heirarchical structures, whether through network effects or an endogenous economic mechanism. Indeed, due to the fact that FX rates and equities are entirely different asset classes, the empirical phenomenon may be more of a statistical phenomenon (such as factor structure) than one that is driven by structural theory. It also remains to be shown whether this feature appears in other panels of volatilities, for example in the volatility of large macroeconomic panels. Finally, our framework here did not accurately account for measurement error in the panels of volatilities. Using frontier theory on the distribution of realized volatility estimators one can extend this work to account for measurement error, and this represents an avenue for future contributions.

1.8 Appendix

1.8.1 Simulation

In this section we confirm the appropriateness of our battery of statistical tests. There are several issues to consider that may warrant skepticism of their use in our environment: (1) Our observed factor volatility (market volatility) is actually observed with measurement error (as it is a realized measure), (2) our panel of interest itself is observed with measurement error (realized measures of idiosyncratic volatility), and (3) our models contain correlated regressors (as the market volatility factor is correlated with the first Principal Component of idiosyncratic volatility).

To assuage our concerns with all three issues, we conduct the following simulation. We generate output using Models 1, 2, and 3 as the data generating processes, for the cases of $N = 10, 100, 200$, $T = 500, 2000$, and intraday observations of 100 and 1000. The log-market volatility is generated as an AR(1) process with AR parameter 0.9 and mean -9. The factor structure (whether Model 1, 2, or 3) is defined in terms of log-volatlities. All factor loadings (for all possible factors) are distributed as absolute value of normals with mean zero and standard deviation 0.5. In Model 2, the PCA factor is generated as the market (log) volatility plus classical measurement error with variance calibrated so that the PCA factor is 75% correlated with the market volatility. Intraday observations are taken as iid draws from a normal distribution with mean 0 and variance the true volatility. Realized volatilties are calculated as in Barndorff-Nielsen and Shephard (2004), the outer product of high-frequency returns. While we acknowledge that the high-frequency generation process is simplistic (and unrealistic), note that the most important object is the signal-to-noise ratio between true and realized volatility. With a more complex DGP, one should use a more sophisticated estimation procedure to maintain a similar amount of information. Factor loadings vary every day as iid noise centered around constant loadings.

We then conduct our battery of tests on each set of data generated for $1,000$ simulations,

and determine if the tests have correct size and power for the respective data generating processes and null hypotheses. The results are very promising and presented in Table 8. Recall Table 1. We expect the LR-1 test to have appropriate power, rejecting the null in all cases[2]. In Model 3, LR-1 has appropriate power even against a correlated regressor, as we regress out σ_{F_t} first. By contrast, LR-2 will under-reject Models 1 and 3, as it is facing an alternative of correlated regressors (similar to a t-test in a simple regression setup). We see this in practice. The approximate Bai and Ng tests behave as expected. Notably, these tests are reasonably robust to measurement error both in the panel and in the observed factor for volatility: In the case of 100 intraday observations, the measurement error volatility in idiosyncratic volatility is 5% of the volatility in the panel, and the tests behave as expected.

Measurement Error And Simulation Results

The exact Bai and Ng tests, as well as the Onatski tests, do not behave as desired in a high frequency simulation setting. We note in particular that when Model 1 is the null, both of these tests strongly over-reject. This suggests that our preference for Models 2 and 3 in the empirical results should potentially be taken with a grain of salt. In this section we explore the role that measurement error in the realized measures of market volatility and idiosyncratic volatility can play in explaining these results.

Recall the construction of realized measures, and suppose we are trying to select between Models 1, 2, and 3. Further suppose we measure idiosyncratic volatility accurately via a direct method. For example, with a high number of intraday observations, our measurement of realized beta will be accurate, so we may construct high frequency idiosyncratic returns directly, resulting in more classical measurement error in idiosyncratic volatilities. We still estimate the factor structure by estimating the regression

$$RIV_{it} = \mu_i + \beta_i RV_{f_t} + u_t^i \tag{1.8.1}$$

[2]Note that even in the case of Model 1, LR-1 should reject $\beta = 0$ since the PCA factor should be the same as the market volatility.

If $RV_{f_t} = \sigma^2_{F_t} + \epsilon_{F_t}$, then the parameter estimate $\widehat{\beta}_i$ will be biased downward relative to the true regression coefficient between σ^2_{it} and $\sigma^2_{F_t}$ due to attenuation bias from measurement error. The result is $\widehat{u}^i_t$ will exhibit factor structure regardless of the nature of u^i_t. Note that in practice error in market volatility realized estimation will be correlated with error in idiosyncratic volatility realized estimation, which will reduce the magnitude of this problem — having correlated errors on LHS and RHS diminishes the impact of attenuation bias from RHS measurement error[3].

We can see this in practice by considering the results of our Onatski test: because the test statistic is constructed by regressing out market volatility and exploring the factor structure of the remaining residuals, it is subject to the above error. As a result, it over-rejects. We can find confirming evidence for this story by running the simulation using true market volatility in place of a realized estimate in the estimation of Equation 1.8.1. When we do this we find that Onatski rejects with the correct rate. We also consider alternative explanations of the phenomenon by running the simulation with different measurement error specifications — in particular we find that classical measurement error on the true market volatility still induces Onatski to over-reject. Thus the over-rejection is simply a matter of having positive measurement error at all, rather than depending on the exact nature of error in the realized estimator.

[3]Consider regressing y against x when we observe $\tilde{y} = y + \varepsilon$ and $\tilde{x} = x + v$. Then

$$y = \beta x + u$$
$$\tilde{y} = \beta \tilde{x} + u - \beta v + \varepsilon$$

Thus a positive correlation between v and ε means the bias in β above is smaller than the bias in the case when $\varepsilon = 0$.

Table 8: Simulation Results

Results of 1000 replications of each model. Columns are labelled M-1, M-2, and M-3 corresponding to Models 1, 2, and 3 respectively. Column and row segments are labelled based on corresponding dimensions N and T and number of intraday observations. Tests LR-1, LR-2 and Onatski report empirical size of 95% cutoff values. Tests A, M, NS and R2 (those from Bai and Ng (2006)) report average values across simulations.

			N = 10			N = 100			N = 200	
			M-1	M-2	M-3	M-1	M-2	M-3	M-1	M-2	M-3
Intraday = 100	T = 500	LR-1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		LR-2	0.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00
		Onatski	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		A	0.64	0.95	0.91	0.88	0.97	0.97	0.91	0.98	0.98
		M	24.40	62.80	115.29	43.57	53.73	143.83	61.32	69.46	189.80
		NS	0.06	8.72	1.09	0.06	3.16	1.06	0.06	3.20	1.06
		R2	0.97	0.81	0.59	0.97	0.69	0.59	0.97	0.68	0.59
	T = 2000	LR-1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		LR-2	0.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00
		Onatski	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		A	0.62	0.94	0.91	0.86	0.97	0.96	0.90	0.97	0.97
		M	29.65	63.73	142.82	48.26	49.97	156.60	65.05	62.25	203.59
		NS	0.05	5.22	1.01	0.05	2.11	0.98	0.05	1.81	0.98
		R2	0.97	0.81	0.57	0.97	0.63	0.57	0.97	0.61	0.57
Intraday = 1000	T = 500	LR-1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		LR-2	0.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00
		Onatski	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		A	0.63	0.99	0.97	0.87	0.99	0.99	0.91	1.00	1.00
		M	23.07	55.09	348.38	41.32	49.87	422.75	58.08	62.42	573.11
		NS	0.01	7.23	0.97	0.01	3.38	0.97	0.01	2.76	0.98
		R2	1.00	0.84	0.61	1.00	0.70	0.61	1.00	0.70	0.61
	T = 2000	LR-1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		LR-2	0.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00
		Onatski	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		A	0.62	0.98	0.97	0.86	0.99	0.99	0.90	0.99	0.99
		M	27.92	54.86	443.58	45.58	45.09	467.36	62.37	58.50	614.67
		NS	0.00	5.16	0.92	0.00	1.95	0.90	0.00	1.81	0.90
		R2	1.00	0.83	0.59	1.00	0.65	0.59	1.00	0.62	0.60

1.8.2 Data Lists

In this section, we present each of the three datasets used in the paper, as well as data descriptions. For the equity datasets we provide ticker and company name, and for the S&P 100 dataset we also present the sector. For the FX dataset we provide data label from FRED as well as the currencies.

- Table 9 - List of companies used for DOW 10 analysis.
- Table 10 - List of companies (and sectors) used for S&P 100 analysis.
- Table 11 - List of currencies used for FX rate analysis.

Table 9: DOW 10 Company List

Ticker	Name
AA	Alcoa Inc
AXP	American Express
BAC	Bank of America
DD	Du Pont
GE	General Electric
IBM	International Business Machines
JPM	JPMorgan Chase
KO	Coca-Cola
MSFT	Microsoft
XOM	Exxon Mobil

Table 10: S&P100 Company List

Ticker	Name	Sector	Ticker	Name	Sector
AAPL	Apple Inc.	Info Tech	IBM	Intl Business Machines Corp	Info Tech
ABT	Abbott Laboratories	Health Care	INTC	Intel Corp	Info Tech
ACN	Accenture plc	Info Tech	JNJ	Johnson & Johnson	Health Care
AIG	American International Group	Financials	JPM	JP Morgan Chase & Co	Financials
ALL	Allstate Corp	Financials	KO	Coca-Cola Co	Cons. Staples
AMGN	Amgen Inc	Health Care	LLY	"Lilly	Health Care
AMZN	Amazon.com Inc	Cons. Discret.	LMT	Lockheed Martin	Industrials
APA	Apache Corp	Energy	LOW	Lowe's Cos Inc	Cons. Discret.
APC	Anadarko Petroleum Corp	Energy	MA	Mastercard Inc A	Info Tech
BA	Boeing Co	Industrials	MCD	McDonald's Corp	Cons. Discret.
BAC	Bank of America Corp	Financials	MDT	Medtronic Inc	Health Care
BAX	Baxter Intl Inc	Health Care	MET	Metlife Inc	Financials
BHI	Baker Hughes Inc	Energy	MMM	3M Co	Industrials
BK	The Bank of New York Mellon Corp	Financials	MO	Altria Group Inc	Cons. Staples
BMY	Bristol-Myers Squibb	Health Care	MON	Monsanto Co.	Materials
C	Citigroup Inc	Financials	MRK	Merck & Co Inc	Health Care
CAT	Caterpillar Inc	Industrials	MS	Morgan Stanley	Financials
CL	Colgate-Palmolive Co	Cons. Staples	MSFT	Microsoft Corp	Info Tech
CMCSA	Comcast Corp	Cons. Discret.	NKE	NIKE Inc B	Cons. Discret.
COF	Capital One Financial	Financials	NOV	National Oilwell Varco Inc	Energy
COP	ConocoPhillips	Energy	NSC	Norfolk Southern Corp	Industrials
COST	Costco Wholesale Corp	Cons. Staples	ORCL	Oracle Corp	Info Tech
CSCO	Cisco Systems Inc	Info Tech	OXY	Occidental Petroleum	Energy
CVS	CVS Caremark Corp.	Cons. Staples	PEP	PepsiCo Inc	Cons. Staples
CVX	Chevron Corp	Energy	PFE	Pfizer Inc	Health Care
DD	"DuPont	Materials	PG	Procter & Gamble	Cons. Staples
DIS	Walt Disney Co	Cons. Discret.	QCOM	QUALCOMM Inc	Info Tech
DOW	Dow Chemical	Materials	RTN	Raytheon Co	Industrials
DVN	Devon Energy Corp	Energy	SBUX	Starbucks Corp	Cons. Discret.
EBAY	eBay Inc.	Info Tech	SLB	Schlumberger Ltd	Energy
EMC	EMC Corp	Info Tech	SO	Southern Co	Utilities
EMR	Emerson Electric Co	Industrials	SPG	Simon Property Group	Financials
EXC	Exelon Corp	Utilities	T	AT&T Inc	Telecom Services
F	Ford Motor Co	Cons. Discret.	TGT	Target Corp	Cons. Discret.
FCX	Freeport McMoRan Copper & Gold	Materials	TWX	Time Warner Inc	Cons. Discret.
FDX	FedEx Corp	Industrials	TXN	Texas Instruments Inc	Info Tech
GD	General Dynamics	Industrials	UNH	Unitedhealth Group Inc	Health Care
GE	General Electric Co	Industrials	UNP	Union Pacific Corp	Industrials
GILD	Gilead Sciences Inc	Health Care	UPS	United Parcel Service Inc B	Industrials
GOOG	Google Inc	Info Tech	USB	US Bancorp	Financials
GS	Goldman Sachs Group Inc	Financials	UTX	United Technologies Corp	Industrials
HAL	Halliburton Co	Energy	VZ	Verizon Communications Inc	Telecom Services
HD	Home Depot Inc	Cons. Discret.	WFC	Wells Fargo & Co	Financials
HON	Honeywell Intl Inc	Industrials	WMT	Wal-Mart Stores	Cons. Staples
HPQ	Hewlett-Packard Co	Info Tech	XOM	Exxon Mobil Corp	Energy

Table 11: Forex List

FRED Label	Currency
exalus	Australia / US
exbzus	Brazil / US
excaus	Canada / US
exdnus	Denmark / US
exjpus	Japan / US
exkous	South Korea / US
exmxus	Mexico / US
exnzus	New Zealand / US
exnous	Norway / US
exsius	Singapore / US
exsfus	South Africa / US
exszus	Switzerland / US
exukus	UK / US
exeuus	EU / US

1.8.3 Forecasting Tables and Figures

In this Appendix we present extra tables and figures from the forecasting exercises.

- Table 12 - DOW 10 forecasting results (average MSE and MAE) using forecasts only after 2009.
- Table 13 - S&P 100 forecasting results (average MSE and MAE) using forecasts only after 2009.
- Table 14 - FX rate forecasting results (average MSE and MAE) using forecasts only after 2009.
- Figure 13 - Plot of squared forecast errors for both equity datasets, post 2009.
- Figure 14 - Plot of squared forecast errors for FX dataset, pre-2008 and post-2009.

Table 12: Average Mean Square Error and Median Absolute Error of DOW 10 Rvariances (post 2009)

All values are relative to BMK forecasts. Bolded value in each row is the minimum, when better than BMK. BMK is benchmark, AR is with univariate autoregressive idiosyncratic volatility, FVOL MKT uses market volatility as a single idiosyncratic vol factor, FVOL PCA uses a single principal component as an idiosyncratic vol factor, FVOL 2 uses both. All models use a 200-day rolling window to estimate parameters, followed by forecasts for 1-12 days ahead. Table presents only forecast errors from predictions after 2009.

	Average MSE				Average MAE			
h	AR	FVOL MKT	FVOL PCA	FVOL2	AR	FVOL MKT	FVOL PCA	FVOL2
1	**0.70**	0.91	0.88	0.87	0.83	0.88	**0.80**	0.81
2	**0.75**	0.93	0.93	0.89	0.87	0.90	**0.82**	0.83
3	**0.81**	0.92	0.94	0.89	0.89	0.90	**0.84**	0.86
4	**0.86**	0.93	0.94	0.90	0.90	0.90	**0.83**	0.86
5	**0.87**	0.92	0.94	0.90	0.92	0.92	**0.85**	0.88
6	**0.86**	0.93	0.94	0.91	0.92	0.92	**0.86**	0.90
7	**0.85**	0.92	0.93	0.90	0.93	0.93	**0.87**	0.90
8	**0.87**	0.93	0.93	0.91	0.93	0.93	**0.85**	0.90
9	**0.86**	0.94	0.94	0.91	0.94	0.93	**0.86**	0.90
10	**0.83**	0.93	0.93	0.91	0.94	0.94	**0.87**	0.90
11	**0.85**	0.94	0.94	0.92	0.94	0.95	**0.88**	0.92
12	**0.85**	0.94	0.93	0.93	0.95	0.93	**0.87**	0.93

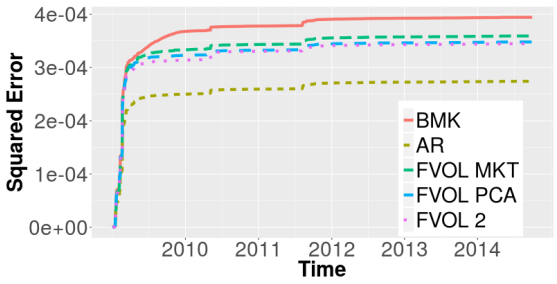
Table 13: Average Mean Square Error and Median Absolute Error of S&P 100 Rvariances (post 2009)

All values are relative to BMK forecasts. Bolded value in each row is the minimum, when better than BMK. BMK is benchmark, AR is with univariate autoregressive idiosyncratic volatility, FVOL MKT uses market volatility as a single idiosyncratic vol factor, FVOL PCA uses a single principal component as an idiosyncratic vol factor, FVOL 2 uses both. All models use a 200-day rolling window to estimate parameters, followed by forecasts for 1-12 days ahead. Table presents only forecast errors from predictions after 2009.

	Average MSE				Average MAE			
h	AR	FVOL MKT	FVOL PCA	FVOL2	AR	FVOL MKT	FVOL PCA	FVOL2
1	**0.75**	0.85	0.87	0.98	0.68	0.78	0.67	**0.65**
2	**0.83**	0.86	0.87	0.93	0.74	0.81	0.71	**0.70**
3	0.88	**0.87**	0.88	0.94	0.77	0.82	**0.72**	0.72
4	0.97	0.87	**0.86**	0.92	0.79	0.83	**0.72**	0.73
5	1.03	**0.85**	0.86	0.90	0.81	0.84	**0.74**	0.75
6	1.19	**0.85**	0.86	0.90	0.82	0.85	**0.75**	0.76
7	1.67	**0.86**	0.86	0.91	0.84	0.86	**0.76**	0.78
8	1.59	0.86	**0.86**	0.90	0.85	0.87	**0.77**	0.81
9	1.89	0.86	**0.85**	0.91	0.86	0.87	**0.77**	0.80
10	2.07	0.88	**0.86**	0.93	0.87	0.88	**0.77**	0.80
11	1.99	0.88	**0.85**	0.95	0.87	0.89	**0.78**	0.83
12	1.81	0.89	**0.86**	0.95	0.88	0.89	**0.79**	0.84

Figure 13: Equities cumulative squared forecast errors - Post 2009

(a) DOW 10



(b) S&P 100

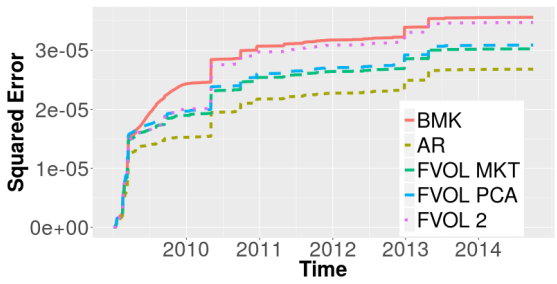
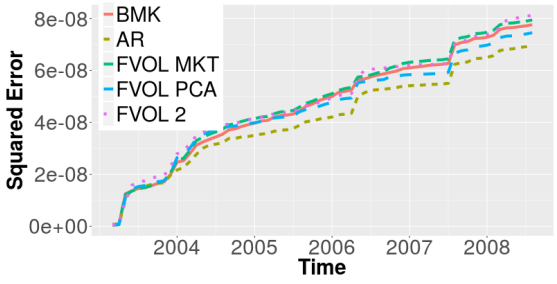


Figure 14: FX cumulative squared forecast errors

(a) Pre- August 2008



(b) Post- 2009

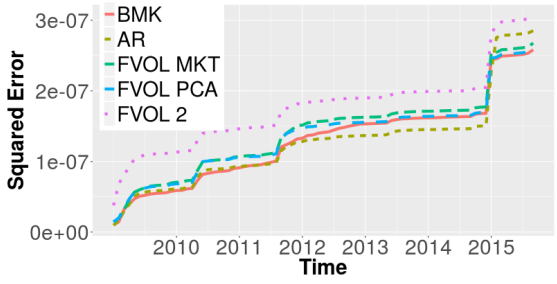


Table 14: Average Mean Square Error and Median Absolute Error of FX rate Rvariances (post 2009)

All values are relative to BMK forecasts. Bolded value in each row is the minimum, when better than BMK. BMK is benchmark, AR is with univariate autoregressive idiosyncratic volatility, FVOL MKT uses market volatility as a single idiosyncratic vol factor, FVOL PCA uses a single principal component as an idiosyncratic vol factor, FVOL 2 uses both. For all models, we use a 50-month rolling window where we estimate the model in every window and then forecast for 1-12 months ahead. Table presents only forecast errors from predictions after 2009.

	Average MSE				Average MAE			
h	AR	FVOL MKT	FVOL PCA	FVOL2	AR	FVOL MKT	FVOL PCA	FVOL2
1	1.10	1.04	1.00	1.19	0.75	0.83	**0.72**	0.75
2	1.06	1.15	1.00	1.29	0.81	0.91	**0.77**	0.83
3	1.12	1.19	**0.96**	1.73	0.92	0.99	**0.83**	0.94
4	1.72	1.19	**0.94**	3.10	0.93	1.03	**0.86**	0.97
5	1.52	1.23	211.40	3.37	0.98	1.04	**0.89**	0.99
6	1.76	1.35	16.09	8.90	1.01	1.06	**0.94**	1.01
7	1.36	1.47	3.33	14.53	0.98	1.03	**0.91**	1.02
8	1.23	1.21	23.83	4.25	0.99	1.05	**0.94**	1.05
9	1.12	1.30	5.26	2.28	0.96	1.02	**0.96**	1.03
10	1.17	1.22	3.33	2.07	0.96	0.98	**0.92**	1.03
11	1.21	1.08	**0.99**	1.36	0.96	0.97	**0.85**	0.98
12	1.23	1.09	**0.94**	1.07	0.95	0.98	**0.86**	0.95

CHAPTER 2

Separating Variances and Correlation; A New Prior for TVP-VARs

2.1 Introduction

Time-Varying Parameter (TVP) Vector Autoregressions are highly parameterized models that have become increasingly popular both from the structural perspective of understanding how the macroeconomy has changed over time, and more recently, from a forecasting perspective.

The literature began with frequentist modeling ala Nyblom (1989) and Stock and Watson (1996a)'s Median Unbiased Estimators. However, the frequentist strategies were computationally intractable, especially as the literature moved to multivariate models. Benati (n.d.) extended the MUB framework to a two dimensional VAR, but could not incorporate heteroskedasticity.

Bayesian TVP models have a long history, starting with Doan et al. (1983) and Sims (1993) who used them at the Minneapolis Federal Reserve for forecasting. Those early models usually made simplifying assumptions, such as only autoregressive coefficients could vary, while all others were fixed. They became more popular when Cogley and Sargent (2002), Cogley and Sargent (2005) and Primiceri (2005) began using models where all parameters were free to vary for retrospective analyses, answering question such as: "How has the economy changed over long periods of time?" While Bayesian TVP-VARs could handle large multivariate models in theory, in practice they were often confined to small systems due to the curse of dimensionality. If the covariance matrix of the parameters is fully dense, then

a VAR with K variables and p lags has $pK^2 + K$ total time-varying parameters, which are driven by a $(pK^2+K) \times (pK^2+K)$ covariance matrix. Thus, even a three-dimensional VAR with two lags would have 21 moving parameters and 210 unique elements of the covariance matrix. As such, the classical papers studied small systems with only a few variables and focused on measuring how the economy was changing by looking at growth rates or impulse response functions at different times.

Recently, TVP models have been adapted to big-data techniques. Stevanovic (2010) applies, and finds, factor structure in the parameters, effectively reducing the number of parameters in the covariance matrix. Stevanovic and Amir-Ahmadi (2015) apply the technique to a 5-variable model with great success. Amisano et al. (2015) creates a TVP model where the covariance matrix has a special kronecker product structure that enables application to higher dimensions (possibly 20 or more variable VAR).

Other methods seek to keep the Bayesian estimation paradigm, but avoid MCMC procedures due to their intense computational needs. Koop and Korobilis (2013) introduce a method called variance discounting, which generates a recursive form for the parameter covariance matrix, which makes filtering trivial. Pettenuzzo et al. (2016) combine variance discounting with a shrinkage method called Compression to estimate and forecast with a 129-variable TVP-VAR with lag length 13.

When it comes to modeling procedures though, all TVP-VAR models are based on the fundamentals from the classical papers (i.e. Primiceri (2005) and others listed above) — especially as related to the choice of priors. All papers select Inverse Wishart distributions as the prior (and posterior) over all covariances. Marginally, that means they select Inverse-Gamma priors on variances. While Inverse-Wishart distributions are a conjugate prior with a Gaussian likelihood, there are a number of negative properties about the distribution, which make them a poor choice for a TVP context. These negative properties will be discussed in more detail in Section 2.2.1, but briefly, they are (1) They give no weight to zero variances, (2) There is only one degree of freedom, (3) For low degrees of freedom,

there is strong comovement between variances and correlations, and (4) They must specify a location/center matrix.

However, TVP models are very similar to Bayesian Random Effects Models/Hierarchical Modeling. Indeed, as Lindley and Smith (1972) showed, Inverse-Wishart distributions can be used as a conjugate prior for estimating an entire covariance matrix in a Hierarchical Model. The connection between Random Effects and TVP models means that innovative priors in the Random Effects models might be helpful for TVP models, and vice versa. For example, Gelman (2006) shows that in the context of Hierarchical Models, Inverse-Gamma distributions can be problematic as an uninformative prior because they place no prior mass on very small values of variance. Instead, they propose using a half-t or half-Cauchy prior. Frühwirth-Schnatter and Wagner (2010) use this observation to show that when considering a model selection problem in TVP regression, using a prior that bounds observations away from 0 will obviously skew variance estimates toward time-variation, even if none exists. As an alternative, they suggest a half-normal prior. Continuing this research, Frühwirth-Schnatter and Bitto (2016) and Belmonte et al. (2014) use various shrinkage methods to shrink both time-varying variances and initial values towards zero. This principle of shrinkage toward constant parameters is especially applicable to the TVP-VAR literature since practitioners are often concerned that the model will overestimate time variation.

Another problem with Inverse Wishart priors is that there is only one degree of freedom parameter for both correlation and variance. Thus, even if one wishes to be informative about variance, but uninformative about correlation (or informative about one variance, but not another), that is impossible with the IW prior. Again in the context of Hierarchical Modeling, Barnard et al. (2000) proposed a solution to this issue by combining independent priors for each into a single covariance matrix. While that prior is fully flexible and easily interpretable, it is computationally intractable. O'Malley and Zaslavsky (2008) relaxed the requirement of full independence and alleviated some of the computational issues by changing the variance prior to one that can be updated by a Metropolis Hastings step. Huang and

Wand (2013) took this one step further by introducing a conjugate prior for variances, so using this prior involves merely an additional draw from a Gamma distribution.

In this paper, I take the prior of Huang and Wand (2013) and further adapt it into an informative prior. This new prior includes 0, and therefore can be thought of as a shrinkage prior for TVP-VARs. It therefore mimics the goals of Frühwirth-Schnatter and Bitto (2016) and Belmonte et al. (2014). Mine differs from their work in that I estimate a fully dense covariance matrix, whereas they assume it is diagonal. My method is conceptually similar to contemporaneous work by Eisenstat et al. (2014), but they are more interested in model selection and use different priors. Like them, I find that the new prior makes a large difference when considering a simulated constant parameter model, especially if one may want to use a threshold rule to set small variances to 0.

In addition to the improved statistical properties, my new prior is also easily interpretable. Statements such as, "My prior is that standard deviations are less than 1 with 90% probability," can be mapped directly into specific hyperparameters via a quantile function.

In Section 2.2, I introduce the model, display the negative properties of IW distributions, and introduce uninformative priors from Hierarchical Modeling. In Section 2.3, I adapt the prior of Huang and Wand (2013) into an informative prior, and introduce two methods of eliciting non-sample information for a practitioner. In Section 2.4, I devise an elaborate simulation study comparing the new priors with their IW counterparts. In general, the simulations show that my priors are better able to estimate the time-varying parameters, but only sometimes improve estimation of the error covariance matrix. This section shows that in addition to being more interpretable and natural priors, my priors also exhibit improved frequentist properties. In Section 2.5, I deal with the model selection issue via two information criteria. In Section 2.6, I apply my prior to the canonical Primiceri (2005) dataset and find differences in impulse response functions and forecasting improvements. I also perform a forecasting exercise with a dataset from Pettenuzzo et al. (2016), but find that all the TVP models do not substantially improve forecasts. Lastly, Section 2.7 concludes.

2.2 Model Setup

The basic setup of my model is the same as the canonical papers in the literature, such as Primiceri (2005), Cogley and Sargent (2002), and Cogley and Sargent (2005). I mostly follow the notation of Primiceri (2005) when applicable. While I acknowledge that the literature has moved towards larger TVP-VARs and developing new techniques to deal with the huge datasets, the newer models are still based on classical priors, which my method improves.

Let $\{y_t\}_{t=1}^T$ be a $K \times 1$ vector of time series observations, where K is the number of objects to predict, and is greater than or equal to one. I consider general Time-Varying Parameter models, given by regressors X_t, where X_t is $J \times 1$. Thus,

$$y_t = X_t' B_t + \varepsilon_t, \qquad t = 1, \dots T.$$

The main application will be to AR and VAR models, in which case

$$X_t = I_n \otimes [1, y_{t-1}', \dots, y_{t-p}'].$$

Let $\boldsymbol{\beta_t} = vec(B_t)$, then the time-varying βs are assumed to follow a random walk, so

$$\boldsymbol{\beta_t} = \boldsymbol{\beta_{t-1}} + \varepsilon_t^{\beta}$$

Primiceri (2005) use the TVP setup to find changes in impulse response functions, and therefore also require a time-varying structural matrix, they therefore use

$$\varepsilon_t \sim N(0, \Omega_t) \qquad t = 1, \dots T$$
$$\Omega_t = A_t^{-1} \Sigma_t \Sigma_t' A_t^{-1\prime}$$

where A_t is a lower triangular matrix of structural coefficients,

$$A_t = \begin{pmatrix} 1 & 0 & \dots & 0 \\ \alpha_{21,t} & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ \alpha_{K1,t} & \dots & \alpha_{KK-1,t} & 1 \end{pmatrix},$$

and Σ_t is a diagonal matrix of standard deviations:

$$\Sigma_t = \begin{pmatrix} \sigma_{1,t} & 0 & \dots & 0 \\ 0 & \sigma_{2,t} & \ddots & 0 \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \sigma_{n,t} \end{pmatrix}.$$

In practice, the time-varying betas are usually the hardest element to estimate, so for simplicity, I fix the structural matrix to be constant over time:

$$\Omega_t = A^{-1}\Sigma_t\Sigma_t' A^{-1\prime}.$$

In order to calculate the time-invariant covariance, I standardize the residuals ($\widehat{\varepsilon}_t$) by $\Sigma_t^{-1/2}$ (so $\widetilde{\varepsilon}_t = \widehat{\varepsilon}_t\Sigma_t^{-1/2}$) and say $\widetilde{\varepsilon}_t \sim N(0, H)$. In practice, since variances are not identified in this specification, H becomes an approximate correlation matrix. I also consider homoskedastic models, despite the known presence of stochastic volatility in many macro variables. In a comment to Cogley and Sargent (2002), Sims (2001) indicated that the amount of time-variation in parameters will be overestimated if the true DGP includes stochastic volatility, but the model does not. Cogley and Sargent (2005) subsequently allowed for stochastic volatility, which I also include.

The standard deviations of the error matrix are also assumed to follow a random walk over

time:

$$\log(\sigma_{i,t}) = \log(\sigma_{i,t-1}) + \varepsilon^{\sigma}_{i,t} \quad i = 1, \dots K.$$

Throughout, the random-walk assumption is made for computational reasons and is standard in the TVP literature.

The errors of the time-varying processes are all normally distributed, with a constant covariance:

$$\begin{pmatrix} \varepsilon^{\beta}_t \\ \varepsilon^{\sigma}_t \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} Q_\beta & 0 \\ 0 & Q_s \end{pmatrix} \right).$$

In the Primiceri (2005) setup, each class of variable (beta, stochastic volatility and structural parameter) errors have dense covariance, but are uncorrelated across classes. Thus, in my setup, Q_β and Q_s are dense covariance matrices. Cogley and Sargent (2002), and Cogley and Sargent (2005) also use this same setup, though others make the simplifying assumption that the parameters should be independent even within parameter type. The dense covariance matrix is also supported by the factor structure on TVP parameters that Stevanovic (2010) and Stevanovic and Amir-Ahmadi (2015) find.

2.2.1 Priors and Model Estimation

To summarize, the TVP setup can be put into state-space form as:

$$\begin{aligned} y_t &= X_t' B_t + \varepsilon_t & \varepsilon_t &\sim N(0, \Omega_t), \\ vec(\beta_t) &= vec(\beta_{t-1}) + \varepsilon^{\beta}_t & \varepsilon^{\beta}_t &\sim N(0, Q_\beta), \\ \log(diag(\Sigma_t)) &= \log(diag(\Sigma_{t-1})) + \varepsilon^{\sigma}_t & \varepsilon^{\eta}_t &\sim N(0, Q_s). \end{aligned}$$

The linear-gaussian state-space form immediately lends itself to Kalman Filtering and smooth-

ing, which is the main method for estimation. While Stock and Watson (1996b) and Benati (n.d.) offer Maximum Likelihood procedures, most of the literature follows Primiceri (2005) and Cogley and Sargent (2005) by performing Bayesian estimation. Bayesian estimation requires a prior, which is combined with the likelihood to obtain the posterior. The class of distributions used as priors are standard in the literature, though for expository purposes I will use the hyperparameters in Primiceri (2005). In all cases they use conjugate priors, which allows for efficient Gibbs Sampling from the posterior distribution. The priors are as follows:

$$\begin{aligned}
\boldsymbol{\beta_0} &\sim N\left(\widehat{\boldsymbol{\beta_{OLS}}}, 4 \cdot V(\widehat{\boldsymbol{\beta_{OLS}}})\right), \\
\log(\sigma_0) &\sim N\left(\log(\widehat{\sigma_{OLS}}), I_K\right), \\
Q_\beta &\sim IW\left(\tau, \kappa_Q^2 \cdot \tau \cdot Q_0\right), \\
Q_s &\sim IW(K+1, I_K \times 10^{-3}).
\end{aligned}$$

Where κ_Q and Q_0 are hyperparameters that will be discussed more later, τ is a burn-in sample size (40), $\widehat{\boldsymbol{\beta_{OLS}}}$ is the vectorized OLS regression estimate, and $V(\widehat{\boldsymbol{\beta_{OLS}}})$ is the covariance of OLS estimators. Both $\widehat{\boldsymbol{\beta_{OLS}}}$ and $V(\widehat{\boldsymbol{\beta_{OLS}}})$ are estimated based on the burn-in sample.

The prior that is most problematic is the Inverse-Wishart prior over the error covariance of the βs, Q_β. Before discussing each of the issues in detail, it is helpful to review some facts about the Inverse Wishart distribution. All these properties can be found in any prominent Multivariate Analysis textbook (such as Marden (2015)) or Wikipedia.

Fact 1. *If a K dimensional positive definite matrix Q has Inverse Wishart distribution with degrees of freedom $\nu > K+1$ and scale matrix Q_0, then we write $Q \sim IW(\nu, Q_0)$ and it has*

distribution function given as:

$$f(Q) = \frac{|Q_0|^{\nu/2}}{2^{\frac{\nu K}{2}}\Gamma_K(\nu/2)}\,|Q_0|^{-\frac{\nu+p+1}{2}}\,e^{-\frac{1}{2}trace(Q_0 Q)}$$

Property 1. *Let $Q^\star$ and $Q_0^\star$ be a conformable partition of Q and Q_0 respectively, such that:*

$$Q = \begin{pmatrix} Q_{11}^\star & Q_{12}^\star \\ Q_{21}^\star & Q_{22}^\star \end{pmatrix} \quad Q_0 = \begin{pmatrix} Q_{0_{11}}^\star & Q_{0_{12}}^\star \\ Q_{0_{21}}^\star & Q_{0_{22}}^\star \end{pmatrix}$$

where Q_{11} and $Q_{0_{11}}$ are dimension $d_1 < K$, and Q_{22} and $Q_{0_{22}}$ are dimension $d_2 = K - d_1$. If $Q \sim IW(\nu, Q_0)$, then $Q_{11} \sim IW(\nu - d_2, Q_{0_{11}})$.

Proof. A Wishart matrix is defined as the sum of outer-product of a multivariate normal vector. So simply partition the vector and take outer products of the partitioned vector. For more details, see Marden (2015) or Wikipedia. ■

Property 2. *Let Q and Q_0 be K dimensional positive definite matrices and $Q \sim IW(\nu, Q_0)$. If $K = 1$, then $Q \sim IG(\nu/2, Q_0/2)$.*

Proof. This follows directly from the distribution functions of the IW and Inverse-Gamma distributions. ■

There are three major issues with using the Inverse-Wishart priors in general, all of which are likely exacerbated in the case of TVP-VAR, since the system is not well identified. The three issues are: (1) The marginal distribution on variances is distributed as Inverse-Gamma, (2) There is only one degree of freedom, and (3) For low degrees of freedom, there is (strong) comovement between correlation and variance.

While these issues are sometimes mitigated when an Inverse-Wishart prior is used for covariance estimation (see Alvarez et al. (2016) for more details), they can be serious issues in a TVP context. Going through them one-by-one:

1. The marginal distribution on variances are Inverse-Gamma: The Inverse-Gamma dis-

tribution has zero mass around 0, which is problematic for estimating the posterior distribution of variances. This problem has been studied in a Hierarchical Modeling context by Gelman (2006) and Gelman and Hill (2007), who show that the problem is severe when forming an uninformative prior, and the true variance is very small. In a Time-Varying-Regression model selection context, Frühwirth-Schnatter and Wagner (2010) shows that using Inverse-Gamma distributions as the priors for the time-varying parameters bounds constant-parameter variance away from 0. They therefore recommend using half-normal distributions for priors, which have positive mass on 0. Although my interest is not specifically model selection, the boundary away from 0 is still an issue for inference.

2. There is only one degree of freedom: The main advantage of Inverse-Wishart priors is that they are priors over all correlations and variances at once. But this is also a disadvantage. One might have different amounts of prior information about correlations and variances. Moreover, it also requires the same amount of information for all parameters. Perhaps the practitioner has considerable nonsample information about variances, but very little about correlations. Or even stronger, perhaps the econometrician has more prior information about the variance of AR(1) parameters than the other variances, or more prior information about one correlation but not another. With only one degree of freedom, the Inverse-Wishart distribution does not provide such flexibility. The literature on Hierarchical modeling, starting with Barnard et al. (2000), developed priors that estimate correlations independently (or at least separately) from variances.
3. For low degrees of freedom, there is strong comovement between correlation and variance: Due to the nature of the distribution function, when one of correlation or standard deviations are high, the highest probability events correspond to the the other also being high.

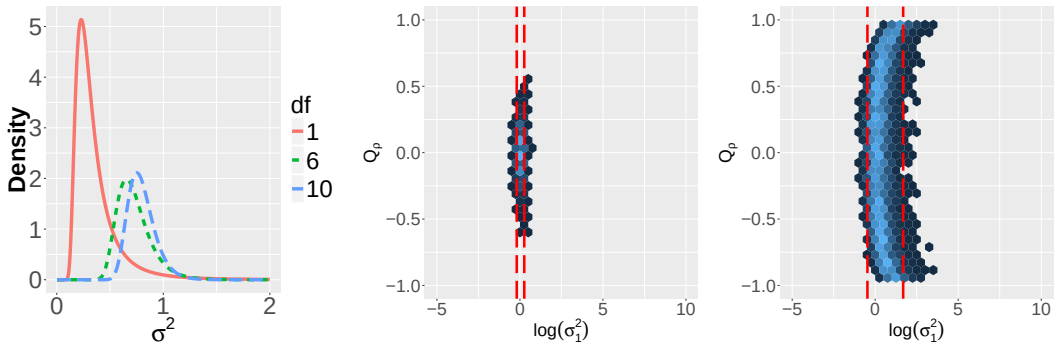
For each of these issues, I present graphical evidence to support the critiques in Figure 15. Issue (1) is shown by plotting the density function for an Inverse Gamma distribution with

various degrees of freedom (1, 6, and 10) and scale parameter chosen such that 90% of the mass is less than 1. For Issues (2) and (3), I repeatedly make draws from an Inverse Wishart distribution and decompose the draw into correlation and standard deviation. I make 10,000 draws from the IW prior centered at the two-dimensional identity matrix with three and 40 degrees of freedom in subfigures 15c and 15b respectively. For Figure 15c, I use the minimal degrees of freedom, which should produce an uninformative prior, yet the figure shows that the "uninformative" prior is actually informative in two dimensions. First, variances are limited to a range close to 1, and second, matrices with high correlation and low variance are impossible. For Figure 15b, I use a higher degree of freedom parameter, which indicates more prior information that the covariance matrix is centered around the identity matrix. But the IW prior cannot differentiate between high prior confidence in variance of 1 or correlation of 0. The Image shows that both are induced via IW priors.

Figure 15: Inverse Wishart Problems

I visually describe each of the three major issues with Inverse-Wishart Distributions as a prior. In Figure 15a, I plot the analytic density function for an Inverse Gamma distribution with degrees of freedom 1, 6, and 10 and location chosen such that 90% of the density is less than 1. For Figures 15b and 15c, I plot 10,000 draws from an Inverse Wishart distribution centered at the two-dimensional Identity Matrix for different degrees of freedom. Figure 15a plots the Inverse-Gamma pdf for various degrees of freedom. Note that for all degrees of freedom, the variances are bounded away from 0. Figure 15c plots draws from and Inverse Wishart distribution with 40 degrees of freedom, while Figure 15b presents draws from an IW distribution with three degrees of freedom. The vertical dashed red bands are 95% bounds on the variance.

(a) IG pdf (b) $\mathbf{IW}(\mathbf{40}, \mathcal{I})$ (c) $\mathbf{IW}(\mathbf{3}, \mathcal{I})$



Returning to the prior for $Q_\beta, Q_\beta \sim IW\left(\tau, \kappa_Q^2 \cdot \tau \cdot Q_0\right)$. In addition to selecting a family of priors (in this case IW), one must also select hyperparameters, which should be done

with care. The TVP-VAR literature has long recognized that the choice of hyperparameter can have a large influence on estimation, especially the choice of κ_Q. Indeed, Stock and Watson (1996b) tests a range of κ_Q hyperparameters and finds that for forecasting purposes, the optimal choice of κ_Q depends on the series to forecast and model choice. Primiceri (2005) proposes a complex, Reversible Jump MCMC procedure over the values $\kappa_Q \in \{0.01, 0.05, 0.1\}$, but also suggests that $\kappa_Q = .01$ is a reasonable choice. Cogley and Sargent (2002) also choose $\kappa_Q = 0.01$, while Cogley and Sargent (2005) use 0.18 and Cogley (2005) use 0.53. Primiceri (2005) acknowledges that estimation can be sensitive to the choice of κ_Q, but ultimately chooses 0.01 for a couple reasons: (1) to be consistent with the extant literature at the time, (2) following Cogley (2005), theoretical considerations paired with long-term trends about consumption growth indicate the variances should not be large, and (3) the MCMC procedure indicated $\kappa_Q = .01$ had highest posterior probability. These specific choices for κ_Q are of course dataset specific, and while they might be optimal for their respective applications, there is no guarantee they will work well in other cases. Indeed, Amir-Ahmadi et al. (2016) propose a hierarchical prior for κ_Q as a parameter to be estimated.

In setting Q_0, the literature has long followed Nyblom (1989) and Stock and Watson (1996b) by setting $Q_0 = V(\hat{\beta}_{OLS})$. As an aside, this indicates another problem with Inverse-Wishart distributions — that is, the necessity of a scale matrix. When the prior scale matrix is far from the likelihood, the posterior will exhibit shrinkage away from the maximum likelihood estimator. This is not always a good thing, and when using the IW distribution, it can produce shrinkage in unexpected directions, which I will further explain later.

2.2.2 Solutions from Bayesian Hierarchical Modeling

The issues with the IW distribution are well documented through many strands of literature, and one particular strand of the statistics literature has offered possible solutions. In the Bayesian Hierarchical Modeling literature, the researcher is interested in estimating regression coefficients, but feels strongly that information may be pooled — that is, one group's

regression coefficient is somehow related to another group's.

The classical setup in the Bayesian Random Effects model is that there are m linear regressions:

$$Y_i = X_i\beta_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma_i I_T), \quad i = 1, \ldots m,$$
$$\beta_i|\bar{\beta}, Q \overset{iid}{\sim} N(\bar{\beta}, Q).$$

Where Y_i is a $T \times 1$ vector, and X_i is a $T \times K$ matrix.

For example, a researcher might be estimating a CAPM model within a given sector. *A priori*, a firm's relationship with the market has mean $\bar{\beta}$, and covariance Q. Moreover, a firm's relationship with the market can be related to that of other firms,' which would be possible with a non-diagonal Q matrix. The econometrician encodes nonsample information about the parameters via priors on $\bar{\beta}$ and Q. I will focus on the solutions for estimating Q, since they are applicable to the TVP application here.

The original approach to modeling this relationship comes from Lindley and Smith (1972), who proposed using an Inverse Wishart prior on Q. Given the shortcomings of the Inverse Wishart prior, other statisticians have recommended alternative priors, to varying degrees of success. One promising line of research attempts to break the covariance matrix into correlation and variance components:

$$Q = diag(S)R\ diag(S)$$

This idea was proposed by Barnard et al. (2000), who used an uninformative prior for correlation coupled with any suitable prior for variance, such as log-normal or Inverse-Gamma to form a prior for the full covariance. Specifically, as a prior for correlation, they used the marginal distribution of an Inverse Wishart centered at the Identity matrix, which has a closed form-kernel:

$$f(R|\nu) \propto |R|^{\frac{-1}{2}(\nu+k+1)} \left(\prod_{i=1}^{k} r^{ii} \right)$$

Where ν is the degrees of the freedom, k is the dimension of the covariance and r^{ii} is the i^{th} diagonal element of R^{-1}. This prior flexibly allows for shrinkage specifically on the correlation matrix, by choosing higher degrees of freedom, ν. For an uninformative prior on correlation, one can simply choose $\nu = K + 1$.

In practice, Barnard et al. (2000) embeds this prior within a Gibbs Sampler, where the covariance is drawn element-by-element via a Griddy Gibbs Sampler (Ritter and Tanner (1992)), which involves evaluating the posterior over a pre-specified grid, and randomly sampling from the estimated posterior. While the procedure can work well (assuming the range of the grid is well-specified), in practice, the process is very slow, especially for larger covariances. Nonetheless, this paper introduced the concept of splitting the covariance into variance and correlation components.

Due to the computational issues, O'Malley and Zaslavsky (2008) proposed using the Inverse Wishart distribution for drawing and approximate correlation while leaving variances drawn log-Normal. Under this paradigm, $Q = diag(S)\widetilde{R}\ diag(S)$, and $\widetilde{R} \sim IW(\nu, I_K)$. While this solves most of the computational issues, it still requires a Metropolis-Hastings step to draw the variances. To solve this issue, Huang and Wand (2013) and Menictas and Wand (2013) instead suggest to use a Gamma hyperprior on variance. Thus, the full prior is:

$$\begin{aligned} Q|\nu, a_1, a_2 \dots a_K &\sim IW(\nu, 2\nu diag(1/a_1, 1/a_2, \dots 1/a_K)) \\ a_k &\overset{iid}{\sim} IG(\alpha, 1/A_k^2), \qquad A_k \text{ large}, \quad k = 1, \dots K \end{aligned}$$

Theorem 1. *When $\alpha = 1/2$, the marginal distribution on each variance is $t_\nu^+(0, A_k)$, where*

A_k is the standard deviation.

Proof. The marginal distribution of each of the variances is distributed as $\sigma_k^2 \sim IG(\nu/2, \nu/a_k)$ by Property 3 of IW distributions. Wand et al. (2011) showed that if

$$x|a \sim IG(\nu/2, \nu/a),$$
$$a \sim IG(1/2, 1/A^2),$$

then $\sqrt{x} \sim t_+(0, A)$.

■

Theorem 2. *The conditional posterior of* $a_k|others \sim IG(\frac{\nu}{2} + \alpha, \nu(Q^{-1})_{[kk]} + \frac{1}{A_k^2})$.

Proof. This follows from Bayes Theorem via the distribution functions of IW and IG random variables. ■

Theorem 3. *The marginal distribution on correlations has density:* $p(\rho_{i,j}) \propto (1-\rho_{i,j}^2)^{\nu/2-1}$

Proof. See Huang and Wand (2013) Property 3 for details ■

Based on Theorem 1, the prior is called Marginally Uninformative (MU). Theorem 2 is important because it maintains conditional conjugacy, and therefore in order to achieve separation between variance and correlation, one only needs one extra step in a Gibbs sampler.

While Barnard et al. (2000) introduced full independence, Huang and Wand (2013) and O'Malley and Zaslavsky (2008) opt instead for merely weak dependence. In the latter papers, the distribution of correlation conditional on variance is the same as an IW distribution. Despite that, the freedom to move across variance space makes the joint distribution substantially different.

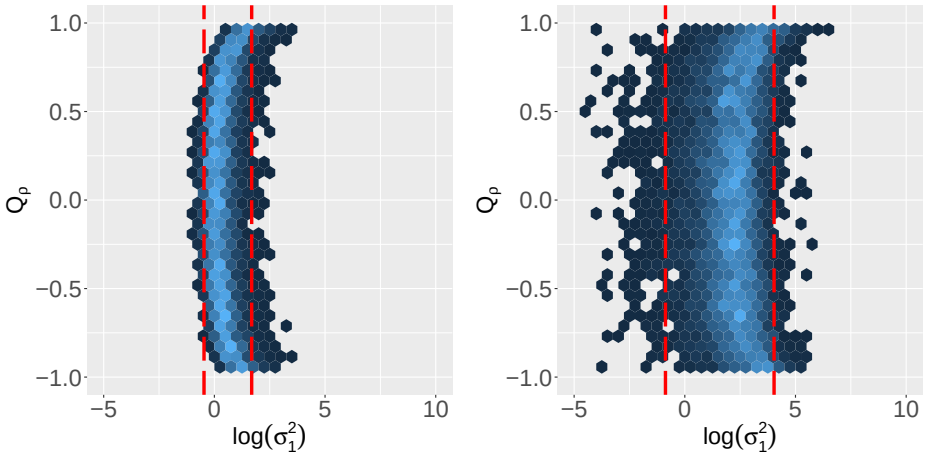
To compare the MU prior with the uninformative IW prior, I make 10,000 draws from the MU prior using $A_k = 10$ for $k = 1, 2$, and again plot correlation against the log-variance.

For comparison, I also include Figure 15c, which is the IW uninformative prior. The draws for the MU prior are in Figure 16b. Notice how the MU prior is indeed true to its name and uninformative over the variances. There is still some comovement between correlation and variance, but it is reduced. More importantly, the range of variances with substantive mass is significantly widened.

Figure 16: MU and IW priors

In this figure I compare the uninformative versions of the IW and MU priors. For each, I make 10,000 draws from the prior distribution and decompose the matrix into variance and correlation. I plot the correlation against the first variance for each of them. The vertical dashed red bands are 95% bounds on the variance.

(a) $\mathbf{IW}(\mathbf{3}, \mathcal{I})$ (b) MU



2.3 An Informative Prior for TVP-VAR

Return to the canonical Bayes Random Effects Model:

$$Y_i = X_i\beta_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma_i I_T), \quad i = 1, \dots m,$$
$$\beta_i | \bar{\beta}, Q \overset{iid}{\sim} N(\bar{\beta}, Q),$$

and observe that by changing the regression from a time-series to cross-sectional in every time period, we get something very similar to a TVP model:

$$Y_t = X_t\beta_t + \varepsilon_t, \quad \varepsilon_t \sim N(0, \Sigma_t), \quad t = 1, \ldots T,$$
$$\beta_t|\bar{\beta}, Q \overset{iid}{\sim} N(\bar{\beta}, Q).$$

It differs only in the evolution of the underlying parameters. The hierarchical model assumes all betas are iid over time, while the TVP model assumes they follow a random walk. As such, one might expect the developments in Hierarchical Modeling to have a (positive) impact on TVP modeling. Since the Huang and Wand (2013) prior blends the separation between correlation and variance with computational efficiency, it is my starting point.

While Huang and Wand (2013) developed their prior as a method for setting an uninformative prior, by changing A_k^2, one can also use the same setup to generate informative priors on variances. Indeed, this would be applicable for TVP-VARs, as the purpose of κ_Q is to shrink parameter variances and ensure they do not grow too large.

In this paper I offer two types of informative priors. The first is an Absolute Threshold Prior, and the second is a Relative Threshold Prior. For the absolute threshold prior, the user specifies a maximum prior variance, and a confidence level in that maximum variance. For instance, "My prior is that, with 90% probability, the error in time-varying parameters ($Q_{\beta[ii]}^{1/2}$) has standard deviation less than 1." The relative threshold prior uses a calculation originating from Cogley (2005), which is as follows: Start with the model assumption that β_t is a normally distributed random walk around β_0, so $\beta_t \sim N(\beta_0, t \times \sigma_\beta^2)$. Then, since the parameters should not move around too much, let the probability of a large change in β be relatively small. The values Cogley (2005) use is that a 20% change in β over 40 years should occur only 5% of the time.

$$
\begin{aligned}
&Pr(\beta_t \geq 1.2\beta_0 \cup \beta_t \leq .8\beta_0) = 0.05 \\
\Rightarrow &Pr(0.8\beta_0 \geq \beta_t) + Pr(1.2\beta_0 \leq \beta_t) = 0.05 \\
\Rightarrow &2\left[1 - \Phi\left(\frac{0.2\beta_0}{\sigma_\beta\sqrt{t}}\right)\right] = 0.05 \\
\Rightarrow &\frac{0.2\beta_0}{1.96\sqrt{t}} = \sigma_\beta
\end{aligned}
$$

Plugging in $t = 40$, and β_0 equal to the long-run mean of quarterly inflation (.03/4) they solve for σ_β. While they use this σ_β to set κ_Q directly, I merely use this value as a threshold for each of the prior variances. In practice, there can be some issues with the relative threshold prior. First of all, it requires a good estimate for β_0, and a fixed-parameter estimate from the burn in sample might be far from the time-varying analog. Second, OLS is scale dependent. That is a good thing for the intercept term — if the units are larger, then the time-varying variance should be larger as well — but there is no way to change the scales of the AR terms. That means that the amount of prior time-variation would be dependent on how autocorrelated the variable is, which is an unrelated metric.

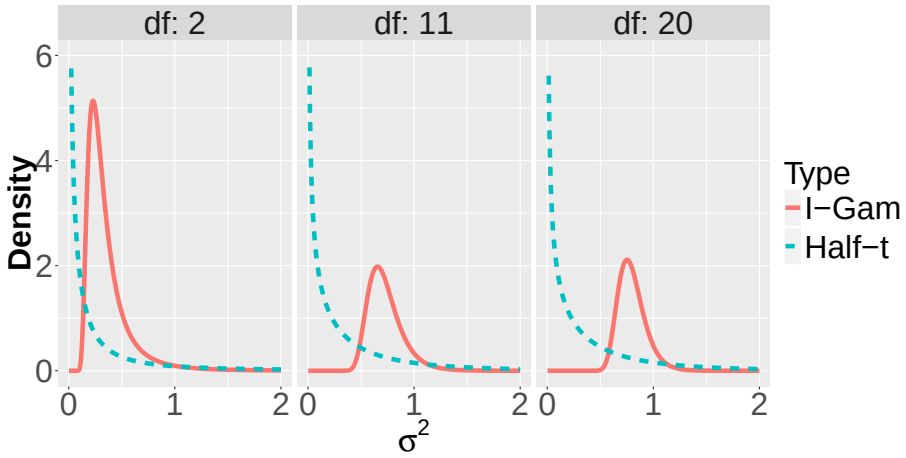
Threshold priors are not unique to the folded-t family and indeed can also be specified via Inverse-Gamma distributions. However, in correcting the statistical properties, the folded-t family also becomes easily interpretable.

Consider again the example above: "My prior is that, with 90% probability, the time-varying standard deviation is less than 1." I choose the rate parameter for Inverse-Gamma and the variance for half-normal distribution that best approximates that statement for three different degrees of freedom, corresponding to three levels of information about the covariance matrix, two, 11 and 20. Based on Properties 1 and 2 of IW distribution, when degrees of freedom is two, that is an uninformative prior. The higher degrees of freedom represent more information, where 20 is the value Primiceri (2005) used. For each of these

degrees of freedom, I plot the distribution function for Inverse-Gamma and Half-t random variables, all of which are in Figure 17.

Figure 17: Problems with Inverse-Gamma Family: Degrees of Freedom

In this figure I plot the distribution function for Inverse-Gamma and Half-t random variables across three degrees of freedom (2, 11 and 20). The higher the degrees of freedom, the more informative the prior on covariances. For the Inverse-Gamma prior, that information is in both variance and correlation space, whereas for the Half-t distribution, it is only in correlation space. The dashed (green) line depicts the Half-t distribution, while the solid (red) line depicts Inverse-Gamma.



In examining the images, notice how the Inverse-Gamma implementation changes the prior statement significantly. For all degrees of freedom, the prior effectively excludes small variances, but the problem gets worse as the degrees of freedom increase. For the highest degrees of freedom, the Inverse-Gamma prior is tightly peaked around 1. On the other hand, the half-t prior does not change as the degrees of freedom changes. This therefore emphasizes the fact that in the separation paradigm, degrees of freedom only have an effect on correlations, and leave variances alone.

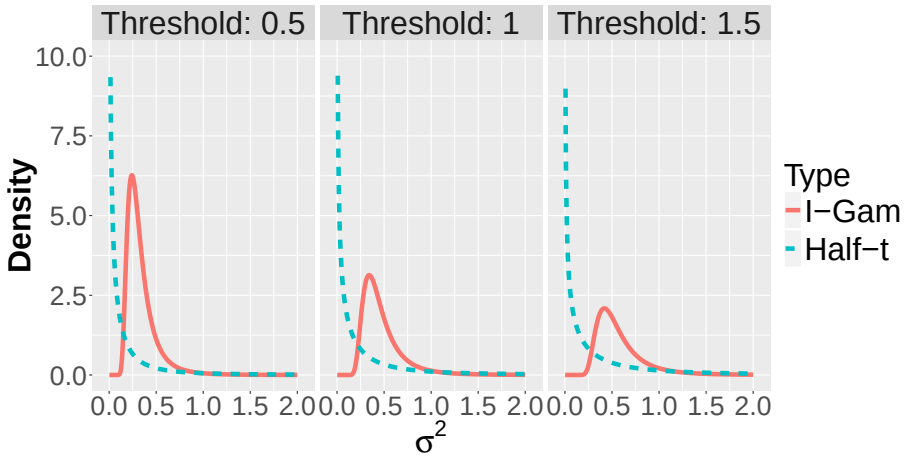
In the case with the highest degrees of freedom, 20, while the prior statement is that variance should be less than 1, mathematically it ends up that the variance should be between 0.6 and 1. Despite information only relating to an upper bound, the Inverse Wishart distribution also imposes a lower bound, and therefore skews the non-sample information provided by the econometrician.[1]

[1]One could also specify a prior such as, "My prior is that standard deviation is centered at 1 with

Next, I also plot the two marginal distributions with varying absolute thresholds to indicate various levels of shrinkage. All priors are calibrated to place 90% of the prior mass less than the threshold value. The plots are displayed in Figure 18.

Figure 18: Problems with Inverse-Gamma Family: Scale

In this figure I plot the distribution function for Inverse-Gamma and Half-t random variables for three scale values and 1 degrees of freedom (uninformative over matrices). Each prior is calibrated to have 90% of the prior mass less than a threshold (0.5, 1, and 1.5). The dashed (green) line depicts the Half-t distribution while the solid (red) line depicts Inverse-Gamma.



As I mentioned previously, Inverse Wishart priors require a pre-specified location matrix, which can have unintended consequences when used for shrinkage. Take for example, the rightmost panel of Figure 18 and observe the Inverse-Gamma distribution function. The prior states that with 90% probability the variance should be less than 1.5. In fact though, the prior forces the mass between 0.2 and 1.5. In a case where the MLE can be approximated beforehand (via an Empirical Bayes-type procedure) one can specify the prior location in order to induce shrinkage (either smaller or larger) as desired. In the case of TVP though, the underlying covariance matrix cannot be approximated from data, so there is a danger of overstating the posterior variance. For instance, in the case where the data has constant parameters, the Inverse-Gamma prior will shrink posterior variances larger than the Half-t

variance 0.02, which roughly corresponds to the Inverse-Gamma prior with 20 degrees of freedom. In that case, one could still use the half-t distribution, but it would have to a *non-centered* half-t. In TVP contexts, we are most concerned with not allowing variances to be too big as opposed to knowing where the variances should be centered, so non-centered half-t distributions are not considered.

prior, despite the fact that both should be doing the same shrinkage. Thus, despite the fact that the econometrician intends to use the prior for (weak) shrinkage towards 0, she may be inflating it instead. I show this property more fully throughout the simulations and data examples.

Lastly, I once again repeat the exercise of making 10,000 draws of a two-dimensional covariance and decomposing it into variance and correlation. This time I use two values for extra degrees of freedom (above the covariance dimension), one and 21, and compare the results for IW and Separation draws. For each of them, I use the Absolute Threshold Prior that 90% of the prior mass should be less than 1. All images are plotted in Figure 19.

The prior draws provide a necessary compliment for the distribution functions and visually display all the concepts discussed above. For both degrees of freedom, the Inverse Wishart prior excludes very small variances, which may generate inflated posterior variances. When degrees of freedom increases, the Inverse Wishart prior draws become very concentrated, both in terms of variances and correlations. On the other hand, true to its name, the separation prior retains its distribution over variances, but its distribution over correlations mimics that of Inverse Wishart.

Thus, the informative Separation prior alleviates the issues associated with the Inverse Wishart prior and is just as easy to include in a Gibbs sample — it requires only one additional step of drawing a vector of Inverse-Gamma random variables. Note also that in order to draw an Inverse Wishart matrix, most algorithms first compute its inverse (a Wishart random variable), so obtaining the inverse of Q_β comes for free.

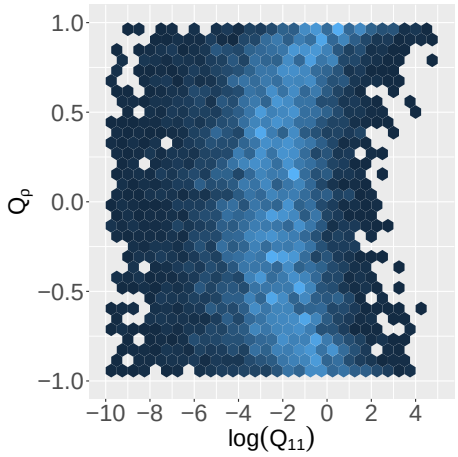
Once parameters, $A_1, A_2, \ldots A_K$ (for each respective class of prior) are set, one can proceed into the Gibbs sampler steps to draw β_t and Q_β

1. *For Separation Prior, draw* $a_k \sim IG(1/2, 1/A_k^2)$. *For IW, set* $a_k = 2 \times A_k$
2. Given other parameters, use Carter-Kohn (Kalman Filter and Smoother) to draw β_t, $t = 1, \ldots T$

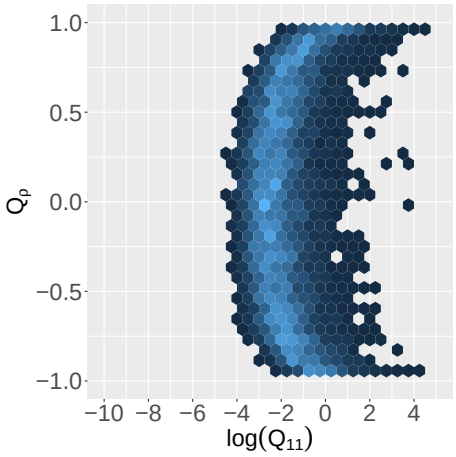
Figure 19: Separation and IW Informative priors

In this figure I compare informative version of the Separation and IW priors. For each of them, I set the prior location using the Absolute Threshold Prior that 90% of the prior mass should be less than 1. The figures in the top row (Figures 19a and 19b) use one extra degree of freedom, while those in the bottom (Figures 19c and 19d) use 21. All images plot 10,000 draws from their respective priors.

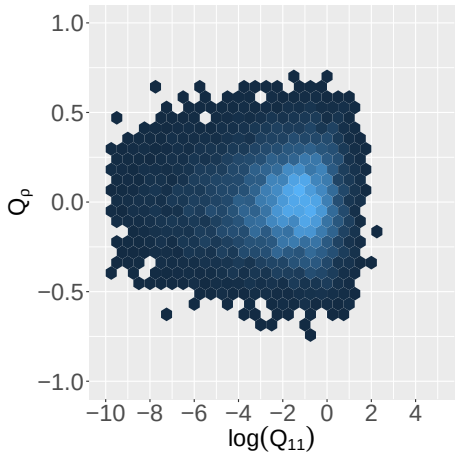
(a) Separation(1)



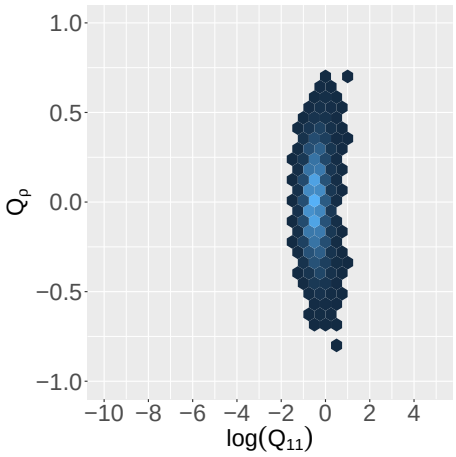
(b) IW(1)



(c) Separation(21)



(d) IW(21)



(a) Define $\widehat{\varepsilon}_t^{\beta} = vec(\beta_t) - vec(\beta_{t-1})$

(b) Define $Q_{\beta}^{MLE} = \widehat{\varepsilon}_t^{\beta\prime}\widehat{\varepsilon}_t^{\beta}$

3. Draw $Q_{\beta} \sim IW(\nu + T, 2\nu diag(a_1, \dots a_K) + Q_{\beta}^{MLE})$

(a) For Separation Prior, draw $a_k \sim IG(1/2 + \nu/2, \nu(Q_{\beta})_{[ii]}^{-1} + 1/A_k^2)$

The rest of the sampler continues in the standard fashion.

My method is also similar to the innovation in Amir-Ahmadi et al. (2016). They proposed a hierarchical prior for κ_β, where one of the hierarchical priors is an Inverse-Gamma distribution. Like mine, each time-varying parameter error variance is also distributed as a half-normal distribution, but while each of the variances have independent priors in my setup, they are highly dependent in Amir-Ahmadi et al. (2016).

2.3.1 Posterior Inference

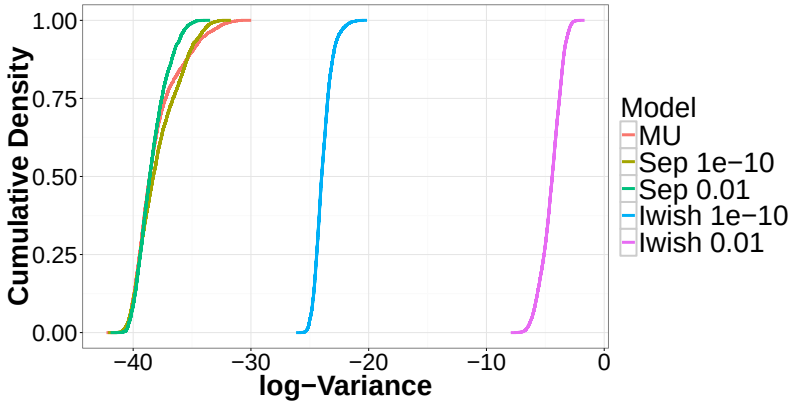
Not surprisingly, the priors impact posterior inference. To display this, I present two DGPs in order to display how the negative properties about Inverse-Wishart distributions effect posterior distributions.

First, consider a DGP where the parameters are actually fixed, yet the econometrician applies a TVP-VAR. The DGP is a two-dimensional VAR(1), but the econometrician runs a TVP-VAR model with 1 lag (6 total parameters). The econometrician is fairly confident that the amount of time-variation is small, if it exists at all, and therefore uses an absolute threshold prior with threshold values of either 0.01 or 10^{-10} and confidence level 90% and the Marginally Uninformative prior. Observe the posterior distributions (in the form of empirical cumulative distribution functions) of the variance of the (misspecified) time-varying parameter for the intercept term of the first series in Figure 20. While all the models produce small variances, essentially agreeing that there is little to no time-variance in the parameter, the separation prior is much more confident. All three separation priors produce posterior distributions that are essentially the same. The posterior mean for all three is around 10^{-16}, which is approximately 0. On the other hand, the Inverse-Wishart priors produce posteriors that are much larger. The Inverse Wishart priors yield posterior means on the order of 10^{-11} and 10^{-2} for the absolute thresholds at 10^{-10} and 0.01 respectively.

Next, I once again consider a VAR(1) in two dimensions, but this time the time-varying variances are quite large — 0.01. While this is an amount considerably larger than one would expect to see in most datasets, it provides an opportunity to see how well each

Figure 20: Posterior in Misspecified TVP Model

In this figure, I compare the posterior distribution of a misspecified TVP-VAR. The true DGP is a two-dimensional VAR(1) with constant parameters, whereas the estimation routines are TVP models with 90% threshold priors with variances 10^{-10} and 0.01 and the Marginally Uninformative prior.

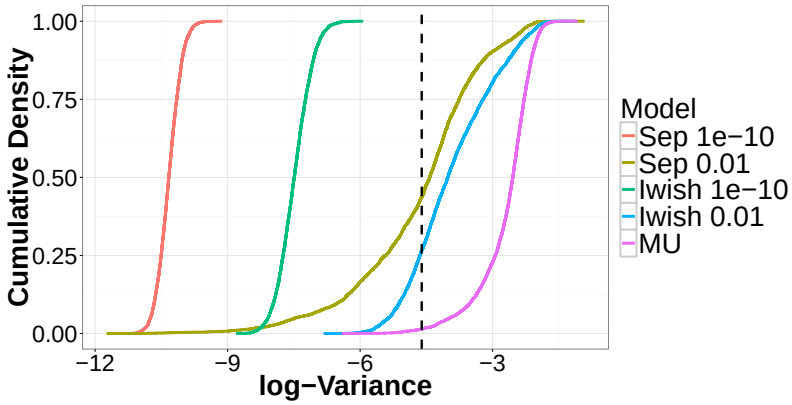


prior acts as a shrinkage prior. I again use the same five priors (Marginally Uninformative, threshold at 10^{-10} and threshold at 0.01) for both Separation and Inverse-Wishart classes. In this case, when the data time-varying variance is as large (or larger) than the prior, then the priors all work as expected. The Marginally Uninformative prior does a decent job estimating the Maximum Likelihood (though slightly overestimates), both 0.01 threshold priors generate posterior distributions centered at nearly the MLE, and both shrinkage priors generate posterior distributions with time-varying variances much smaller than the likelihood suggests. The separation prior has more mass at smaller values though, so the posterior distribution associated with that prior generates more shrinkage.

These examples show the danger of specifying a location for the Inverse-Wishart prior in a TVP context. Whereas the Half-t prior always includes 0 and is adjusted by increasing or decreasing the variance, the Inverse-Wishart prior must also specify a centering location (even if one wishes to use the distribution as a shrinkage prior). If that location is larger than the likelihood covariance, then the prior will inflate posterior variance instead of shrinking it. On the other hand, if the likelihood covariance is larger than the prior center, then both Inverse-Wishart and Separation priors will act as traditional shrinkage priors (that is,

Figure 21: Posterior in TVP Model with Large Variances

In this figure, I compare the posterior distribution of a misspecified TVP-VAR. The true DGP is a two-dimensional TVP-VAR(1) with time varying variances 0.01, and the estimation routines are TVP models with 90% threshold priors with variances 10^{-10} and 0.01 and the Marginally Uninformative prior. The dashed vertical line represents the true value.



shrinking towards smaller values). Lastly, the Marginally Uninformative prior usually does a pretty good job of estimating the Maximum Likelihood values, though it can slightly inflate them.

2.4 Simulation Study

In the previous sections I have depicted the deficiencies of the Inverse-Wishart distribution as a prior, and shown how the Separation prior can alleviate these issues. Moreover, by using an informative separation prior, with shrinkage towards constant parameter models, my prior can better estimate those constant parameters.

In this and the following sections, I show that my new prior also has improved frequentist properties, as seen via an extensive simulation study and multiple forecasting exercises. While the gains from changing the prior are not always very large, their cost (in computation time and coding difficulty) is correspondingly small as well.

In the following section, I outline a set of simulations to test my new priors. Most of the datasets I use are low dimensional VARs due to computation-time considerations, but some

are larger. For each dataset, I test the Primiceri style prior, the Marginally Uninformative prior, three absolute threshold priors (both IW and Separation style) and one relative threshold prior (again, for each of the IW and Separation families). I repeat each DGP 100 times, and record the average MSE of Q_β, its diagonal elements, and the time paths of stochastic volatility and regression parameters.

I use six different DGPs, four of which are 2-dimensional VAR(1)s and two of which are 4-dimensional VAR(1)s. Five of them have time-varying parameters of varying degrees, and one has constant parameters. All have the same amount of stochastic volatility ($Q_s = I_d \times 10^{-4}$, where d is the cross-sectional dimension) and constant covariance ($H = I_d \times 0.01$).

For the two-dimensional VAR(1)s, Q_β has dimension 6, while for the four-dimensional VAR(1)s, Q_β has dimension 20. In Simulation 1, Q_β is diagonal, with all variances 10^{-4}. In Simulation 2, Q_β is diagonal and reduced rank, with four zero variances, and the other two 10^{-4}. In Simulation 3, Q_β is also reduced rank, with two zero variances, but also has some variances 0.1 and other 10^{-4}. There are also four nonzero correlations. In Simulation 4, Q_β is full rank with all variances 10^{-4}, and five pairwise correlations. In Simulation 5, Q_β is a random draw from an Inverse Wishart distribution centered at a diagonal matrix with variances 10^{-4} and 25 degrees of freedom. In Simulation 6, Q_β is the zero matrix, so all parameters are constant. For more details on each of the Q_β used in the simulations, please see the Appendix. Simulations 1-4 use a two dimensional dataset, while 5 and 6 are four dimensional.

I choose these values of variances since Primiceri (2005) found that time-varying variances were between 10^{-6} and 10^{-4}, so my reflect realistic amounts of parameter time-variation. I am also interested in how the methods perform on DGPs with more time-variation (simulation 3) and constant parameters (simulations 2, 3, 6). Moreover, dense error covariance matrices are essential based on the factor structure found on Stevanovic (2010). Some DGPs retain diagonal covariances as a baseline.

For Simulations 1-4, I generate 50,000 draws with 30,000 discarded as burn-in. For Simulation 5, I produce 80,000 with 50,000 discarded as burn-in and for Simulation 6, I generate 100,000 draws and discard the first 50,000. All DGPs are summarized in Table 15.

Table 15: Simulation Designs

Sim	$dim(Y_t)$	$dim(Q_\beta)$	$rank(Q_\beta)$	Variances	Unique Correlations	Posterior Draws
1	2	6	6	10^{-4}	0	50,000
2	2	6	2	10^{-4}	0	50,000
3	2	6	4	10^{-4}, 0.1	5	50,000
4	2	6	6	10^{-4}	6	50,000
5	4	20	20	$\sim 10^{-5}$	190	80,000
6	4	20	0	0	0	100,000

2.4.1 Priors

For each simulated dataset, I utilize ten different priors. Five are from the Inverse-Wishart class of priors and five are from the Separation hierarchical prior. From each class of priors, four are the same, only differing in the class, and one is specific to that class. I use three Absolute Threshold priors, where the prior probability is less than the cutoff is 90%. The cutoffs are 10, 0.01 and 10^{-10}. I use one Relative Threshold prior, which states that there is a 90% prior probability that the probability of a 20% move from the full-sample OLS is less than 95% over 40 periods. Lastly, for the Inverse Wishart prior, I also test the Primiceri prior, and for the Separation class, I try the Marginally Uninformative (MU) prior. For a summary of each, see Table 16. Details on each of the full posterior samplers can be found in the Appendix. The basic structure for all samplers was derived from code from Dmitris Koribilis's website, but I translated the code to C++ for speed, made some other speed improvements, and changed the OLS estimation from a pre-sample to the full-sample.

In Figure 22, I plot kernel density estimates of each of the absolute threshold priors and the true values used in the simulations. Since the relative threshold priors change based on the OLS estimates for each simulation replication, I only plot the absolute thresholds. The plot provides expectations for how each of the priors should perform.

Table 16: Posterior Samplers

	Sampler	Probability of 20% move	% Less than threshold	Absolute Threshold
1	Primiceri (P)	-	-	-
2	IW Absolute (IW1)	-	90%	10
3	IW Absolute (IW2)	-	90%	0.01
4	IW Absolute (IW3)	-	90%	10^{-10}
5	IW Relative (IWR)	10%	90%	-
6	MU	-	-	-
7	Sep Absolute (S1)	-	90%	10
8	Sep Absolute (S2)	-	90%	0.01
9	Sep Absolute (S3)	-	90%	10^{-10}
10	Sep Relative (SR)	10%	90%	-

When true error variance is larger than 0, both of the larger-threshold separation priors (S1 and S2) should deliver posteriors that are nearly identical. The true values are covered by both prior distributions, though there might be some moderate shrinkage towards larger variances in S1. On the other hand, IW1 should shrink variances larger for both variances, and both IW1 and IW2 should shrink variances larger for the small variance (10^{-4}). The very small threshold priors (IW3 and S3) place all prior mass substantially less than the true variances, and posterior variance should therefore be shrunk smaller.

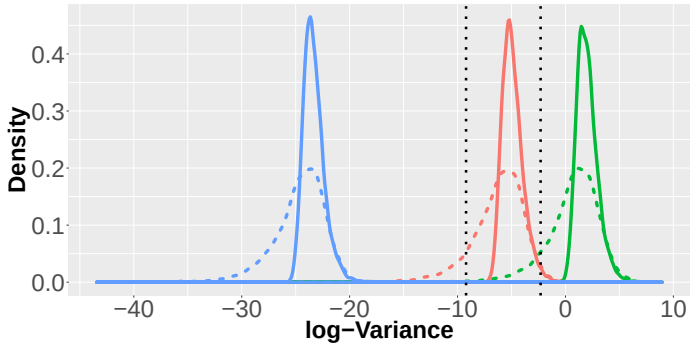
When the true error variance is exactly 0, all priors will slightly inflate posterior variances as compared to the MLE, but the amount of inflation will depend on the threshold. Moreover, since the Separation priors have a longer tail towards smaller values, they should better approximate the true variances.

2.4.2 Results

The purpose of these simulations is twofold. For one, I want to compare across IW priors to Separation priors. Second, I want to see which versions of the priors work best in different DGPs. I record the average MSE of β_t in Table 17, the average MSE of stochastic volatility in Table 18, the average Frobenius norm of Q_β in Table 19, and the average MSE of the diagonal of Q_β in Table 20. In order to normalize the results, I record all results relative to

Figure 22: Prior distributions used in simulations

Kernel density estimate of 10,000 draws from each absolute threshold prior distribution. The solid lines represent the IW priors, while dashed lines depict Separation priors. Vertical dotted lines designate the true values used in the simulations. Green lines represent thresholds at 10, pink lines represent thresholds at 0.01, and blue lines depict thresholds at 10^{-10}.



the Primiceri prior.

Table 17: β_t

Average MSE of $\hat{\beta}_t$ relative to true β_t. All values are normalized to make the average MSE of the Primiceri Prior 1.

	Sampler									
Sim #	P	MU	IW1	S1	IW2	S2	IW3	S3	IWR	SR
1	1.00	0.54	1.13	0.54	0.49	0.52	1.08	1.01	0.78	0.54
2	1.00	0.45	1.98	0.44	0.38	0.42	0.96	0.95	0.88	0.44
3	1.00	0.14	0.34	0.14	0.14	0.14	1.00	0.87	0.23	0.14
4	1.00	0.80	3.13	0.80	0.67	0.76	1.05	1.26	1.58	0.78
5	1.00	0.89	3.37	0.89	0.74	0.87	0.65	1.29	2.12	0.89
6	1.00	2.27	3.86	2.24	2.03	2.24	0.87	0.90	4.58	2.25

Table 18: Stochastic Volatility

Average MSE of time-varying volatility ($diag(H) \times \Sigma_t$) relative to the true value. All values are normalized to make the average MSE of the Primiceri Prior 1.

	Sampler									
Sim #	P	MU	IW1	S1	IW2	S2	IW3	S3	IWR	SR
1	1.00	1.05	1.94	1.04	1.09	1.06	1.15	1.30	1.06	1.05
2	1.00	1.00	1.34	1.00	0.96	0.97	0.95	0.98	1.02	0.99
3	1.00	0.06	0.07	0.06	0.06	0.06	1.01	0.07	0.06	0.06
4	1.00	0.89	0.78	0.89	0.85	0.87	0.95	1.36	0.98	0.90
5	1.00	0.94	1.69	0.92	0.89	0.90	0.91	1.13	1.40	0.91
6	1.00	1.18	2.40	1.19	1.22	1.18	1.00	0.99	1.32	1.18

Table 19: Error Covariance of β_t

Average Frobenius norm of the Error Covariance of β_t (Q_β) relative to its estimated value. All values are normalized to make the average Frobenius norm of the Primiceri Prior 1.

	Sampler									
Sim #	P	MU	IW1	S1	IW2	S2	IW3	S3	IWR	SR
1	1.00	4.69	816	4.67	3.13	4.11	0.65	3.44	57.30	4.65
2	1.00	3.15	510	3.16	2.13	2.77	0.59	2.64	37.52	3.11
3	1.00	0.98	8.10	0.98	0.98	0.98	0.99	4.03	2.30	0.98
4	1.00	4.70	748	4.68	2.81	4.25	0.80	4.48	61.78	4.92
5	1.00	1.43	501	1.43	0.82	1.32	0.07	2.68	31.96	1.41
6	1.00	11.81	4430	11.81	6.54	10.92	0.00	0.00	448	11.70

Table 20: Error Variance of β_t

Average MSE of the diagonal of the Error Covariance of β_t ($diag(Q_\beta)$) relative to its estimated value. All values are normalized to make the average MSE of the Primiceri Prior 1.

	Sampler									
Sim #	P	MU	IW1	S1	IW2	S2	IW3	S3	IWR	SR
1	1.00	18.96	5.9E5	18.86	8.04	14.42	0.54	51.81	3066	18.75
2	1.00	4.67	1.2E5	4.75	2.01	3.57	0.19	13.24	789	4.55
3	1.00	0.75	91.93	0.74	0.74	0.75	0.96	103.66	11.78	0.75
4	1.00	22.44	4.8E5	21.17	6.61	22.21	0.67	52.65	4666	33.40
5	1.00	0.58	7.3E4	0.58	0.19	0.49	0.00	5.87	324	0.57
6	1.00	9.30	1.3E6	9.30	2.91	7.97	0.00	0.00	3.5E4	9.14

First, comparing different specifications of the prior within each of the classes, notice that for the IW class, the results vary substantially in all categories depending on which prior hyperaparameters are used. The absolute threshold at 10 shrinks the error variance towards too large a value, which in turn produces poor estimates for β_t and Q_β. The other IW hyperparameters also shrink in a specific direction, which in turn produces different MSEs. On the other hand, the Separation class is remarkably consistent across various hyperparameters. That again makes sense since the larger thresholds merely indicate less prior information that the likelihood variance is close to 0. Since most simulation DGPs all have parameter error variances that are 10^{-4} or smaller, they fit well within the prior mass, so all priors act the same. On the other hand, the threshold prior at 10^{-10} does shrink towards 0, which increases average MSE of β_t and Q_β for all simulation designs except the constant

parameter model, where it does much better than the others.

Within each hyperparameter and comparing across IW vs Separation classes, the separation priors do better in terms of the time path of β_t, but interestingly often do slightly worse in estimating the error covariance matrix. This is even true for the constant parameter model, where the IW prior with threshold at 10^{-10} does a better job estimating the error variances than the separation prior.

2.5 Hyperparameter Selection

Since TVP models are so overparametrised, the posterior distributions can be very sensitive to the choice of hyperparameter. Indeed, as described above, the hyperparameter selection for κ_Q can be either computationally cumbersome (Primiceri (2005) Jump MCMC procedure), or loosely based on theory (Cogley (2005)). However, most papers merely use their suggested hyperparameters, despite the fact that they are tailored for specific databases. As an alternative, Amir-Ahmadi et al. (2016) develop a portable selection routine, treating κ_Q as an additional parameter in a hierarchical setup.

One advantage the separation priors have over Inverse Wishart priors is that when the absolute threshold is large enough, the prior will act as an uninformative prior, and posteriors will not be sensitive to prior hyperparameter choice. Indeed, this was evident in the simulations. However, when one wishes to use the prior as a shrinkage prior, some formal model selection is necessary is decide between levels of shrinkage.

In this paper, I use two methods, the Divergence Information Criterion (DIC) and the Widely Applicable Information Criterion (WAIC), to choose between levels of absolute shrinkage (for both IW and Separation priors). The main advantage for using these methods over other model selection methods (Jump MCMC, Marginal Data Density, etc) is that they are easily produced based on the output of a Gibbs Sampler (conditional likelihood) and are applicable to nonlinear models.

2.5.1 Traditional Bayesian Methods

Traditionally, Bayesian model selection utilizes the Marginal Data Density (MDD) for model selection, which can be found by inverting Bayes Rule. For data Y and parameter vector θ, by Bayes Rule,

$$p(\theta|Y) = \frac{p(Y|\theta)p(\theta)}{p(Y)},$$
$$\implies p(Y) = \frac{p(Y|\theta)p(\theta)}{p(\theta|Y)}.$$

For a set of candidate models, $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_J$ endowed with prior probabilities, $p(\mathcal{M}_1)$, $p(\mathcal{M}_2), \ldots, p(\mathcal{M}_J)$, one can calculate posterior model probabilities by using the MDD and integrating out the parameter vector:

$$p(Y|\mathcal{M}_j) = p(\mathcal{M}_j) \int p(Y|\theta, \mathcal{M}_j)p(\theta|\mathcal{M}_j)d\theta.$$

These posterior model probabilities can then be used for model averaging or selection. The different models may merely be different hyperparameters, so this framework also includes hyperparameter selection.

In the case of a linear model with Gaussian priors and likelihood, the data density can be calculated in closed form, which makes hyperparameter selection straightforward. Nonetheless, rigorous model selection is rather rare. A few exceptions are: Del Negro and Schorfheide (2004), who used output from a DSGE model to select priors for VARs, and in turn improve forecasts. Similarly, Del Negro and Schorfheide (2011), Carriero et al. (2012) and others have used the MDD to select the variance of a Minnesota prior from a grid of possible values. More recently, Giannone et al. (2012) generalized the approach into a hierarchical modeling framework, which avoids the necessity of a grid.

However, those papers use the fact that some (or all) of the MDD can be calculated in closed form. More general methods for estimating the MDD (especially based on output

from a Gibbs Sampler), have been proposed by Newton and Raferty (1994), Chib (1995), Geweke (1999) and Sims et al. (2008). Nonetheless, most of the methods are based on the Harmonic Mean of the likelihood, and therefore can be numerically unstable, especially in high-dimensional environments. See the textbook treatment in Herbst and Schorfheide (2016) for their application to DSGE models.

While traditional Bayesian model selection methods are based on the MDD, DIC and WAIC approach the model selection problem from a different perspective. Thus, while DIC and WAIC are easily computed based on output from Bayesian estimation, they are more similar to frequentist methods. DIC (Spiegelhalter et al. (2002)) and WAIC (Watanabe (2010)) impose a bias correction on in-sample fit in order to estimate out-of-sample prediction error. In that way the methods are similar to the well-known AIC as an alternative to cross-validation. It is worthwhile to review the main concepts of model selection via minimizing out-of-sample prediction error as it applies to DIC and WAIC. My introduction of DIC and WAIC mainly follow exposition in Gelman et al. (2014).

2.5.2 Bias Correction Methods

For a general treatment of model selection based on bias correction methods from in-sample measures of fit, see Hastie et al. (2001), Chapter 7. For data y, and parameters, θ, the in-sample prediction error is:

$$Err_{in} = \frac{1}{T}\sum_{t=1}^{T} L(y_t, \theta),$$

where L is some general loss function (for instance, the negative likelihood). On the other hand, the out-of-sample error rate (for new data $\tilde{y}$, with distribution $f(y)$) can be found by taking expectations over over the DGP:

$$Err_{oos} = \frac{1}{T}\sum_{t=1}^{T} E_f[L(\tilde{y}_t, \hat{\theta})],$$

where $\hat{\theta}$ is the estimator that minimizes in-sample prediction error, Err_{in}.

In general, in-sample prediction error will be overconfident, but that bias can often be approximated so the model selection criterion will have the form of:

$$IC = insamplefit + biascorrection,$$

where the in-sample measure and bias correction varies by the assumptions made by the Information Criterion. Note that while the Schwarz Information Criterion (SIC/BIC) has this general form, it is actually an approximation to the MDD for a specific loss function.

One of the canonical examples of a bias correction method is the Akaike Information Criterion (AIC). Under certain assumptions, AIC uses asymptotic arguments to show that the bias correction is precisely the number of parameters in a linear model. Thus,

$$AIC = -2\left[p(y|\hat{\theta}_{MLE}) - p\right],$$

where p is the number of parameters in the linear model, and $p(y|\theta)$ is the log-likelihood function. The candidate model that minimizes AIC will therefore maximize the expected log-likelihood (asymptotically).

DIC

DIC is the information criterion introduced by Spiegelhalter et al. (2002), and is an ad-hoc Bayesian version of AIC. While AIC asymptotically selects the model that minimizes Kullback-Liebler divergence, it is only valid for linear models (where the degrees of freedom are easily calculated).

In order to adapt AIC into a Bayesian version, and to account for nonlinearities, DIC makes two changes to AIC. It replaces the Maximum Likelihood Estimator $\hat{\theta}_{MLE}$ with the posterior mean, $\hat{\theta}_{Bayes} = E(\theta|Y)$, and introduces an "effective number of parameters:"

$$p_{D1} = 2\left(\log p(y|\hat{\theta}_{Bayes}) - E_{post}(\log p(y|\theta))\right),$$

where E_{post} represents expectation over the posterior distribution. The expectation can be estimated by averaging over draws from the posterior distribution,

$$E_{post}(\log p(y|\theta)) \approx \frac{1}{S}\sum_{s=1}^{S} \log p(y|\theta_s),$$

where $\theta_1, \theta_2, \ldots \theta_S$ are draws from the posterior distribution. These draws are a direct byproduct of the Gibbs Sampler, which makes DIC trivial to implement.

Gelman et al. (2004) propose an alternative penalty parameter based on the variance:

$$p_{D2} = 2\mathbb{V}_{post}(\log p(y|\theta)),$$

where $\mathbb{V}_{post}$ represents variance over the posterior distribution, and can similarly be estimated based on the Monte Carlo average of the posterior draws.

In either case, the full IC is:

$$DIC_1 = -2\log p(y|\hat{\theta}) + 2p_{D1},$$
$$DIC_2 = -2\log p(y|\hat{\theta}) + 2p_{D2}.$$

While this IC is ad-hoc, there is some theoretical justification. Under a uniform prior and linear model, DIC is asymptotically the same as AIC, and therefore equivalent to leave-one-out cross validation.

WAIC

WAIC is similar to DIC, but more fully Bayesian, in that its measure of in-sample fit is the correct Bayesian object of interest. Similar to DIC, it adds on a correction term to adjust for overfitting. WAIC also has stronger theoretical justifications. Watanabe (2010) introduced WAIC, and showed that it is a valid approximation to leave-one-out cross validation even for singular statistical models (such as hierarchical models). Similar to DIC, under a uniform prior and linear model, the WAIC penalty reduces to the number of parameters, and is therefore equivalent to AIC.

Start by defining an appropriate measure of in-sample fit. Given data $y = (y_1, y_2, \dots y_T)'$ and parameter vector θ, define $p(y_t|\theta)$ as the density of data at time t given the parameters. The predictive density (of the fitted model) is simply the average over the posterior distribution:

$$\log(p_{post}(y_t)) = \log \int p(y_t|\theta)p(\theta)d\theta.$$

In practice, the posterior distribution is unknown, but can be estimated based on draws from the posterior distribution (such as the output from a Gibbs Sampler), $\theta_1, \theta_2, \dots \theta_S$:

$$\log(\widehat{p_{post}}(y_t)) = \log \left(\frac{1}{S} \sum_{s=1}^{S} p(y_t|\theta_s) \right).$$

In order to aggregate across time, use the simplification that $p(y|\theta) = \sum_{t=1}^{T} p(y_t|\theta)$, where the date-$t$ likelihoods could be only conditional likelihoods, in which case the full likelihood would only be a quasi-likelihood.

Similar to the frequentist likelihood function, the Bayesian measure of in-sample fit will be overconfident, and a bias-correction term is needed to construct an Information Criterion.

Watanabe (2010)'s proposed penalty term is:

$$p_{WAIC1} = 2\sum_{t=1}^{T}\left(\log(E_{post}[p(y_t|\theta)]) - E_{post}[\log(p(y_t|\theta))]\right),$$

but Vehtari et al. (2016) showed that a penalty term of

$$p_{WAIC2} = \sum_{t=1}^{T}\mathbb{V}_{post}(\log p(y_t|\theta))$$

is second-order equivalent to Bayes leave-one-out cross-validation (as opposed to only first-order equivalent). Once again, in order to estimate the posterior expectations and variances, one can use the Monte Carlo average of the draws from the posterior.

Thus, there are two version of WAIC:

$$WAIC_1 = -2\sum_{t=1}^{T}\log\left(\frac{1}{S}\sum_{s=1}^{S}p(y_i|\theta_s)\right) + 2p_{WAIC1}$$
$$WAIC_2 = -2\sum_{t=1}^{T}\log\left(\frac{1}{S}\sum_{s=1}^{S}p(y_i|\theta_s)\right) + 2p_{WAIC2}$$

2.5.3 Practical Consideration

While both WAIC and DIC (and both penalty terms for each IC) are equivalent to cross-validation, I find that in TVP models, the second version of the penalty (the one based on variance) works much better. The penalty is considerably stricter than the first version, which in turn favors simpler models. In terms of TVP models, these IC therefore support models with parameters closer to constant than the first penalty terms.

Between DIC and WAIC, DIC is more well-known and and has a longer usage history. DIC has been integrated into the output of BUGS since its inception, and more recently, a version of WAIC has been included in STAN.

Given its longevity, DIC has been applied to TVP contexts in a few situations. Chan and Eisenstat (2015) use DIC to decide between a VAR with and without Time-Varying parameters and/or stochastic volatility on a three-variable VAR applied to US and Australian data. For their US dataset, they find that both time-varying parameters and stochastic volatility provide substantial improvements in the likelihood, but that DIC prefers the constant parameter model with constant volatility.

Chan and Grant (2016) considered multiple forms of the DIC based on different types of likelihoods. The conditional likelihood, which conditions on all variables and latent states, the complete-data likelihood, which integrates out the latent states, and the integrated likelihood, which integrates out the parameter vector as well. They apply all three methods of computing DIC to a four variable TVP-VAR model where they test restrictions that each of the coefficient terms can be held constant. They find that the choice of likelihood can impact model selection, and that the IC values that use more conditional (as opposed to integrated) information have higher numerical standard errors. Moreover, Chan and Grant (2014) find, via simulation, that when the conditional likelihood is used for model selection of stochastic volatility models, the criterion selects overly complex models. In the context of hierarchical models, Millar (2009) also find that the conditional likelihood performs poorly with DIC. On the other hand, Berg et al. (2004) use DIC on a series of stochastic volatility models and find that in both real and simulated data, DIC produces the same model selections as classical Bayesian selection routines based on the MDD. However, they do not explicitly state which likelihood they use.

While I acknowledge the computational issues raised in Chan and Grant (2016), I use the conditional likelihood as it is the easiest and fastest to compute. As well, calculating the full integrated likelihood when both parameters and volatility can time-vary is computationally difficult, but see Chan and Eisenstat (2015) for an importance sampling algorithm. Moreover, in my applications, I find that while the original versions of WAIC and DIC prefer more complicated models, the alternative penalties offered by Gelman et al. (2004) and Vehtari

et al. (2016) prefer models with less time variation.

2.6 Empirical Work

In this section, I apply my new prior to two datasets. The first is a classical three-variable VAR from Primiceri (2005) and the second is a seven-variable VAR that was analyzed in Pettenuzzo et al. (2016). For the Primiceri data, I explore forecasting performance and impulse response functions, while in the other I only perform a forecasting exercise.

2.6.1 Primiceri Data

The data come from Dmitris Koribilis's website and consist of Quarterly Unemployment, Inflation and Interest Rates from 1953Q1 to 2006Q2, and all series are in levels. While Primiceri (2005) used a pre-sample period to estimate OLS parameters, I use the full sample for two reasons. One, this makes the expected value of the parameter at all times the OLS value, and two, so that the shrinkage priors shrink towards the full-sample OLS, rather than a pre-sample OLS. They also use 19 additional degrees of freedom (above dimension of 21) on the prior for the error covariance matrix, while I use 2 in order to remain uninformative.

Forecasting

D'Agostino et al. (2013) applied the Primiceri (2005) method to their three-dimensional dataset and found that the TVP model slightly outperformed random-walk benchmarks, and the gains were larger at longer horizons. The forecasting gains were largest in predicting Inflation, though all series showed some improvement.

That is my starting point as well. I apply both classes of TVP models to an expanding window of data, starting with an initial 25-year period. The first window runs from 1953Q1 - 1973Q1, and the last window is from 1953Q1-2006Q2. There are therefore 104 quarters for prediction.

At each window, I estimate 16 total models. For each of the IW and Separation classes

I use the relative threshold prior in addition to a series of absolute threshold priors. All absolute threshold priors are 90% threshold priors, where the thresholds are from a grid of 6 possible values (10, 1, 0.01, 10^{-6}, 10^{-10} and 10^{-15}). For the Separation class I also use the MU prior, and for the IW class I also use the Primiceri prior. I then further select the optimal model based on each of the four information criteria. This exercise therefore reflects pseudo-results from real-time forecasting. In addition to the 16 TVP Models, I estimate 3 constant parameter benchmarks — a VAR(1), independent AR(1)s and a random walk.

For all models, I estimate a TVP-VAR(2) with stochastic volatility. In order to estimate the distribution, I generate 20,000 draws from the posterior and discard the first 10,000. I generate forecasts from 1-12 steps ahead, though only report statistics for horizons 1-4 quarters ahead since forecast quality deteriorates as horizon increases. For forecast evaluation, I report both point, interval and density forecast statistics. The point forecast statistics are in Table 22, while interval and density forecast statistics are in Table 23.

For the point forecasts, I record the root mean-squared-error (RMSE) relative to the random-walk benchmark. I also record the geometric mean of the relative RMSEs and the log-determinant (ln-det) of forecast errors. The ln-det measure was first proposed by Doan et al. (1983) and has gained popularity in the DSGE forecasting literature. As described by Del Negro and Schorfheide (2004), the ln-det statistic is the natural logarithm of the determinant of the error covariance matrix of the forecasts. The determinant of the covariance matrix is the product of eigenvalues, which in turn are the variances of the uncorrelated forecast errors. Therefore, the ln-det can be seen as the average in the improvements for the individual variables, adjusted to take into account the joint forecasting performance. In my table, I present the exponential of the difference of each ln-det with the benchmark (random walk), so the units are percentage improvement (or deterioration) of the forecasts, where 1 is the benchmark.

For the interval forecasts, I present the empirical 90% forecast interval coverage probability in addition to its average length. For density forecasts, I present the average log-predictive

likelihood.

In order to get a sense of how each of the IC perform, I present the model selection results from the first window (first 100 quarters) in Table 21. For these values, I generate 75,000 draws from the posterior and discard the first 25,000 as a burn-in sample, leaving 50,000 to calculate the likelihood and penalty values.

Table 21: Information Criteria based on first 100 observations of Primiceri Data

DIC and WAIC values applied to eight different priors for each of the Separation and IW class of priors. The right panel presents the Monte-Carlo average (negative) log-likelihood, its variance and the average posterior variance of Q_β.

Prior	DIC-1	DIC-2	WAIC-1	WAIC-2	$-l(y\|\theta)$	$Var(l(y\|\theta))$	$mean(diag(Q_\beta))$
Separation							
MU	-2127	25028	-1878	906361	-1188	6912.81	0.01
R	-2163	27549	**-1895**	887116	**-1197**	7543.26	0.01
10	-2057	27663	-1892	898594	-1195	7596.03	0.02
1	**-2182**	28123	-1885	901528	-1195	7680.04	0.02
0.01	-2149	28451	-1878	917934	-1190	7765.78	0.02
10^{-6}	-756	10728	-448	690976	-394	2887.33	1.1E-4
10^{-10}	-250	-2	-259	349512	-175	112.09	2.0E-9
10^{-15}	-229	**-26**	-259	**349001**	-175	111.37	2.6E-14
Inverse-Wishart							
P	-276	496	-265	361704	-190	244.51	4.2E-6
R	-1854	27381	-1501	900403	-1051	7432.64	0.22
10	-1593	52439	-1251	1060854	-975	13686.83	2.97
1	-1749	34959	-1488	932404	**-1052**	9354.13	0.30
0.01	**-2129**	29124	**-1537**	1007809	-1050	7798.31	0.01
10^{-6}	-269	694	-246	351939	-181	287.75	8.0E-7
10^{-10}	-231	**-18**	-260	346168	-175	112.97	8.1E-11
10^{-15}	-247	-14	-260	**345687**	-175	109.42	4.9E-15

There is a rather large break between the strict and lax penalties for each IC. For both DIC and WAIC, when the penalty term is based on the variance of the likelihood, they prefer absolute thresholds 10^{-10} and 10^{-15}. On the other hand, the more lax IC prefer models with substantially more time variation (absolute thresholds 1, 0.01, and the relative threshold). Comparing across classes of priors, both DIC and WAIC-1 all prefer Separation models over the Inverse-Wishart options, whereas WAIC-2 prefers Inverse-Wishart.

There are also large breaks in the Information Criterion values between absolute thresholds 0.01 and 10^{-6}, and between thresholds 10^{-6} and 10^{-10}. The breaks in the IC are due to the factors displayed on the right-hand panel of Table 21. For those priors, both the average and variance of the log-likelihood drop significantly. This drop is mainly driven by error covariance of the time-varying parameters, Q_β, which correspondingly shrinks.

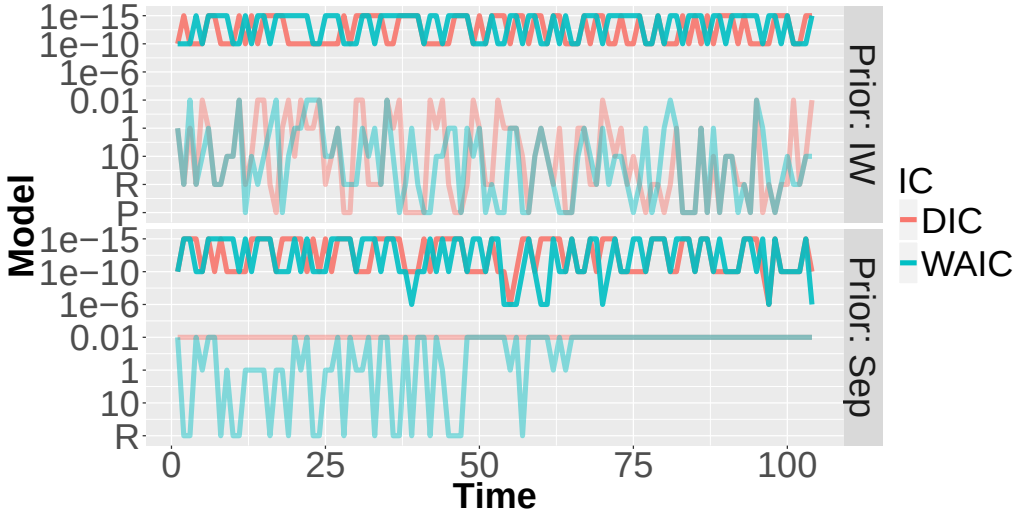
This major theme — that DIC-1 and WAIC-1 prefer more complex models while DIC-2 and WAIC-2 prefer simpler models — is not confined to these particular first 100 quarters of data. Indeed, observe Figure 23. For all forecasting periods, there is a clear divide between the two sets of information criteria, and little differentiation between DIC and WAIC when the penalty term is similar (based on variance of likelihood or not). When the penalty is based on the variance of the likelihood, the information criteria almost exclusively prefer models with absolute threshold 10^{-10} or 10^{-16}. On the other hand, when the penalty is not based on the variance term, the information criteria broadly select from the complex models which higher absolute thresholds. Indeed, based on Table 21, the models with high absolute thresholds are fairly similar in terms of likelihood value and average posterior variance of time-varying parameters.

Moving on to the point forecasts of each model, I present Table 22, which displays the point forecast statistics of models chosen by each of the IC, and Table 29 (in the Appendix) of each of the individual priors. The discrepancies between information criteria in terms of which models are selected not surprisingly also effects forecasting performance. The strict information criteria, which mostly select the priors using absolute thresholds 10^{-10} and 10^{-15} do very well, and substantially beat the benchmarks, whereas the lax criteria, which select models with more time variation do much worse than the benchmarks. Further, the Inverse-Wishart class apparently generates parameters that imply a nonstationary system, as forecasts for DIC-1 and WAIC-1 for the Inverse-Wishart prior become explosive when the forecasting horizon exceeds one quarter.

As mentioned, the strict IC improve forecasts, often substantially. Similar to D'Agostino

Figure 23: Model selection in each of the expanding windows for Primiceri Forecast

The top panel displays selected models from the IW class, while the bottom panel displays those for the Separation class. The opaque lines represent the strict penalties (DIC-2 and WAIC-2), whereas the translucent lines represent the lax penalties (DIC-1 and WAIC-1).



et al. (2013), I find that the forecast gains are largest for unemployment, though interest rate forecasts also improve by about 5% at all horizons. Inflation forecasts also improve by about 5%, but only at horizon 1. At all other horizons, the forecasts underperform the benchmark. Comparing across prior classes, the Separation class forecasts better than the Inverse-Wishart class for almost all horizons and series. More importantly, the Separation prior improves performance according to the multivariate statistics, which indicates that the point forecasts errors are both smaller and less correlated.

It is also important to compare the information criteria forecasting performance with the individual prior forecasting performance. While it is difficult to say whether the IC select the correct (lowest MSE) model at each forecasting period, it is simple to compare IC forecasting results with the underlying model performances in Table 29. I will focus on the performance of DIC-2 and WAIC-2, which select between the low time-variation models, and also perform best. For the Inverse-Wishart class, DIC-2 and WAIC-2 select between absolute thresholds 10^{-10} and 10^{-15}, and the selected models do better than the full-sample results for either of those priors individually. On the other hand, for the Separation class, DIC-2 and WAIC-2

Table 22: RMSE from forecasting Primiceri Data with Model Selection

Forecasting results for models chosen by each of the Model Selection criteria. At the end of each quarter, a model is selected by each of the four IC, and used to forecast 1-4 quarters ahead. All results are relative to Random Walk forecasts. AVG is the average relative RMSEs and ln-det is log-determinant of forecast error covariance. For brevity, I use only the first letter of each of the two information criteria (D for DIC and W for WAIC). Values larger than 1000 are replaced with a dash. The values in column RW display the units.

	Inverse-Wishart				Separation				Constant		
	D1	D2	W1	W2	D1	D2	W1	W2	VAR	AR	RW
h=1											
INFL	1.11	0.95	1.00	0.96	1.10	0.94	1.10	0.94	1.13	1.02	0.33
UNEMP	0.90	0.88	0.93	0.88	0.93	0.86	0.92	0.87	1.05	1.03	0.31
INTRT	1.01	0.97	0.98	0.97	1.01	0.98	1.02	0.97	1.06	1.04	1.02
AVG	1.01	0.93	0.97	0.94	1.01	0.93	1.01	0.93	1.08	1.03	0.55
ln-det	1.28	0.74	1.01	0.76	1.26	0.70	1.26	0.72	1.25	1.21	-4.93
h=2											
INFL	16	1.07	20	1.08	1.08	1.06	1.08	1.06	1.20	1.03	0.59
UNEMP	7.85	0.84	14	0.84	0.92	0.82	0.92	0.82	1.03	1.04	0.54
INTRT	6.47	0.97	6.82	0.97	1.01	0.96	1.02	0.96	1.08	1.05	1.41
AVG	10	0.96	13	0.96	1.01	0.95	1.01	0.95	1.10	1.04	0.84
ln-det	—	0.76	—	0.78	1.13	0.72	1.14	0.73	1.26	1.35	-2.20
h=3											
INFL	22	1.15	25	1.15	1.08	1.12	1.09	1.13	1.25	1.05	0.81
UNEMP	13	0.84	22	0.84	0.99	0.82	0.99	0.82	0.98	1.05	0.74
INTRT	6.9	0.95	8.77	0.95	0.99	0.94	0.99	0.94	1.10	1.05	1.67
AVG	14	0.98	18	0.98	1.02	0.96	1.02	0.96	1.11	1.05	1.08
ln-det	—	0.87	—	0.88	1.30	0.82	1.31	0.82	1.14	1.47	-0.68
h=4											
INFL	—	1.24	—	1.24	1.13	1.20	1.13	1.20	1.29	1.06	1.02
UNEMP	—	0.84	—	0.83	1.09	0.81	1.10	0.80	0.92	1.06	0.92
INTRT	829	0.96	879	0.96	1.03	0.94	1.03	0.94	1.11	1.04	1.95
AVG	—	1.01	—	1.01	1.08	0.98	1.09	0.98	1.11	1.05	1.30
ln-det	—	1.03	—	1.04	1.87	0.94	1.97	0.93	1.11	1.58	0.39

also mostly select between absolute thresholds 10^{-10} and 10^{-15}, but there are some periods when they select absolute threshold 10^{-6}. The absolute threshold 10^{-6} prior forecasts much worse than the other (smaller) time-variation models, and perhaps that poor forecasting performance effects the IC performance. Both IC underperform the absolute thresholds 10^{-10} and 10^{-15} priors by 1%-5%, depending on the metric.

Next, I present the interval and density forecasts from the models selected by the information

criteria in Table 23, with the results for all 16 candidate models in Table 30 in the Appendix. In general, the models (and correspondingly, the information criteria) that perform better in terms of point forecasts perform worse in terms of interval forecasts. The first set of information criteria (DIC-1 and WAIC-1) have empirical hit probabilities between 85% and 95% for their 90% coverage intervals, which is very good. On the other hand, the second set of information criteria (DIC-2 and WAIC-2) have hit probabilities only around 35%. Of the models that produce good interval forecasts (high variance individual priors, DIC-1, and WAIC-1), the separation class produces forecasts that are better calibrated (both in terms of empirical hit probability and interval length) than their IW counterparts.

However, the second set of information criteria outperform the first set in terms of density forecasts. Moreover, the Separation class performs better in terms of density forecasts than the Inverse-Wishart class. However, all TVP models significantly underperform all the constant parameter benchmarks. This is not surprising though. Since the predictive distribution is Gaussian, the estimated in-sample error covariance matrix will be underestimated in the case of a TVP model, since the model, by definition, will over-fit in sample. Thus, when making predictions, the true realization will often be far from the predictive distribution, therefore decreasing the predictive likelihood. The density forecasts of the IC are also better than their full-sample analogs, which indicates that while the IC perform slightly worse in terms of point forecasts, they perform better for density forecasts.

Thus, to summarize the findings, I have shown that the TVP models can generate large gains to point forecasts, but not interval or density forecasts. Moreover, the point forecasts generated by the Separation class of priors outperforms their Inverse-Wishart counterparts at nearly all forecast horizons and forecast evaluation metrics. On the other hand, the models that perform worse in terms of point forecasts generate interval forecasts that match their proposed hit probability. The Separation class also substantially beats their Inverse-Wishart counterparts in terms of density forecasts, though all TVP models perform much worse than the constant parameter benchmarks.

Table 23: Interval and Density Forecasts for Primiceri Data with Model Selection

Empirical 90% interval forecast hit-probabilities and average length (in brackets) of the interval forecast. Log-score metrics are relative to the RW density forecasts (negative means underperform benchmark).

	Inverse-Wishart				Separation				Constant		
	D1	D2	W1	W2	D1	D2	W1	W2	VAR	AR	RW
					h=1						
INFL	0.84	0.35	0.86	0.34	0.93	0.34	0.93	0.36	0.65	0.77	0.76
	[5.66]	[0.25]	[7.39]	[0.25]	[1.16]	[0.24]	[1.2]	[0.26]	[0.58]	[0.75]	[0.76]
UNEMP	0.88	0.34	0.85	0.33	0.9	0.33	0.9	0.34	0.57	0.55	0.61
	[5.59]	[0.23]	[7.28]	[0.23]	[1.22]	[0.23]	[1.26]	[0.25]	[0.58]	[0.75]	[0.76]
INTRT	0.84	0.32	0.81	0.32	0.71	0.29	0.72	0.33	0.88	0.9	0.9
	[6.42]	[0.48]	[7.97]	[0.45]	[1.49]	[0.44]	[1.51]	[0.47]	[0.58]	[0.75]	[0.76]
log-score	-2.0E6	-355	-2.7E6	-334	-9.6E4	-144	-1.1E5	-149	-0.07	-0.11	0.00
					h=2						
INFL	0.85	0.34	0.87	0.34	0.94	0.36	0.95	0.36	0.37	0.56	0.59
	[45]	[0.58]	[66]	[0.57]	[2.28]	[0.57]	[2.36]	[0.61]	[0.41]	[0.51]	[0.53]
UNEMP	0.91	0.4	0.87	0.44	0.88	0.44	0.89	0.46	0.37	0.38	0.42
	[44]	[0.48]	[64]	[0.49]	[2.17]	[0.48]	[2.25]	[0.51]	[0.41]	[0.51]	[0.53]
INTRT	0.8	0.28	0.83	0.29	0.66	0.25	0.67	0.29	0.73	0.81	0.85
	[47]	[0.82]	[67]	[0.79]	[2.62]	[0.76]	[2.68]	[0.8]	[0.41]	[0.51]	[0.53]
log-score	-2.5E8	-1434	-3.5E8	-1487	-2.2E5	-649	-2.8E5	-677	-0.86	-0.36	0.00
					h=3						
INFL	0.86	0.41	0.94	0.42	0.95	0.44	0.95	0.45	0.28	0.42	0.43
	[521]	[0.95]	[800]	[0.95]	[3.64]	[0.94]	[3.81]	[0.99]	[3.18]	[3.36]	[3.49]
UNEMP	0.89	0.42	0.87	0.45	0.92	0.46	0.93	0.46	0.35	0.24	0.31
	[501]	[0.72]	[775]	[0.73]	[3.51]	[0.73]	[3.68]	[0.77]	[3.18]	[3.36]	[3.49]
INTRT	0.81	0.29	0.81	0.28	0.77	0.27	0.78	0.31	0.63	0.71	0.74
	[473]	[1.17]	[712]	[1.13]	[4.09]	[1.09]	[4.24]	[1.14]	[3.18]	[3.36]	[3.49]
log-score	-5E10	-2393	-7E10	-2169	-4.9E5	-908	-5.9E5	-936	-1.85	-0.72	0.00
					h=4						
INFL	0.85	0.41	0.93	0.42	0.94	0.45	0.94	0.47	0.23	0.38	0.38
	[7001]	[1.35]	[10890]	[1.35]	[5.15]	[1.32]	[5.44]	[1.39]	[0.58]	[0.75]	[0.76]
UNEMP	0.88	0.43	0.88	0.43	0.93	0.47	0.94	0.5	0.32	0.18	0.22
	[6571]	[0.93]	[10389]	[0.95]	[4.91]	[0.94]	[5.18]	[1.01]	[0.58]	[0.75]	[0.76]
INTRT	0.8	0.28	0.79	0.27	0.81	0.3	0.81	0.33	0.58	0.66	0.66
	[5816]	[1.49]	[8946]	[1.45]	[5.82]	[1.39]	[6.09]	[1.45]	[0.58]	[0.75]	[0.76]
log-score	-1E13	-4720	-2E13	-4750	-9.7E5	-1971	-1.2E6	-2065	-3.08	-1.22	0.00

Impulse Response Functions

The primary motivation for Primiceri (2005) was to use a TVP model to examine how the economy changed (or did not change) from the 1950s to 2000s. One of their main methods was to compute impulse response functions over various time periods and determine whether they change. They find that the IRFs do not change very much, which leads them to conclude that despite changes in the parameters, those changes did not effect the overall economy

very much. In this section I perform the same analysis and attempt to answer whether Primiceri (2005)'s conclusions were dependent of their choice of prior or not.

It should be noted that despite the long tradition of using TVP models for Impulse Response Functions, they cannot be interpreted as the IRF at date t conditional on information at that time. Since the estimation procedure involves sampling the underlying parameters using the Kalman Smoother, the IRFs are actually conditional on full-sample information.

I run the Gibbs samplers and present both the time-paths of parameters and the Impulse Response Functions. For all priors I make 1.5 million draws from the posterior distribution and discard the first 500,000 as burn-in. The posterior samplers take around 5 hours to complete all the draws. I compute Impulse Response Functions at dates 1975Q1 and 1996Q1. Throughout, I compute the responses of Inflation and Unemployment to a unit shock in Interest Rates.

In order to decide which models to run in full, I consult both information criteria and point forecasting performance from Table 29. I consider the same 16 candidate models, and present IC results in Table 24. In order to calculate the information criteria, I generate 75,000 draws from the posterior and discard the first 25,000 as a burn-in sample, leaving 50,000 to calculate the likelihood and penalty values.

The table confirms many of the themes from above. First of all, the separation class is much less sensitive to choice of hyperparameter than the IW class. All thresholds larger than 10^{-6} (including MU and R) have likelihood values within 20 points of each other, and all the information criteria score them approximately the same. The strict information criteria (penalty functions based on the Monte Carlo variance of the likelihood) all prefer the models with very small thresholds (10^{-10} or 10^{-15}), while the more lax information criteria often prefer some of the more complex models. The information criteria can also be applied between classes, and universally prefer the Separation class to the IW class. There are once again large breaks in the IC values between thresholds 0.01 and 10^{-6} and between 10^{-6} and

Table 24: DIC and WAIC for Full Primiceri dataset

DIC and WAIC values applied to eight different priors for each of the Separation and IW class of priors. The right panel presents the Monte-Carlo average (negative) log-likelihood, its variance and the average posterior variance of Q_β.

Prior	DIC-1	DIC-2	WAIC-1	WAIC-2	$-l(y\|\theta)$	$Var(l(y\|\theta))$	$mean(diag(Q_\beta))$
Separation							
MU	-4009	51955	-3051	933554	-2155	14141.88	0.01
R	**-4068**	46820	**-3094**	916792	**-2165**	12852.58	0.01
10	-4037	49972	-3040	941328	-2151	13634.98	0.01
1	-4008	58237	-3071	932502	-2162	15719.06	0.01
0.01	-4019	50002	-3056	945593	-2155	13650.91	0.01
10^{-6}	-1802	54746	-1097	759422	-966	14202.47	2.7E-4
10^{-10}	-674	**-303**	-746	375108	-449	204.75	6.0E-9
10^{-15}	-697	-281	-745	**375106**	-448	203.95	1.8E-14
Inverse-Wishart							
P	-1356	20970	-1013	606306	-763	5666.91	3.0E-4
R	-3326	67592	-2444	875181	-1967	18033.24	0.13
10	-2664	105322	-1470	1062951	-1642	27306.15	2.12
1	-3397	76226	-2425	874301	-1959	20166.11	0.17
0.01	**-3612**	42221	**-2742**	1076689	**-1989**	11641.29	3.6E-3
10^{-6}	-1234	21394	-944	588570	-713	5752.62	1.4E-4
10^{-10}	-1100	11877	-917	572663	-667	3361.86	1.0E-4
10^{-15}	-878	**2395**	-873	**485156**	-566	945.34	2.5E-5

10^{-10}, and this again seems to be driven by a sharp decrease in in the Monte-Carlo variance of the likelihood function and the posterior variance of time-varying parameters.

In Figure 26 (in the Appendix), I plot the time-paths of some of the samplers (relative threshold, absolute threshold 0.01 and 10^{-10} for both Separation and IW classes, and the MU and Primiceri priors).

The IW class of priors are nicely nested in terms of time-variation of parameters. The absolute threshold at 10^{-10} suppresses just about any time-variation and the parameters are largely constant over time. The Primiceri prior allows some time-variation, but the parameters are still well centered at the unconditional mean. The 0.01 threshold allows considerably more time-variation, where now parameters freely move far away from their unconditional mean. The relative threshold allows slightly more time-variation.

On the other hand, the Separation priors do not show similar nesting patterns. All three of the MU, relative threshold and absolute threshold at 0.01 show the same amount of time-variation, and the absolute threshold at 10^{-10} shrinks parameters to remain constant. This result is consistent with the simulation and IC results above — that if the prior time-variation is sufficiently large compared to the likelihood time-variation, then all separation priors will return the likelihood time-variation.

Turning to the IRFs, Primiceri (2005) found that while the underlying parameters vary, the corresponding IRFs do not. I find that these results are rather sensitive to prior choice.

For the Inverse-Wishart class, I plot IRFs for the Primiceri prior (as a benchmark), and absolute thresholds 0.01, 10^{-10} and 10^{-15}. Absolute threshold 10^{-15} is selected by both DIC-2 and WAIC-2, and performs the very well in terms of forecasting. Absolute threshold 0.01 is selected by DIC-1 and WAIC-1, and absolute threshold 10^{-10} also performs well in the point forecasting exercise.

Starting with the Inverse-Wishart class of priors (plotted in Figure 24), I confirm that for Primiceri's specification of the prior, the IRFs do not display much time-variation. However, the absolute threshold priors do show much more time variation than the Primiceri version. For the Inflation response, both the absolute thresholds at 0.01 and 10^{-15} show substantial time-variation after 10 quarters. There are also large deviations in the effect to Unemployment for the 10^{-15} absolute threshold prior at horizons 2-15. Lastly, the absolute threshold at 0.01 displays substantial time-variation at horizons 1-4 for both Inflation and Unemployment (but also IRFs that are inconsistent with the SVAR literature), and a convergence as the horizon grows.

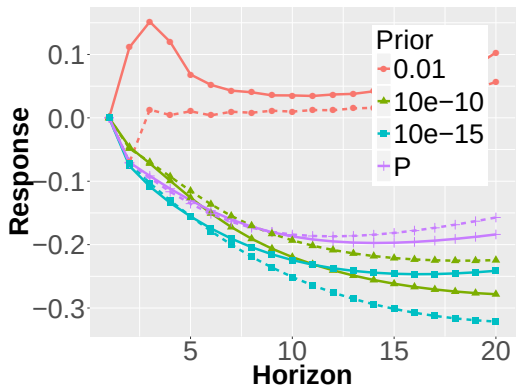
For the Separation class, I plot IRFs for the MU prior as a benchmark, the relative threshold prior, and absolute threshold priors at 10^{-10} and 10^{-15}. I use the Relative threshold prior because it is selected by DIC-1 and WAIC-1, and use the other threshold priors because they are selected by DIC-2 and WAIC-2, and perform the best in the point forecasting exercise.

For the Separation priors (plotted in Figure 25), the effect of time-variation depends on the choice of prior (or equivalently, a choice between the the second of first version of DIC and WAIC). The relative threshold and MU priors (selected by WAIC-1 and DIC-1) display nearly the exact same responses over all horizons, but there is substantial time-variation. On the other hand, the absolute threshold priors at 10^{-10} and 10^{-15} (selected by DIC-2 and WAIC-2) are exactly the same both across time and over all horizons. These results are consistent with the time-plots of coefficients in the Appendix. While MU and the relative threshold display substantial time-variation in parameters, the more strict absolute thresholds display (visually) constant parameters. The IRFs further suggest that with increased time-variation (MU and R priors), some dates might have parameters that imply an explosive system. One could correct for that using the adapted sampler developed in Cogley and Sargent (2005) and Cogley (2005) that does not allow explosive parameters. Another option could be to discard posterior draws that imply an explosive system, but that might be computationally infeasible since there is a parameter vector for every time period.

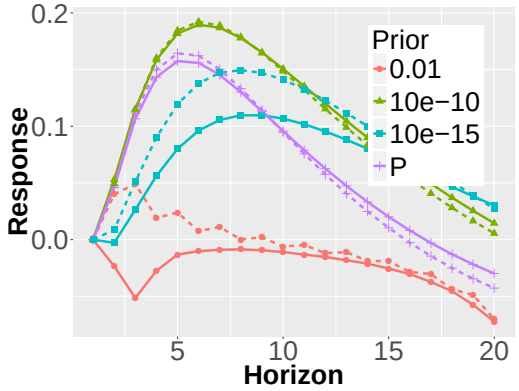
Figure 24: IRFs for IW Class of Priors, Response to Unit Shock in Interest Rates

All plots show Impulse Response Functions to a unit shock in Interest Rates. The solid line depicts the IRF at 1975Q1, while the dashed line is for 1996Q1. The color and shapes represent different priors within the IW class.

(a) Inflation



(b) Unemployment

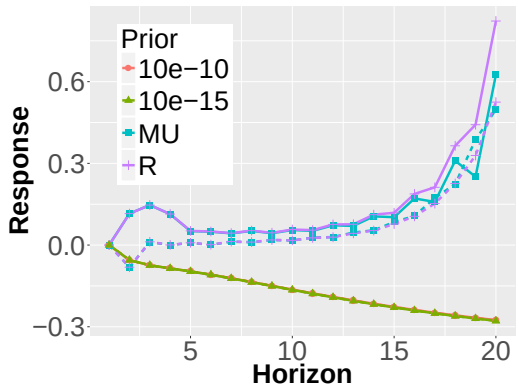


In terms of economic interpretations of the IRFs, the Inverse-Wishart priors show a sustained reduction in Inflation and an increase in Unemployment that dies out after about 20 quarters. Primiceri (2005) found a short price puzzle in the response of Inflation to a monetary policy

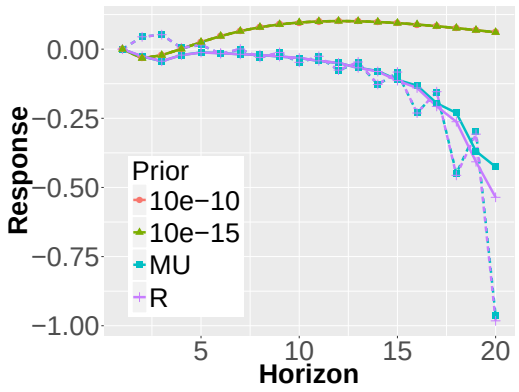
Figure 25: IRFs for Separation Class of Priors, Response to Unit Shock in Interest Rates

All plots show Impulse Response Functions to a unit shock in Interest Rates. The solid line depicts the IRF at 1975Q1, while the dashed line is for 1996Q1. The color and shapes represent different priors within the Separation class.

(a) Inflation



(b) Unemployment



shock, so that issue has apparently been mitigated by the two changes I made. On the other hand, he found that unemployment drops for 2-3 quarters after a shock before rising, whereas all priors except absolute thresholds 10^{-15} and 0.01 increase immediately following the shock. For the Separation priors (with small absolute thresholds), there is also a sustained reduction in Inflation, while unemployment drops for 2-3 quarters after a shock before rising, which is the same as in Primiceri (2005).

Primiceri (2005)'s broad conclusions were that nonlinearities in the parameter vector did not contribute to changes in the overall economy. By computing the IRFs for the other priors, I see that his conclusions are somewhat sensitive to his choice in prior. Had he stayed within the Inverse-Wishart class of prior, he would have seen substantially more variation over time with one of the absolute threshold priors. On the other hand, if Primiceri had used one of the separation priors (with small absolute threshold), he would have found no time-variation in IRFs.

2.6.2 Pettenuzzo Data

Next, I also apply my method to a larger system, which is (part of) the dataset analyzed in Pettenuzzo et al. (2016). They were introducing a method called compression, which is a shrinkage method that randomly selects regressors and averages resulting predictions together. They then apply the method to a medium sized VAR (19 variables) large VAR (46 variables) and huge VAR (129 variables), and have an extension to time-varying parameters as in Koop and Korobilis (2013). The TVP model in Koop and Korobilis (2013) uses a simplifying assumption called variance discounting, which does not use MCMC. In addition, it also requires a deterministic time-path for the Kalman Filter measurement error covariance, so it does not produce optimal (MSE) filtered values. All VARs are estimated with up to 13 lags. Despite the size of the regressions, they are mainly interested the performance on seven key variables — (1) Total non-farm payroll (PAYEMS, change in log), (2) Consumer price inflation (CPIAUCSL, change in log), (3) Change in Federal Funds Rate (FEDFUNDS, first difference), (4) Industrial Production growth (INDPRO, change in log), (5)Unemployment Rate (UNRATE, level), (6) Producer good price inflation (WPSFD49207, change in log), and (7) Change in 10 year T-bill rate (GS10, first difference).

Without the shrinkage of compression and the simplifications of Koop and Korobilis (2013)'s method, I compute a much smaller seven-variable VAR with one lag, so there are 56 time-varying parameters. All data is available monthly, and I use data from August 1954 through November 2016, which makes 749 total forecasting periods. This dataset is therefore substantially larger than the previous Primiceri (2005) dataset.

Due to computational considerations, I estimate all covariance matrices in a pre-sample window of 150 months, and then use an expanding window to filter for the regression coefficients and stochastic volatility holding the covariance matrices constant at the posterior mean. I generate 45,000 draws to estimate the covariance matrices, discarding the first 20,000 as burn in, and use 1000 draws to update the regression coefficients, discarding the first 500. As before, I use the same 16 priors to forecast, and select between them using the four model

selection criteria.

Once again, in order to get a sense of the information criteria, I perform the model-selection routine on the pre-sample window of 150 observations. The model selection results are in Table 25, and the overall results are similar to those from the Primiceri dataset. As before, I use the same eight candidate priors for each of the classes. The priors for the separation class produce largely similar likelihood values, as long as the prior threshold is large enough. This again supports the idea that for large priors, the separation prior produces results that are not as sensitive to choice of prior hyperparameter. There is also the same result that the information criteria based on the variance of the likelihood (strict IC) prefer the models with very small thresholds, while the ones not based on the variance prefer models with more time-variation. For this dataset, the Inverse-Wishart relative threshold prior induces less time-variation in parameters than its Separation counterpart.

Table 25: Information Criteria based on first 150 observation of Pettenuzzo data

DIC and WAIC values applied to eight different priors for each of the Separation and IW class of priors. The right panel presents the Monte-Carlo average (negative) log-likelihood, its variance and the average posterior variance of Q_β.

Prior	DIC-1	DIC-2	WAIC-1	WAIC-2	$-l(y\|\theta)$	$Var(l(y\|\theta))$	$mean(diag(Q_\beta))$
Separation							
MU	-7787	74191	-7095	328986	-4356	20956.66	0.02
R	-7827	57697	-7194	335543	**-4377**	16844.20	0.02
10	-7762	53510	-7121	339903	-4356	15792.78	0.02
1	-7685	58480	-7114	337335	-4337	17036.27	0.02
0.01	**-7894**	53903	-7165	332589	-4352	15854.68	0.02
10^{-6}	-7634	6837	**-7607**	204746	-4167	3967.44	6.0E-5
10^{-10}	-7408	-5776	-7481	179339	-3919	622.98	1.2E-8
10^{-15}	-7341	**-6529**	-7421	**168684**	-3859	391.15	5.5E-13
Inverse-Wishart							
P	-7556	-5669	-7560	176430	-3959	653.18	0.26
R	-7361	**-6613**	-7423	**167764**	-3858	364.80	3.2E-14
10	-5219	154382	-3903	470158	-3154	40445.08	5.8
1	-6229	126506	-5141	405408	-3616	33685.47	0.71
0.01	-7683	52990	-7022	328213	**-4278**	15604.94	0.01
10^{-6}	**-7779**	-2560	**-7785**	195938	-4144	1559.64	1.3E-6
10^{-10}	-7355	-6430	-7426	170185	-3865	418.62	2.3E-10
10^{-15}	-7358	-6531	-7422	168307	-3858	385.68	6.1E-14

However, while for the Primiceri data the information criteria were split based on the penalty parameters (DIC-2 and WAIC-2 only selected absolute thresholds 10^{-10} or 10^{-15}, while DIC-1 and WAIC-1 ranged between the high-variance models, though settling on absolute threshold 0.01 for the Separation class), the same is not true for the Pettenuzzo data. In Table 26, I present the distribution of selected models for each of the information criteria. There is very little consistency across the different classes, where the only similarity is that WAIC-2 selects absolute threshold 10^{-6} about 90% of the time for both classes. For the Inverse-Wishart class, DIC-1 and WAIC-1 select absolute threshold 10^{-6} 100%, while DIC-2 is evenly split between the relative threshold and absolute thresholds 10^{-10} and 10^{-15}. For the Separation class, DIC-1 most often selects the relative threshold (47%), though also absolute threshold 0.01 (14%). DIC-2 selects absolute threshold 10^{-15} over 90% of the time, while WAIC-1 is evenly split between absolute thresholds 10^{-6} and 10^{-10}.

Table 26: Distribution of selected models for Pettenuzzo Data

Distribution of priors selected by each of the information criteria over all forecasting periods. Numbers may not add up to 1 due to rounding.

	Inverse-Wishart				Separation			
	DIC-1	DIC-2	WAIC-1	WAIC-2	DIC-1	DIC-2	WAIC-1	WAIC-2
P/MU	0.00	0.03	0.00	0.05	0.05	0.00	0.00	0.00
R	0.00	0.34	0.00	0.01	0.47	0.00	0.00	0.00
10	0.00	0.00	0.00	0.00	0.18	0.00	0.00	0.00
1	0.00	0.00	0.00	0.00	0.09	0.00	0.00	0.00
0.01	0.00	0.00	0.00	0.00	0.14	0.00	0.00	0.00
10^{-6}	1.00	0.00	1.00	0.92	0.01	0.00	0.44	0.90
10^{-10}	0.00	0.33	0.00	0.01	0.06	0.09	0.56	0.04
10^{-15}	0.00	0.30	0.00	0.01	0.00	0.91	0.00	0.05

The model selection tables in Table 25 and Table 26 show that there is either substantially more time-variation in the dataset, or that the information criteria require more data and/or posterior draws to obtain better estimates of the information criteria.

Nonetheless, I continue with the forecasting results. For the point forecasts, I again report RMSE for each of the seven series individually (relative to AR(1) benchmarks), average RMSE and the log-determinant of forecast error covariance (relative to AR(1) benchmark).

The point forecasting results are in Table 27. For interval and density forecasts, I again record 90% coverage rate and log-predictive scores, which is in Table 28. I also record the forecast performance for each of the priors individually (in the Appendix), with point forecasts in Table 31 and interval/density forecasts in Table 32.

Next I move on to interval and density forecasts, which are in Table 28 for models selected by information criteria, and Table 32 (in the Appendix) for all 16 models. Since results are largely similar to those from the Primiceri dataset, I only present one-step ahead interval and density forecasting results. As before, models that produced better point forecasts produce worse interval forecasts. One major difference is that for this dataset, the constant parameter models are not as well calibrated, and their empirical coverage rate is essentially zero for all variables. This is also likely due to dimensionality. As the number of predictions increases, the empirical hit probability decreases.

While in the case of the Primiceri (2005) data, the Separation class almost always outperformed the Inverse-Wishart class, this dataset presents a somewhat different story. First of all, forecasting gains (for both classes and all information criteria) are mainly concentrated in the forecasts for PPI. For example, at horizon 1, while the other series show forecasting improvements and declines of around 2%-5%, forecasts of PPI improve by 13%. These values are also consistent with improvements by the VAR over the AR benchmark. But while both TVP classes (paired with DIC-2 Information Criterion) outperform the VAR in terms of average RMSE and log-determinant, in this dataset, the Inverse-Wishart class outperform the Separation class. In addition, while DIC-2 and WAIC-2 forecasted very similarly for the Primiceri data, they perform quite differently this time. Again, this is evident from Table 26 where they select between the different thresholds very differently.

Pettenuzzo et al. (2016) made significant forecasting gains over the benchmark in most series and at most horizons, and their gains were substantially larger than mine. That means that the forecasting gains they found were likely due to a combination of the longer lag-length or the additional variables. Moreover, they also found that including time-variation (via

Table 27: Relative RMSE of forecasts in Pettenuzzo et al. (2016) data

RMSE for 7-dimensional VAR(1), for the variables of interest from Pettenuzzo et al. (2016). Each RMSE is recorded relative to an AR(1) benchmark. I also display the mean of all relative RMSEs and the ln-det statistic. The TVP models considered are selected via information criteria

	Inverse-Wishart				Separation				Constant	
	D1	D2	W1	W2	D1	D2	W1	W2	VAR	AR
h=1										
PAYEMS	1.08	1.01	1.08	1.08	3.67	1.03	1.05	1.16	1.03	0.00
CPI	0.97	0.97	0.97	0.97	1.11	0.97	0.99	0.99	1.00	0.52
FEDFUNDS	1.02	1.00	1.02	1.02	1.19	1.00	1.04	1.06	1.03	0.29
IP	1.03	0.97	1.03	1.03	1.57	0.96	1.08	1.09	0.98	0.01
UNRATE	0.96	1.00	0.96	0.96	5.26	1.00	1.03	1.03	1.02	0.00
PPI	0.89	0.87	0.89	0.89	1.04	0.87	0.92	0.95	0.87	0.18
GS10	1.08	0.99	1.08	1.07	1.93	1.02	1.06	1.15	1.04	0.01
Average	1.01	0.97	1.01	1.00	2.25	0.98	1.02	1.06	1.00	0.14
ln-det	1.32	0.65	1.32	1.28	—	0.72	1.69	2.62	0.85	1.00
h=2										
PAYEMS	1.02	1.05	1.02	1.02	—	1.07	0.98	1.13	1.01	0.00
CPI	0.97	0.97	0.97	0.97	1.24	0.97	1.00	1.00	1.00	0.57
FEDFUNDS	1.01	0.99	1.01	1.01	1.39	0.99	1.03	1.06	1.03	0.31
IP	1.05	0.97	1.05	1.04	—	0.96	1.08	1.08	0.97	0.01
UNRATE	0.96	1.02	0.96	0.95	—	1.03	0.99	0.97	1.03	0.00
PPI	0.83	0.81	0.83	0.83	1.03	0.81	0.85	0.90	0.80	0.28
GS10	1.03	0.98	1.03	1.03	—	0.99	1.00	1.09	1.00	0.01
Average	0.98	0.97	0.98	0.98	—	0.97	0.99	1.03	0.98	0.17
ln-det	1.33	0.73	1.33	1.23	—	0.79	1.53	2.89	0.78	1.00
h=3										
PAYEMS	1.03	1.03	1.03	1.03	95.05	1.05	1.00	1.15	1.01	0.00
CPI	0.99	0.99	0.99	0.99	1.62	0.99	1.03	1.03	1.01	0.57
FEDFUNDS	1.02	1.00	1.02	1.02	2.18	1.00	1.04	1.06	1.02	0.30
IP	1.05	0.99	1.05	1.04	37.76	0.98	1.08	1.06	0.99	0.01
UNRATE	0.95	1.04	0.95	0.95	181.43	1.04	1.13	1.22	1.02	0.00
PPI	0.81	0.80	0.81	0.80	2.1	0.80	0.83	0.88	0.78	0.39
GS10	1.05	0.98	1.05	1.04	60.92	0.98	1.02	1.12	1.00	0.01
Average	0.98	0.97	0.98	0.98	54.44	0.98	1.02	1.07	0.98	0.18
ln-det	1.48	0.83	1.48	1.42	—	0.88	2.69	6.02	0.79	1.00
h=4										
PAYEMS	1.00	1.03	1.00	1.00	721.1	1.02	0.99	1.13	1.01	0.00
CPI	1.00	0.99	1.00	1.00	16.68	0.99	1.05	1.05	1.02	0.56
FEDFUNDS	1.02	1.00	1.02	1.02	22.56	1.00	1.06	1.09	1.01	0.30
IP	1.07	0.98	1.07	1.07	316.29	0.98	1.11	1.12	1.00	0.01
UNRATE	0.99	1.05	0.99	0.99	981.92	1.04	1.07	1.15	1.04	0.00
PPI	0.79	0.79	0.79	0.79	10.52	0.80	0.80	0.86	0.77	0.49
GS10	1.05	0.99	1.05	1.05	411.08	0.99	1.03	1.14	1.00	0.01
Average	0.99	0.97	0.99	0.99	354.31	0.97	1.02	1.08	0.98	0.20
ln-det	1.54	0.80	1.54	1.48	—	0.78	2.65	6.12	0.78	1.00

Table 28: Interval and Density Forecasts for Pettenuzzo Data with Model Selection

Empirical 90% interval forecast hit-probabilities and average length (in brackets) of the interval forecast. Log-score metrics are relative to the RW density forecasts (negative means underperform benchmark).

	Inverse-Wishart				Separation				Constant	
	D1	D2	W1	W2	D1	D2	W1	W2	VAR	AR
PAYEMS	0.83	0.23	0.83	0.79	0.98	0.19	0.81	0.94	0	0
	[0.01]	[0]	[0.01]	[0.01]	[0.44]	[0]	[0.02]	[0.02]	[0]	[0]
CPI	0.35	0.24	0.35	0.35	0.77	0.24	0.35	0.54	0.52	0.66
	[0.16]	[0.11]	[0.16]	[0.17]	[0.8]	[0.11]	[0.26]	[0.32]	[0]	[0]
FEDFUNDS	0.27	0.12	0.27	0.26	0.59	0.12	0.27	0.44	0.34	0.36
	[0.16]	[0.08]	[0.16]	[0.16]	[0.61]	[0.08]	[0.22]	[0.34]	[0]	[0]
IP	0.8	0.26	0.8	0.78	0.98	0.25	0.69	0.91	0.01	0.01
	[0.02]	[0]	[0.02]	[0.02]	[0.44]	[0]	[0.02]	[0.03]	[0]	[0]
UNRATE	0.95	0.35	0.95	0.93	1	0.32	0.94	0.95	0	0
	[0.01]	[0]	[0.01]	[0.01]	[0.46]	[0]	[0.02]	[0.03]	[0]	[0]
PPI	0.35	0.2	0.35	0.36	0.81	0.2	0.35	0.55	0.2	0.28
	[0.14]	[0.08]	[0.14]	[0.14]	[0.6]	[0.08]	[0.17]	[0.26]	[0]	[0]
GS10	0.76	0.21	0.76	0.73	0.97	0.19	0.68	0.85	0.01	0.01
	[0.01]	[0]	[0.01]	[0.01]	[0.46]	[0]	[0.02]	[0.03]	[0]	[0]
log-score	-45	-23	-45	-43	-2.8E4	-9.79	-153	-215	-0.05	0

Koop and Korobilis (2013) further improved density forecasts while marginally improving point-forecasts. It would be interesting to see if the separation prior could be adapted into the Koop and Korobilis (2013) method to find if the forecasting gains would be any different.

2.7 Conclusion

In this paper I have introduced a new prior for TVP-VARs. Classically, Bayesian TVP-VAR models have used priors on covariances from the Inverse-Wishart distribution with scale matrices and degrees of freedom chosen in an ad-hoc fashion. Based on recent work in Bayesian Random Effects model, I adapt a prior that specifies information over variances and correlations separately.

This new class of priors, which I call Separation Priors, have many benefits over Inverse-Wishart priors, some of which have been studied before, and some which are new and especially applicable to TVP contexts.

First of all, this new class of priors was previously used mainly in order to be uninformative on both variances and correlations at the same time, but I show that it can also be used as a more intuitive *informative* prior as well. The separation prior has marginal distribution on variances as half-t distributions, as compared to the Inverse-Wishart, whose marginal distribution on variances are Inverse-Gamma. This distinction is important when it comes to describing prior information into a proper distribution. While Inverse-Gamma distributions bound variances away from zero, half-t distributions do not. Thus, if one has a prior that variances should not be "too big," there is no way to effectively describe that via an Inverse-Gamma distribution. The statement effectively becomes: "The variance should not be too big, but also not too small," which may not be the econometricians intention. On the other hand, the half-t distribution includes zero, so by changing the prior variance, the econometrician can easily translate prior information into the distribution.

Secondly, because the Separation prior encodes information about variance and correlation separately, the econometrician can describe nonsample information about each (almost) independently. When using the Inverse-Wishart prior, all information about the covariance is encoded at once, so it is impossible to encode substantial information about variance without encoding information about correlations. The Separation priors makes this possible.

Lastly, in the TVP context, we are usually concerned that the model will induce too much time-variation. In other words, practitioners think that constant parameter models are more likely to explain the data well. Indeed, the long tradition on inference for structural break models uses the null hypothesis of no-break. I show that the bounds away from zero are a problem not only for the prior, but also the posterior of a constant parameter model. Thus, an econometrician using an IW prior might conclude that the data indicates time-variation in the parameters, when really it is a product of the prior. Separation priors are better able to nest constant parameter models.

In addition to advantages from the Bayesian perspective, the Separation prior also exhibits positive frequentist properties. In a series of simulations, I show that the Separation priors

do a much better job estimating the true time-path of the time-varying parameters. I also apply the new prior to two forecasting datasets. While it substantially improves point-forecasts for a small model (2-variable VAR(2)), it slightly underperforms a larger system (7-variable VAR(1)). Lastly, the new prior is computationally tractable, as it requires merely an additional step to draw Inverse-Gamma random variables, which adds negligent computation time.

2.8 Appendix

2.8.1 Gibbs Samplers

- Primiceri

  1. Start with ν (degree of freedom for IW distribution)
  2. Take sample $t = 1, \dots \tau$ and estimate $\hat{\beta}_{OLS}$, $\mathbb{V}[\hat{\beta}_{OLS}]$, and $\hat{\sigma}_{OLS}$ (I use $\tau = T$)
  3. Set prior:

$$Q_\beta \sim IW(\tau, \kappa_\beta \mathbb{V}[\hat{\beta}_{OLS}])$$
$$H \sim IW(n+1, I_n)$$

  4. Gibbs Sample
     (a) Given Q_β, H and Σ_t, use Carter-Kohn to filter β_t
     (b) Given β_t, draw $Q_\beta \sim IW$
     (c) Given β_t, draw Σ_t using Stochastic Smoother (Kim et al. (1998))
     (d) Given β_t and Σ_t, draw $H \sim IW$

- Inverse Wishart

  1. Start with ν (degree of freedom for IW distribution)
  2. Take sample $t = 1, \dots \tau$ and estimate $\hat{\beta}_{OLS}$, $\mathbb{V}[\hat{\beta}_{OLS}]$, and $\hat{\sigma}_{OLS}$ (I use $\tau = T$)
  3. Use absolute or relative prior to find $a_k, k = 1, \dots K$
     - Absolute Threshold
       (a) Given thresholds, $t_1, t_2, \dots t_K$ and confidence levels $\alpha_1, \alpha_2, \dots \alpha_K$
       (b) Find $a_1, a_2, \dots a_K$ that best satisfies the threshold
     - Relative Threshold
       (a) Given $vec(\hat{\beta}_{OLS}) = b_1, b_2, \dots b_K$, time-change amounts τ_{b_k} and confidence levels, $\alpha_1, \alpha_2, \dots \alpha_K$
       (b) Calculate thresholds, $t_1, t_2, \dots t_K$

(c) Find $a_1, a_2, \ldots a_K$ that best satisfies the threshold

4. Set prior:

$$Q_\beta \sim IW(\nu, diag(a_1, a_2, \ldots a_K))$$
$$H \sim IW(n+1, I_n)$$

5. Gibbs Sample

(a) Given Q_β, H and Σ_t, use Carter-Kohn to filter β_t

(b) Given β_t, draw $Q_\beta \sim IW$

(c) Given β_t, draw Σ_t using Stochastic Smoother (Kim et al. (1998))

(d) Given β_t and Σ_t, draw $H \sim IW$

- Separation:

1. Take sample $t = 1, \ldots \tau$ and estimate $\hat{\beta}_{OLS}$, $\mathbb{V}[\hat{\beta}_{OLS}]$, and $\hat{\sigma}_{OLS}$ (I use $\tau = T$)

2. Set ν (prior information on correlation ($\nu > K + 1$)

3. Use absolute or relative prior to find $A_k, k = 1, \ldots K$

– Absolute Threshold

(a) Given thresholds, $t_1, t_2, \ldots t_K$ and confidence levels $\alpha_1, \alpha_2, \ldots \alpha_K$

(b) Find $A_1, A_2, \ldots A_K$ that best satisfies the threshold

– Relative Threshold

(a) Given $vec(\hat{\beta}_{OLS}) = b_1, b_2, \ldots b_K$, time-change amounts τ_{b_k} and confidence levels, $\alpha_1, \alpha_2, \ldots \alpha_K$

(b) Calculate thresholds, $t_1, t_2, \ldots t_K$

(c) Find $A_1, A_2, \ldots A_K$ that best satisfies the threshold

4. Set prior:

$$Q_\beta \sim IW(\nu, 2\nu diag(1/a_1, 1/a_2, \ldots 1/a_n))$$
$$a_k \sim IGamma(1/2, 1/A_k^2)$$

$$H \sim IW(n+1, I_n)$$

5. Gibbs Sample

   (a) Given Q_β, H and Σ_t, use Carter-Kohn to filter β_t

   (b) Given β_t, draw $Q_\beta \sim IW$

   (c) Given Q_β, draw $a_k \sim Gamma$

   (d) Given β_t, draw Σ_t using Stochastic Smoother (Kim et al. (1998))

   (e) Given β_t and Σ_t, draw $H \sim IW$

2.8.2 Simulation Designs

All simulations follow the DGP:

$$\begin{aligned} Y_t &= \beta_t Y_{t-1} + \varepsilon_t \qquad \varepsilon_t \sim N(0, \Sigma_t) \\ vec(\beta_t) &= vec(\beta_{t-1}) + \varepsilon_t^\beta \qquad \varepsilon_t^\beta \sim N(0, Q_\beta) \\ diag(\Sigma_t) &= diag(\Sigma_{t-1}) + \varepsilon_t^\sigma \qquad \varepsilon_t^\sigma \sim N(0, Q_\sigma) \end{aligned}$$

In all cases, where $d = dim(Y_t)$, $Q_\sigma = I_d \times 10^{-4}$

In simulation 1, $d = 2$, and

$$Q_\beta = I_6 \times 10^{-4}$$

In simulation 2, $d = 2$, and

$$Q_\beta = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \times 10^{-4}$$

In simulation 3, $d = 2$, and

$$Q_\beta = \begin{pmatrix} 1000.00 & 600.00 & 12.65 & 15.81 & 0.00 & 0.00 \\ 600.00 & 1000.00 & 9.49 & 0.00 & 0.00 & 0.00 \\ 12.65 & 9.49 & 1.00 & 0.00 & 0.00 & 0.00 \\ 15.81 & 0.00 & 0.00 & 1.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \end{pmatrix} \times 10^{-4}$$

In simulation 4, $d = 2$ and

$$Q_\beta = \begin{pmatrix} 1.00 & 0.60 & 0.40 & 0.50 & 0.00 & 0.00 \\ 0.60 & 1.00 & 0.30 & 0.00 & 0.00 & 0.00 \\ 0.40 & 0.30 & 1.00 & 0.00 & 0.00 & 0.00 \\ 0.50 & 0.00 & 0.00 & 1.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 1.00 & 0.80 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.80 & 1.00 \end{pmatrix} \times 10^{-4}$$

In simulation 5, $d = 4$ and

$$Q_\beta \sim IW(25, I_{20} \times 10^{-4}),$$

with $seed = 100$.

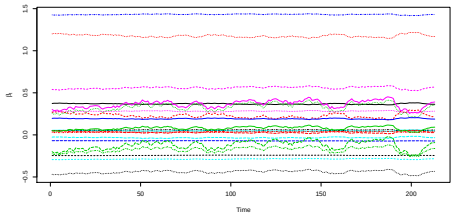
In simulation 6, $Q_\beta = 0_{20}$, the 20-dimensional zero matrix.

2.8.3 Additional Figures and Tables

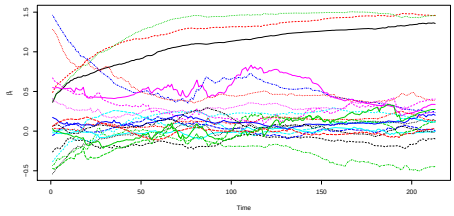
Figure 26: Time varying parameters in Primiceri Data

Posterior mean of time-varying parameters of Primiceri data. The time span ranges from 1953Q1 to 2006Q2.

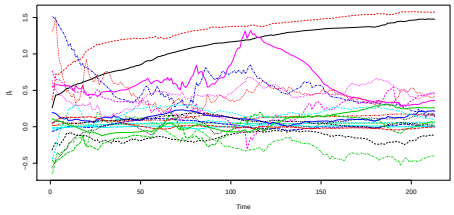
(a) Primiceri Prior



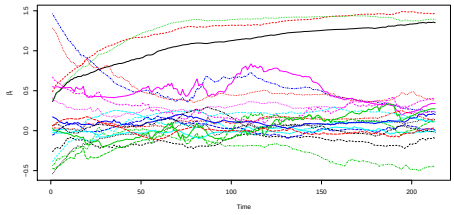
(b) MU Prior



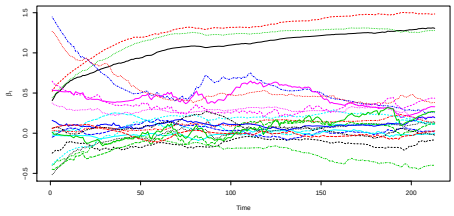
(c) IW Relative Prior



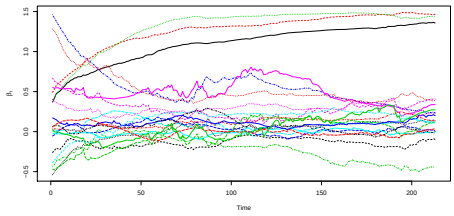
(d) Separation Relative Prior



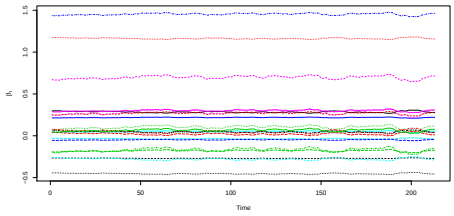
(e) IW Threshold (0.01) Prior



(f) Sep Threshold (0.01) Prior



(g) IW Threshold (10^{-10}) Prior



(h) Sep Threshold (10^{-10}) Prior

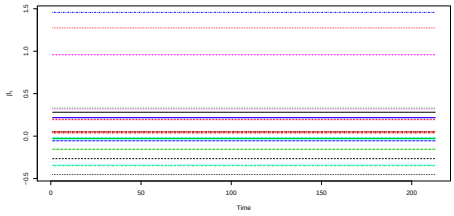


Table 29: RMSE from forecasting Primiceri data for all 16 models

Forecasting results for all 16 models considered above. All results are RMSE relative to Random Walk forecast. AVG is the average relative RMSEs and ln-det is log-determinant of forecast error covariance. Values larger than 1000 are replaced with a dash.

	Inverse-Wishart								Separation								Constant		
	h=1																		
	P	R	10	1	0.01	10^{-6}	10^{-10}	10^{-15}	MU	R	10	1	0.01	10^{-6}	10^{-10}	10^{-15}	VAR	AR	RW
INFL	0.96	1.03	1.06	1.09	1.09	0.95	0.95	0.96	1.09	1.09	1.09	1.09	1.10	0.98	0.93	0.93	1.13	1.02	0.33
UNEMP	0.88	0.92	0.92	0.93	0.92	0.86	0.88	0.88	0.93	0.92	0.92	0.92	0.93	0.89	0.86	0.86	1.05	1.03	0.31
INTRT	0.99	0.99	1.05	1.04	0.99	0.98	0.98	0.97	1.02	1.02	1.02	1.03	1.01	0.98	0.97	0.97	1.06	1.04	1.02
AVG	0.95	0.98	1.01	1.02	1.00	0.93	0.94	0.94	1.01	1.01	1.01	1.01	1.01	0.95	0.92	0.92	1.08	1.03	0.55
ln-det	0.73	1.05	1.28	1.30	1.18	0.72	0.75	0.75	1.25	1.26	1.25	1.27	1.26	0.84	0.68	0.68	1.25	1.21	1.00
	h=2																		
INFL	1.10	4.2	28	3.01	1.09	1.08	1.08	1.08	1.09	1.08	1.08	1.08	1.08	1.10	1.05	1.05	1.20	1.03	0.59
UNEMP	0.84	3.14	28	3.04	0.92	0.81	0.84	0.83	0.93	0.92	0.92	0.93	0.92	0.84	0.82	0.82	1.03	1.04	0.54
INTRT	1.03	2.17	14	1.94	1.00	0.98	0.98	0.96	1.02	1.02	1.02	1.02	1.01	0.99	0.96	0.96	1.08	1.05	1.41
AVG	0.99	3.17	23	2.66	1.00	0.96	0.96	0.96	1.01	1.01	1.01	1.01	1.01	0.98	0.94	0.94	1.10	1.04	0.84
ln-det	0.79	—	—	504	1.13	0.74	0.78	0.77	1.15	1.14	1.15	1.16	1.13	0.78	0.71	0.70	1.26	1.35	1.00
	h=3																		
INFL	1.19	4.78	45	4.53	1.08	1.16	1.16	1.15	1.09	1.09	1.09	1.09	1.08	1.11	1.12	1.12	1.25	1.05	0.81
UNEMP	0.83	3.99	42	4.72	0.95	0.81	0.84	0.83	0.99	0.99	0.99	0.99	0.99	0.83	0.82	0.82	0.98	1.05	0.74
INTRT	1.03	2.76	16	1.97	0.99	0.97	0.97	0.95	0.99	0.99	0.99	1.00	0.99	0.99	0.94	0.94	1.10	1.05	1.67
AVG	1.02	3.84	35	3.74	1.01	0.98	0.99	0.98	1.03	1.02	1.02	1.02	1.02	0.98	0.96	0.96	1.11	1.05	1.08
ln-det	0.87	—	—	—	1.19	0.83	0.92	0.87	1.34	1.34	1.33	1.33	1.30	0.76	0.81	0.80	1.14	1.47	1.00
	h=4																		
INFL	1.28	68	—	45	1.11	1.23	1.25	1.24	1.14	1.14	1.14	1.13	1.13	1.14	1.19	1.19	1.29	1.06	1.02
UNEMP	0.82	40	—	48	0.98	0.80	0.84	0.83	1.11	1.11	1.11	1.11	1.09	0.82	0.82	0.81	0.92	1.06	0.92
INTRT	1.08	19	—	20	0.99	0.98	0.98	0.96	1.04	1.03	1.03	1.04	1.03	1.01	0.94	0.94	1.11	1.04	1.95
AVG	1.06	42	—	38	1.03	1.00	1.02	1.01	1.10	1.09	1.09	1.09	1.08	0.99	0.98	0.98	1.11	1.05	1.30
ln-det	1.04	—	—	—	1.36	1.00	1.10	1.03	2.09	2.06	2.06	2.03	1.87	0.84	0.93	0.92	1.11	1.58	1.00
	h=8																		
INFL	1.54	—	—	—	1.40	1.39	1.45	1.48	2.41	2.45	2.61	2.39	2.17	1.20	1.38	1.39	1.52	1.10	1.64
UNEMP	0.70	—	—	—	1.23	0.75	0.77	0.78	3.34	3.24	3.27	3.24	2.74	0.76	0.75	0.75	0.68	1.07	1.40
INTRT	1.24	—	—	—	1.25	1.00	1.02	0.99	1.95	1.89	1.91	1.93	1.73	1.13	0.93	0.94	1.15	1.03	2.79
AVG	1.16	—	—	—	1.29	1.04	1.08	1.09	2.57	2.53	2.60	2.52	2.21	1.03	1.02	1.03	1.12	1.07	1.94
ln-det	1.46	—	—	—	3.38	1.30	1.36	1.38	124	116	133	112	78	0.95	1.03	1.03	0.86	1.62	1.00

Table 30: Interval and Density Forecasts for Primiceri Data

Empirical 90% interval forecast hit-probabilities and average length of the interval forecast. Log-score metrics are relative to the RW density forecasts (negative means underperform benchmark).

	Inverse Wishart								Constant		
	P	R	10	1	0.01	10^{-6}	10^{-10}	10^{-15}	VAR	AR	RW
IINFL	0.41	1	1	1	0.83	0.43	0.35	0.33	0.65	0.77	0.76
	[0.31]	[4.62]	[18]	[5.85]	[0.81]	[0.32]	[0.25]	[0.25]	[0.58]	[0.75]	[0.76]
UNEMP	0.44	1	1	1	0.81	0.38	0.33	0.33	0.57	0.55	0.61
	[0.32]	[3.66]	[18]	[5.87]	[0.83]	[0.3]	[0.23]	[0.23]	[0.58]	[0.75]	[0.76]
INTRT	0.51	0.97	1	0.96	0.65	0.38	0.34	0.31	0.88	0.9	0.9
	[0.97]	[6.4]	[18]	[5.93]	[1.33]	[0.58]	[0.48]	[0.45]	[0.58]	[0.75]	[0.76]
log-score	-3E3	-5E5	-6E6	-8E5	-7E4	-736	-984	-318	-0.07	-0.11	0
	Separation								Constant		
	MU	R	10	1	0.01	10^{-6}	10^{-10}	10^{-15}	VAR	AR	RW
INFL	0.93	0.92	0.93	0.93	0.93	0.49	0.34	0.33	0.65	0.77	0.76
	[1.23]	[1.23]	[1.23]	[1.23]	[1.16]	[0.45]	[0.24]	[0.24]	[0.58]	[0.75]	[0.76]
UNEMP	0.9	0.91	0.91	0.9	0.9	0.6	0.3	0.31	0.57	0.55	0.61
	[1.29]	[1.28]	[1.29]	[1.29]	[1.22]	[0.44]	[0.23]	[0.22]	[0.58]	[0.75]	[0.76]
INTRT	0.71	0.71	0.71	0.72	0.71	0.5	0.28	0.28	0.88	0.9	0.9
	[1.55]	[1.55]	[1.53]	[1.54]	[1.49]	[1.05]	[0.41]	[0.41]	[0.58]	[0.75]	[0.76]
log-score	-5E3	-1E5	-1E5	-1E5	-9E4	-800	-630	-282	-0.07	-0.11	0

Table 31: RMSE from forecasting Pettenuzzo data for all 16 models

Forecasting results for all 16 models considered above. All results are RMSE relative to Random Walk forecast. AVG is the average relative RMSEs and ln-det is log-determinant of forecast error covariance. Values larger than 1000 are replaced with a dash.

	Inverse-Wishart								Separation								Constant	
	P	R	10	1	0.01	10^{-6}	10^{-10}	10^{-15}	MU	R	10	1	0.01	10^{-6}	10^{-10}	10^{-15}	VAR	AR
h=1																		
PAYEMS	0.94	1.02	84	24	2.51	1.08	0.95	1.02	4.30	3.63	3.55	3.63	4.19	1.16	0.97	1.02	1.03	0.00
CPI	0.98	0.97	1.30	1.28	1.09	0.97	0.97	0.97	1.09	1.10	1.09	1.11	1.10	0.99	0.97	0.97	1.00	0.52
FEDFUNDS	1.01	1.00	1.45	1.34	1.16	1.02	1.00	1.00	1.21	1.20	1.19	1.20	1.19	1.06	1.01	1.00	1.03	0.29
IP	1.04	0.96	23	8.51	1.35	1.03	0.96	0.96	1.61	1.56	1.68	1.67	1.66	1.11	0.97	0.96	0.98	0.01
UNRATE	0.98	1.01	100	37	3.40	0.96	0.96	1.01	4.79	4.78	5.17	5.01	5.49	1.04	0.92	1.00	1.02	0.00
PPI	0.89	0.87	1.41	1.11	1.06	0.89	0.87	0.87	1.06	1.06	1.07	1.06	1.07	0.95	0.88	0.87	0.87	0.18
GS10	0.98	1.00	42	9.19	1.73	1.08	0.99	1.00	1.88	2.01	1.85	2.01	1.84	1.16	0.99	1.00	1.04	0.01
Average	0.98	0.98	36	12	1.76	1.01	0.96	0.98	2.28	2.19	2.23	2.24	2.36	1.07	0.96	0.98	1.00	0.14
ln-det	0.70	0.70	—	—	—	1.32	0.51	0.70	—	—	—	—	—	2.98	0.58	0.70	0.85	1.00
h=2																		
PAYEMS	0.93	1.08	—	—	—	1.02	0.96	1.08	—	—	—	—	—	1.13	0.93	1.08	1.01	0.00
CPI	0.97	0.97	—	—	1.18	0.97	0.97	0.97	1.3	1.28	1.22	1.26	1.26	1.00	0.97	0.97	1.00	0.57
FEDFUNDS	1.00	0.99	—	—	1.26	1.01	0.99	0.99	1.4	1.38	1.42	1.4	1.41	1.06	0.99	0.99	1.03	0.31
IP	0.99	0.97	—	—	—	1.05	0.96	0.97	—	—	—	—	—	1.11	0.97	0.97	0.97	0.01
UNRATE	0.98	1.04	—	—	—	0.96	0.97	1.04	—	—	—	—	—	1.00	0.93	1.03	1.03	0.00
PPI	0.82	0.81	—	—	1.04	0.83	0.81	0.81	1.07	1.06	1.08	1.04	1.07	0.90	0.81	0.81	0.80	0.28
GS10	0.96	1.00	—	—	—	1.03	0.97	1.00	10	—	10	10	10	1.10	0.95	1.00	1.00	0.01
Average	0.95	0.98	—	—	—	0.98	0.95	0.98	—	—	10	—	—	1.04	0.94	0.98	0.98	0.17
ln-det	0.57	0.79	—	—	—	1.33	0.55	0.78	—	—	—	—	—	3.35	0.57	0.78	0.78	1.00
h=3																		
PAYEMS	0.91	1.08	—	—	40.57	1.03	0.94	1.08	106	95	113	111	113	1.16	0.90	1.07	1.01	0.00
CPI	0.99	0.99	—	155	1.2	0.99	0.99	0.99	1.71	1.72	1.69	1.76	1.67	1.03	0.99	0.99	1.01	0.57
FEDFUNDS	1.01	1.00	—	223	1.39	1.02	1.00	1.00	2.73	1.99	2.81	2.1	2.37	1.07	1.00	1.00	1.02	0.30
IP	1.00	0.98	—	—	13	1.05	0.98	0.98	43	36	45	45	36	1.08	0.99	0.98	0.99	0.01
UNRATE	0.96	1.05	—	—	60	0.95	0.97	1.05	187	188	197	178	168	1.22	0.92	1.04	1.02	0.00
PPI	0.80	0.80	—	200	1.21	0.81	0.80	0.80	2.61	2.06	2.22	2.37	2.1	0.88	0.79	0.80	0.78	0.39
GS10	0.96	1.00	—	—	21	1.05	0.97	1.00	55	60	58	56	59	1.13	0.95	1.00	1.00	0.01
Average	0.95	0.98	—	—	19	0.98	0.95	0.98	57	55	60	56	54	1.08	0.93	0.98	0.98	0.18
ln-det	0.57	0.85	—	—	—	1.48	0.59	0.85	—	—	—	—	—	6.69	0.61	0.84	0.79	1.00

Table 32: Interval and Density Forecasts for Pettenuzzo Data

Empirical 90% interval forecast hit-probabilities and average length of the interval forecast. Log-score metrics are relative to the AR density forecasts (negative means underperform benchmark).

	Inverse Wishart								Constant	
	P	R	10	1	0.01	10^{-6}	10^{-10}	10^{-15}	VAR	AR
PAYEMS	0.42	0.17	1	1	1	0.83	0.37	0.17	0	0
	[0]	[0]	[9.7]	[3.11]	[0.32]	[0.01]	[0]	[0]	[0]	[0]
CPI	0.5	0.23	1	0.98	0.77	0.35	0.24	0.24	0.52	0.66
	[0.25]	[0.11]	[9.76]	[3.17]	[0.7]	[0.16]	[0.11]	[0.11]	[0]	[0]
FEDFUNDS	0.25	0.12	1	0.98	0.61	0.27	0.12	0.13	0.34	0.36
	[0.15]	[0.08]	[9.75]	[3.12]	[0.58]	[0.16]	[0.08]	[0.08]	[0]	[0]
IP	0.64	0.23	1	1	1	0.8	0.3	0.24	0.01	0.01
	[0.01]	[0]	[9.79]	[3.1]	[0.32]	[0.02]	[0]	[0]	[0]	[0]
UNRATE	0.8	0.27	1	1	1	0.95	0.53	0.27	0	0
	[0]	[0]	[9.66]	[3.11]	[0.32]	[0.01]	[0]	[0]	[0]	[0]
PPI	0.52	0.2	1	1	0.78	0.35	0.2	0.19	0.2	0.28
	[0.22]	[0.08]	[9.85]	[3.13]	[0.49]	[0.14]	[0.08]	[0.08]	[0]	[0]
GS10	0.33	0.17	1	1	1	0.76	0.25	0.18	0.01	0.01
	[0]	[0]	[9.67]	[3.11]	[0.32]	[0.01]	[0]	[0]	[0]	[0]
log-score	-66.36	-11.6	-2E6	-5E5	-1E4	-45.06	-38.5	-23.43	-0.05	0
	Separation								Constant	
	MU	R	10	1	0.01	10^{-6}	10^{-10}	10^{-15}	VAR	AR
PAYEMS	1	1	1	1	1	0.99	0.65	0.17	0	0
	[0.5]	[0.44]	[0.5]	[0.5]	[0.49]	[0.03]	[0]	[0]	[0]	[0]
CPI	0.81	0.81	0.82	0.8	0.8	0.56	0.25	0.24	0.52	0.66
	[0.84]	[0.85]	[0.85]	[0.85]	[0.84]	[0.35]	[0.11]	[0.11]	[0]	[0]
FEDFUNDS	0.62	0.62	0.63	0.62	0.62	0.47	0.15	0.12	0.34	0.36
	[0.65]	[0.65]	[0.65]	[0.65]	[0.65]	[0.36]	[0.08]	[0.08]	[0]	[0]
IP	1	1	1	1	1	0.94	0.53	0.23	0.01	0.01
	[0.5]	[0.43]	[0.5]	[0.5]	[0.49]	[0.04]	[0.01]	[0]	[0]	[0]
UNRATE	1	1	1	1	1	1	0.84	0.28	0	0
	[0.5]	[0.47]	[0.5]	[0.5]	[0.49]	[0.03]	[0]	[0]	[0]	[0]
PPI	0.85	0.85	0.85	0.85	0.85	0.58	0.22	0.2	0.2	0.28
	[0.64]	[0.63]	[0.64]	[0.63]	[0.62]	[0.27]	[0.09]	[0.08]	[0]	[0]
GS10	1	1	1	1	1	0.91	0.5	0.17	0.01	0.01
	[0.5]	[0.48]	[0.5]	[0.5]	[0.49]	[0.03]	[0.01]	[0]	[0]	[0]
log-score	-2E4	-2E4	-2E4	-1E4	-2E4	-226.52	-14.02	-9.36	-0.05	0

BIBLIOGRAPHY

Aguilar, O. and M. West (2000): "Bayesian Dynamic Factor Models and Portfolio Allocation," *Journal of Business & Economic Statistics*, 18, 338–357.

Ait-Sahalia, Y., P. Mykland, and L. Zhang (2005a): "How Often to Sample a Continuous-Time Process in the Presence of Market Microstructure Noise," *Review of Financial Studies*, 18, 351–416.

Alvarez, I., J. Niemi, and M. Simpson (2016): "Bayesian Inference for a Covariance Matrix," .

Amir-Ahmadi, P., C. Mathes, and M.-C. Wang (2016): "Choosing Prior Hyperparameters," Federal reserve bank of richmond working paper series.

Amisano, G., D. Giannone, and M. Lenza (2015): "Large time varying parameter VARs for macroeconomic forecasting," Online.

Andersen, T., T. Bollerselv, and F. Diebold (2007): *Handbook of Financial Econometrics*, Amsterdam; North Holland, chap. Parametric and Nonparametric Volatility Measurement.

Andersen, T., T. Bollerslev, F. Diebold, and P. Labys (2001): "The Distribution of Realized Exchange Rate Volatility," *Journal of the American Statistical Association*, 96.

Andersen, T. G., T. Bollerslev, F. X. Diebold, and P. Labys (2003): "Modeling and Forecasting Realized Volatility," *Econometrica*, 71, 579–625.

Andersen, T. G., T. Bollerslev, F. X. Diebold, and J. Wu (2006): "Realized beta: Persistence and predictability," *Advances in Econometrics: Econometric Analysis of Economic and Financial Time Series in Honor of R.F. Engle and C.W.J. Granger*, B, 1–40.

Bai, J. and S. Ng (2002): "Determining the Number of Factors in Approximate Factor Models," *Econometrica*, 70, 191–221.

——— (2006): "Evaluating Latent and Observed Factors in Macroeconomics and Finance," *Journal of Econometrics*, 507–537.

Bali, T. G. and R. F. Engle (2010): "The intertemporal capital asset pricing model with dynamic conditional correlations," *Journal of Monetary Economics*, 57, 377 – 390.

Bali, T. G., R. F. Engle, and Y. Tang (2013): "Dynamic Conditional Beta is Alive and Well in the Cross-Section of Daily Stock Returns," KoÃ§ University-TUSIAD Economic Research Forum Working Papers 1305, Koc University-TUSIAD Economic Research Forum.

Barigozzi, M. and M. Hallin (2014): "Generalized Dynamic Factor Models and Volatilities. Recovering the Market Volatility Shocks," Working Papers ECARES ECARES 2014-52, ULB – Universite Libre de Bruxelles.

Barnard, J., R. McCulloch, and X. Meng (2000): "Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage," *Statistica Sinica*, 10, 1281–1311.

Barndorff-Nielsen, O. E., P. R. Hansen, A. Lunde, and N. Shephard (2011): "Multivariate realised kernels: Consistent positive semi-definite estimators of the covariation of equity prices with noise and non-synchronous trading," *Journal of Econometrics*, 162, 149–169.

Barndorff-Nielsen, O. E. and N. Shephard (2004): "Econometric Analysis of Realized Covariation: High Frequency Based Covariance, Regression, and Correlation in Financial Economics," *Econometrica*, 72, 885–925.

Belmonte, M. A., G. Koop, and D. Korobilis (2014): "Hierarchical Shrinkage in Time-Varying Parameter Models," *Journal of Forecasting*, 33, 80–94.

Benati (n.d.): "How Fast Can Advanced Economies Grow?" .

Berg, A., R. Meyer, and J. Yu (2004): "Deviance Information Criterion for Comparing Stochastic Volatility Models," *Journal of Business & Economic Statistics*, 22, 107–120.

Braun, P. A., D. B. Nelson, and A. M. Sunier (1995): "Good News, Bad News, Volatility, and Betas," *The Journal of Finance*, 50, 1575–1603.

Carriero, A., G. Kapetanios, and M. Marcellino (2012): "Forecasting government bond yields with large Bayesian vector autoregressions," *Journal of Banking & Finance*, 36, 2026 – 2047.

Chan, J. and E. Eisenstat (2015): "Bayesian model comparison for time-varying parameter VARs with stochastic volatility," Cama working papers, Centre for Applied Macroeconomic Analysis, Crawford School of Public Policy, The Australian National University.

Chan, J. C. and A. L. Grant (2014): "Issues in Comparing Stochastic Volatility Models Using the Deviance Information Criterion," CAMA Working Papers 2014-51, Centre for Applied Macroeconomic Analysis, Crawford School of Public Policy, The Australian National University.

——— (2016): "Fast computation of the deviance information criterion for latent variable models," *Computational Statistics & Data Analysis*, 100, 847 – 859.

Chen, Z. and R. Petkova (2012): "Does Idiosyncratic Volatility Proxy for Risk Exposure?" *The Review of Financial Studies*, 30.

Chib, S. (1995): "Marginal Likelihood from the Gibbs Output," *Journal of the American Statistical Association*, 90, 1313–1321.

Christoffersen, P., A. Lunde, and K. Olesen (2014): "Factor Structure in Commodity Futures Return and Volatility," *CREATES Research Papers*.

Cogley, T. (2005): "How fast can the new economy grow? A Bayesian analysis of the evolution of trend growth," *Journal of Macroeconomics*, 27, 179 – 207.

Cogley, T. and T. Sargent (2002): "Evolving Post-World War II U.S. Inflation Dynamics," Working Papers 2132872, Department of Economics, W. P. Carey School of Business, Arizona State University.

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, 262–302.

D'Agostino, A., L. Gambetti, and D. Giannone (2013): "Macroeconomic forecasting and structural change," *Journal of Applied Econometrics*, 28, 82–101.

Del Negro, M. and F. Schorfheide (2004): "Priors from General Equilibrium Models for VARS," *International Economic Review*, 45, 643–673.

——— (2011): *Bayesian Macroeconometrics*, Oxford University Press, 293–389.

Diebold, F. and M. Nerlove (1989): "The Dynamics of Exchange Rate Volatility: A Multivariate Latent Factor ARCH Model," *Journal of Applied Econometrics*, 4, 1–21.

Doan, T., R. B. Litterman, and C. A. Sims (1983): "Forecasting and Conditional Projection Using Realistic Prior Distributions," Working Paper 1202, National Bureau of Economic Research.

Duarte, J., A. Kamara, S. Siegel, and C. Sun (2014): "The Systematic Risk of Idiosyncratic Volatility," Manuscript.

Eisenstat, E., J. C. Chan, and R. W. Strachan (2014): "Stochastic Model Specification Search for Time-Varying Parameter VARs," CAMA Working Papers 2014-23, Centre for Applied Macroeconomic Analysis, Crawford School of Public Policy, The Australian National University.

Ferson, W. E. and C. R. Harvey (1993): "The Risk and Predictability of International Equity Returns," *Review of Financial Studies*, 6, 527–566.

Frühwirth-Schnatter, S. and A. Bitto (2016): "Achieving Shrinkage in a Time-Varying Parameter Model Framework," Tech. rep.

Frühwirth-Schnatter, S. and H. Wagner (2010): "Stochastic model specification search for Gaussian and partial non-Gaussian state space models," *Journal of Econometrics*, 154, 85 – 100.

Gelman, A. (2006): "Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper)," *Bayesian Analysis*, 1, 515–534.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2004): *Bayesian Data Analysis*, CRC Texts in Statistical Science, Chapman and Hall, second ed.

Gelman, A. and J. Hill (2007): *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Analytical Methods for Social Research, Cambridge University Press.

Gelman, A., J. Hwang, and A. Vehtari (2014): "Understanding predictive information criteria for Bayesian models," *Statistics and Computing*, 24, 997–1016.

Geweke, J. F. (1999): "Using simulation methods for bayesian econometric models: inference, development and communication," *Econometric Reviews*, 18, 1–126.

Giannone, D., M. Lenza, and G. E. Primiceri (2012): "Prior Selection for Vector Autoregressions," Working Papers ECARES ECARES 2012-002, ULB – Universite Libre de Bruxelles.

Hastie, T., R. Tibshirani, and J. Friedman (2001): *The Elements of Statistical Learning*, Springer Series in Statistics, New York, NY, USA: Springer New York Inc.

Hautsch, N., L. Kuj, and P. Malec (2011): "The Merit of High-Frequency Data in Portfolio Allocation," .

Hautsch, N. and L. Kyj (2009): "A Blocking and Regularization Approach to High Dimensional Realized Covariance Estimation," .

Herbst, E. P. and F. Schorfheide (2016): *Bayesian Estimation of DSGE Models*, no. 10612 in Economics Books, Princeton University Press.

Herskovic, B., B. T. Kelly, H. Lustig, and S. V. Nieuwerburgh (2014): "The Common Factor in Idiosyncratic Volatility: Quantitative Asset Pricing Implications," Working Paper 20076, National Bureau of Economic Research.

Huang, A. and M. P. Wand (2013): "Simple Marginally Noninformative Prior Distributions for Covariance Matrices," *Bayesian Anal.*, 8, 439–452.

Jacquier, E., N. G. Polson, and P. E. Rossi (1994): "Bayesian Analysis of Stochastic Volatility Models," *Journal of Business & Economic Statistics*, 12, 371–389.

Jagannathan, R. and Z. Wang (1996): "The Conditional CAPM and the Cross-Section of Expected Returns," *The Journal of Finance*, 51, 3–53.

Kalnina, I. and K. Tewou (2015): "Cross-Sectional Dependence in Idiosyncratic Volatility," .

Kim, S., N. Shephard, and S. Chib (1998): "Stochastic Volatility: Likelihood Inference and Comparison with ARCH Models," *The Review of Economic Studies*, 65, 361–393.

Koop, G. and D. Korobilis (2013): "Large time-varying parameter {VARs}," *Journal of Econometrics*, 177, 185 – 198, dynamic Econometric Modeling and Forecasting.

Lewellen, J. and S. Nagel (2006): "The conditional CAPM does not explain asset-pricing anomalies," *Journal of Financial Economics*, 82, 289–314.

Lindley, D. V. and A. F. M. Smith (1972): "Bayes Estimates for the Linear Model," *Journal of the Royal Statistical Society. Series B (Methodological)*, 34, 1–41.

Liu, L., A. Patton, and K. Sheppard (2015): "Does Anything Beat 5-Minute RV?" *Journal of Econometrics*, 187, 293–311.

Marden, J. I. (2015): *Multivariate Statistics*, CreateSpace Independent Publishing Platform.

Menictas, M. and M. Wand (2013): "Variational inference for marginal longitudinal semiparametric regression," *Stat*, 2, 61–71.

Millar, R. B. (2009): "Comparison of Hierarchical Bayesian Models for Overdispersed Count Data using DIC and Bayes' Factors," *Biometrics*, 65, 962–969.

Newton, M. A. and A. E. Raferty (1994): "Approximate Bayesian Inference by the Weighted Likelihood Bootstrap," *Journal of the Royal Statistical Society, Series B*, 56, 3–48.

Nyblom, J. (1989): "Testing for the Constancy of Parameters Over Time," *Journal of the American Statistical Association*, 84, 223–230.

O'Malley, A. J. and A. M. Zaslavsky (2008): "Domain-Level Covariance Analysis for Multilevel Survey Data With Structured Nonresponse," *Journal of the American Statistical Association*, 103, 1405–1418.

Onatski, A. (2009): "Testing Hypotheses About the Number of Factors in Large Factor Models," *Econometrica*, 77, 1447–1479.

Pettenuzzo, D., G. Koop, and D. Korobilis (2016): "Bayesian Compressed Vector Autoregressions," Working Papers 103, Brandeis University, Department of Economics and International Businesss School.

Pitt, M. and N. Shephard (1999): *Time-Varying Covariance: A Factor Stochastic Volatility Approach*, Oxford University Press, chap. 6.

Primiceri, G. E. (2005): "Time Varying Structural Vector Autoregressions and Monetary Policy," *The Review of Economic Studies*, 72, 821–852.

Ritter, C. and M. A. Tanner (1992): "Facilitating the Gibbs Sampler: The Gibbs Stopper and the Griddy-Gibbs Sampler," *Journal of the American Statistical Association*, 87, 861–868.

Sheppard, K. and W. Xu (2014): "Factor High-Frequency Based Volatility (HEAVY) Models," .

Sims, C., D. Waggoner, and T. Zha (2008): "Methods for inference in large multiple-equation Markov-switching models," *Journal of Econometrics*, 146, 255–274.

Sims, C. A. (1993): "A Nine-Variable Probabilistic Macroeconomic Forecasting Model," in *Business Cycles, Indicators and Forecasting*, National Bureau of Economic Research, Inc, NBER Chapters, 179–212.

——— (2001): "COMMENT ON SARGENT AND COGLEYâS âEVOLVING US POST-WAR INFLATION DYNAMICSâ," *NBER Macroeconomics Annual*, 373–379.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. Van Der Linde (2002): "Bayesian measures of model complexity and fit," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64, 583–639.

Stevanovic, D. (2010): "Common Sources of Parameter Instability in Macroeconomic Models: A Factor-TVP Approach," .

Stevanovic, D. and P. Amir-Ahmadi (2015): "Common Sources of Instabilities in Macroeconomic Dynamics," Tech. rep., ESG-UQAM.

Stock, J. H. and M. W. Watson (1996a): "Asymptotically Median Unbiased Estimation of Coefficient Variance in a Time Varying Parameter Model," NBER Technical Working Papers 0201, National Bureau of Economic Research, Inc.

——— (1996b): "Evidence on Structural Instability in Macroeconomic Time Series Relations," *Journal of Business & Economic Statistics*, 14, 11–30.

Vehtari, A., A. Gelman, and J. Gabry (2016): "Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC," *Statistics and Computing*, 1–20.

Wand, M. P., J. T. Ormerod, S. A. Padoan, and R. Frühwirth (2011): "Mean Field Variational Bayes for Elaborate Distributions," *Bayesian Anal.*, 6, 847–900.

Watanabe, S. (2010): "Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory," *J. Mach. Learn. Res.*, 11, 3571–3594.