Vector autoregression models with skewness and heavy tails

Sune Karlsson[a], Stepan Mazur[a,b] and Hoang Nguyen[a]

[a]Unit of Statistics, School of Business, Örebro University, Sweden

[b]Department of Economics and Statistics, School of Business and Economics, Linnaeus University, Sweden

December 1, 2022

Abstract

With uncertain changes of the economic environment, macroeconomic downturns during recessions and crises can hardly be explained by a Gaussian structural shock. There is evidence that the distribution of macroeconomic variables is skewed and heavy tailed. In this paper, we contribute to the literature by extending a vector autoregression (VAR) model to account for more realistic assumptions on the multivariate distribution of macroeconomic variables. We propose a general class of generalized hyperbolic skew Student's t distribution with stochastic volatility for the innovations in the VAR model that allows us to take into account both skewness and heavy tails. Tools for Bayesian inference and model selection using a Gibbs sampler are provided. In an empirical study, we present evidence of skewness and heavy tails for monthly macroeconomic variables. The analysis also gives a clear message that skewness is a value-added feature to VAR models with heavy tails.

JEL-codes: C11, C15, C16, C32, C52

Keywords: Vector autoregression; Skewness and heavy tails; Generalized hyperbolic skew Student's t distribution; Stochastic volatility; Markov Chain Monte Carlo

1 Introduction

Since the seminal work of Sims (1980), the vector autoregression (VAR) model has become one of the key macroeconomic models for policy makers and forecasters, see Karlsson (2013). The utility of the basic VAR model of Sims has been greatly enhanced by extensions allowing for time-varying parameters (Primiceri, 2005; Cogley and Sargent, 2005) and stochastic volatility (SV) (Uhlig, 1997; Clark, 2011; Clark and Ravazzolo, 2015). These can, however, not fully account for features in the data such as heavy tailed or skewed distributions.

Acemoglu et al. (2017) gives a theoretical motivation for the non-Gaussian distribution of macroeconomic variables and the presence of heavy tails and asymmetries is well documented in the literature. For example, Christiano (2007) found evidence against Gaussianity by inspecting the skewness and kurtosis properties of residuals from a Gaussian VAR model and Fagiolo et al. (2008) find that the distribution of the output growth rates of OECD countries can be approximated by symmetric exponential-power densities with Laplace tails even after accounting for outliers, autocorrelation and heteroscedasticity. To model the heavy tails, Ni and Sun (2005) propose a VAR model with a multivariate Student's t distribution, while Cúrdia et al. (2014) and Chib and Ramamurthy (2014) impose a similar heavy tailed structural shock in Dynamic Stochastic General Equilibrium (DSGE) models. Karlsson and Mazur (2020), on the other hand, propose a general class of multivariate heavy tailed distributions which includes the normal, t and Laplace distributions as well as their mixture for the innovations in VAR models. Stochastic volatility can also lead to a heavy tailed marginal distribution as in Cross and Poon (2016), Chiu et al. (2017), Liu (2019) and Carriero et al. (2020).

As noted by, among others, Cúrdia et al. (2014) the largest shocks occur during recessions, and the skewness of the distribution should be taken into account. Skew-normal and skew-t distributions are common choices for modelling data with skewed distributions. An early application in the VAR literature is Panagiotelis and Smith (2008) who proposed the use of a multivariate skew-t distribution. A different approach to modelling skewness is represented by Carriero et al. (2020) who apply a VAR model with conditionally symmetric innovations and skewness induced by mean and variance shifts driven by a financial conditions indicator and find evidence of skewness in the unemployment rate and the financial conditions indicator. Similarly, Carriero et al. (2021b) account for extreme Covid-19 observations using a VAR model with an outlier-augmented stochastic volatility. They show that the model performs on par with a VAR with Student's t distribution. In a univariate context, Liu (2019) estimates different asymmetric and heavy tailed distributions for macroeconomic variables, even though the symmetric Student's t distribution is preferred for monthly data, Delle Monache et al. (2021) model the conditional distribution of GDP using a skew-t distribution with time-varying location, scale and shape parameters and Nakajima and Omori (2012) combine a generalized hyperbolic skew Student's t distribution with stochastic volatility to model stock returns. Montes-Galdón and Ortega (2022) allow the time-varying parameter of the skewness innovations in the structural VAR model to be driven by other economic factors.

There are thus two strands in the literature on skewness, one modelling skewness as an unconditional phenomenon and one modelling (conditional) skewness as time-varying with the skewness driven by economic factors. The second approach is typically motivated by the observation that recessions are often associated with (large) negative shocks. This in itself does not necessarily mean that skewness is time-varying. It could also be the random occurrence of a sequence of negative shocks that drives the economy into a recessionary state. A theoretical argument for this is given by Acemoglu et al. (2017) who show that macroeconomic tail risk can be due to idiosyncratic microeconomic shocks to heterogeneous sectors of the economy. In line with this our contribution is within the first stand of the literature and we model skewness as an essentially unconditional phenomenon.

Table 1 provides some preliminary empirical support for our approach. The first panel of Table 1 reports summary statistics for the variables in our application (see Section 4 for details on the data). We find significant skewness and excess kurtosis for most of the variables. As this does not necessarily imply that the innovations have skewed or heavy tailed distributions the second panel reports on the residuals from an OLS-fit of a homoskedastic VAR(4) model. As can be expected the amount of skewness is reduced, for the residuals only the VIX displays significant skewness. While, on the other hand, the excess kurtosis is now significant for all variables. There are thus signs of (unconditional) skewness as well as fat tails in the data and consequently a need for flexible modelling tools that can capture these features of the data.

Table 1: Summary statistics

	Mean	Standard deviation	Skewness	Kurtosis	Minimum	Maximum	J-B test skewness	J-B test kurtosis
Summary statistics of macroeconomic variables								
Industrial production	0.169	0.729	-1.139	5.410	-4.434	2.377	130.329***	739.383***
Inflation	0.321	0.326	0.184	4.431	-1.786	1.794	3.398*	496.314***
Unemployment	6.202	1.594	0.665	-0.188	3.500	10.800	44.447***	0.795
VIX	2.924	0.329	0.482	0.487	2.081	4.207	23.366***	6.221**
Summary statistics for the residuals from the OLS-fit of a homoskedastic VAR(4)								
Industrial production	0.000	0.628	-0.125	3.627	-3.545	2.728	1.549	330.623***
Inflation	0.000	0.237	0.108	4.645	-1.149	1.501	1.164	541.791***
Unemployment	0.000	0.151	0.041	0.491	-0.568	0.502	0.171	6.275**
VIX	0.000	0.161	1.129	2.814	-0.418	0.738	127.316***	199.363***

Table reports the summary statistics of monthly macroeconomic variables for the period 01/1970 to 12/2019 from the Federal Reserve Bank of St. Louis, see McCracken and Ng (2016). ***,**,* denote significant at 1%, 5%, 10% level of the skewness and kurtosis components of the Jarque-Bera test.

More specifically, we contribute to the literature by extending the VAR model to account for more realistic assumptions on the multivariate distribution of the variables. We propose a general class of skewed distributions with heavy tails and stochastic volatility for the innovations in the VAR. Crucially, we do this by allowing the skewness and heavy tailedness to differ between the variables. In doing so we take the generalized hyperbolic skew Student's t (GHSkew-t) distribution as our starting point and we refer to this as a class of VAR models with the GHSkew-t-SV innovation. The GHSkew-t-SV distribution can be represented as a normal variance-mean mixture and lends itself to straightforward Bayesian inference using a Gibbs sampler with a few Metropolis-Hastings steps. Model comparison and marginal likelihood calculations can be done using the cross-entropy method of Chan and Eisenstat (2018) or the Chib and Jeliazkov (2001) method.

In an application to monthly US macro data we compare the in-sample and out of sample forecast performance of 14 VAR models with different assumptions on the tail distribution and stochastic volatility. We find strong support for VAR models with skewness and heavy tails. Stochastic volatility, heavy tails and skewness all contribute to the in-sample fit. In general, the VAR model with stochastic volatility improves the point and density out-of-sample forecasts. Furthermore, allowing for heavy tailed distributions enhances the out-of-sample forecast which is in agreement with current findings in the literature, see Chiu et al. (2017) and Liu (2019). An interesting finding is that the asymmetric distribution is more important in the VAR models with SV than that in the VAR model without SV. We recommend that skewness as well as heavy tails should be taken into account for better predictions and in-sample fit.

The rest of the paper is organized as follows. Section 2 introduces the GHSkew-t-SV models. Section 3 presents the Bayesian algorithm for inference and discuss how to estimate the marginal likelihood. Section 4 illustrates the usefulness of the proposed models for potentially skew and heavy-tailed data using both in-sample and out-of-sample evidence. Section 5 concludes.

2 VAR Models with skewness and heavy tails

The Student's t distribution is a natural choice for modelling data with heavy tails and has been used quite extensively with VAR models. The Student's t distribution has been extended in several different ways to allow for skewness and asymmetric behaviour. Among these, Ferreira and Steel (2007) propose a multivariate skew-t distribution via an affine linear transformation of independent skew-t variables while Sahu et al. (2003) use a hidden truncation model to construct a multivariate skew-t distribution where the heavy tail behavior is captured by only one parameter.

We will, however, take the GHSkew-t distribution as our starting point. It is commonly used as it is a general class of distribution which nests the Gaussian distribution and the

Student's t distribution as special cases, see McNeil et al. (2015). A particularly appealing feature is that it has a convenient representation as a variance-mean mixture of normal distributions. That is

$$y_t = \gamma\xi_t + \sqrt{\xi_t} z_t$$

with $\xi_t \sim \mathcal{IG}(\frac{\nu}{2}, \frac{\delta}{2})$ independent of $z_t \sim \mathcal{N}(0,1)$ has a GHSkew-t distribution with skewness parameter γ, scale parameter δ and ν degrees of freedom. As we will consider models with stochastic volatility we modify this slightly by fixing the scale at $\delta = \nu$ and letting z_t have a time-varying variance, h_t.[1]

Using the conditional normality of y_t it is straightforward, but tedious, to derive the first few moments of y_t. We have

$$\begin{aligned} E(y_t) &= \frac{\gamma\nu}{\nu-2}, \quad V(y_t) = \frac{h_t\nu}{\nu-2} + \frac{2\gamma^2\nu^2}{(\nu-2)^2(\nu-4)}, \qquad (1)\\ E\left(y_t - E(y_t)\right)^3 &= \frac{6\gamma h_t \nu^2}{(\nu-2)^2(\nu-4)} + \frac{16\gamma^3\nu^3}{(\nu-2)^3(\nu-4)(\nu-6)},\\ E\left(y_t - E(y_t)\right)^4 &= \frac{3h_t^2\nu^2}{(\nu-2)(\nu-4)} + \frac{12\gamma^2 h_t \nu^3(\nu+2)}{(\nu-2)^3(\nu-4)(\nu-6)} + \frac{12\gamma^4\nu^4(\nu+10)}{(\nu-2)^4(\nu-4)(\nu-6)(\nu-8)}, \end{aligned}$$

with the variance, (absolute) third and fourth moments increasing in the (absolute) skewness (γ) and dispersion (h_t) parameters and decreasing in the degrees of freedom (ν). Looking at the standardized measures skewness and kurtosis, the variance, absolute skewness and kurtosis are decreasing in the degrees of freedom and approaches those of a normal distribution as ν increases. The absolute skewness and kurtosis are increasing in the absolute value of γ and decreasing in h_t. It is also worth noting that the existence of the k^{th} moment requires that $\nu > 2k$ when $\gamma \neq 0$.

Another useful property of the GHSkew-t distribution is the difference in tail behavior. Aas and Haff (2006) show that, for $\gamma < 0$, the left tail decays as $|y|^{-\nu/2-1}$ and is heavier

[1] Nakajima and Omori (2012) used a different formulation of the GHSkew-t combined with stochastic volatility in a univariate context.

than the right tail which decays as $|y|^{-\nu/2-1}\exp(-2|\gamma y|)$ and vice versa for a right skewed distribution.

In the following we develop two multivariate extensions of the univariate GHSkew-t distribution which are suitable for use with VAR models. In doing this we take a VAR with Gaussian volatility as the starting point as we are concerned with both the skewness and heavy tails of the innovations in the VAR. While we are focusing on the distribution of the innovations conditional on the volatility, the unconditional distribution is also of interest and will be heavy tailed even with Gaussian innovations when the variance is time-varying and stochastic as with stochastic volatility or GARCH-type conditional variances (Carriero et al., 2020).

2.1 A VAR Model with Gaussian-SV innovations

Following Primiceri (2005) we write the VAR model with Gaussian stochastic volatility (Gaussian-SV) as

$$\begin{aligned} \mathbf{y}_t &= \mathbf{c} + \mathbf{B}_1\mathbf{y}_{t-1} + \ldots + \mathbf{B}_p\mathbf{y}_{t-p} + \mathbf{u}_t, \quad t = 1, \ldots, T, \\ &= \mathbf{B}\mathbf{x}_t + \mathbf{u}_t, \end{aligned} \tag{2}$$

where $\mathbf{y}_t$ is a k-dimensional vector of endogenous variables; $\mathbf{c}$ is a k-dimensional vector of constants; $\mathbf{B}_j$ is a $k \times k$ variate matrix of regression coefficients with $j = 1, \ldots, p$; $\mathbf{u}_t$ is a k-dimensional vector of reduced-form heteroskedastic innovations associated with the VAR equations. To simplify the notation, let $\mathbf{B} = (\mathbf{c}, \mathbf{B}_1, \ldots, \mathbf{B}_p)$ be a $k \times (1 + kp)$ matrix, and $\mathbf{x}_t = (1, \mathbf{y}'_{t-1}, \ldots, \mathbf{y}'_{t-p})'$ be $(1+kp)$-dimensional vector. In the VAR model with Gaussian-SV innovations $\mathbf{u}_t$ is given by

$$\mathbf{u}_t = \mathbf{A}^{-1}\mathbf{H}_t^{1/2}\boldsymbol{\epsilon}_t, \tag{3}$$

where $\mathbf{A}$ is a $k \times k$ lower triangular matrix with ones on the diagonal that describes the contemporaneous interaction of the endogenous variables;[2] $\mathbf{H}_t$ is a $k \times k$ diagonal matrix that captures the heteroskedastic volatility; $\boldsymbol{\epsilon}_t$ is a k-dimensional vector of innovations that follows a multivariate Gaussian distribution with zero mean vector and identity covariance matrix, i.e. $\boldsymbol{\epsilon}_t \sim \mathcal{N}_k(\mathbf{0}, \mathbf{I})$. We assume that the log volatilities follow a random walk for $\mathbf{H}_t = \text{diag}(h_{1t}, \ldots, h_{kt})$ with

$$\log h_{it} = \log h_{it-1} + \sigma_i \eta_{it}, \quad i = 1, \ldots, k, \tag{4}$$

where $\eta_{it} \sim \mathcal{N}(0, 1)$. The VAR model without stochastic volatility can be obtained by fixing zero values of $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_k^2)'$ and assuming that $\log h_{it} = \log h_{i0}$ for $i = 1, \ldots, k$ and $t = 1, \ldots, T$.

2.2 A VAR Model with Orthogonal Skew-t-SV innovations

Next, we allow for skewness and heavy tails by defining the reduced form innovations $\mathbf{u}_t$ as a rotation of GHSkew-t innovations. We refer to this as the VAR with an orthogonal skew-t innovation (OST). Given the recursive structure in $\mathbf{A}$, we let the "structural" innovations, $\mathbf{e}_t$, be a vector of zero mean independent generalized hyperbolic skew t random variables,

$$\mathbf{u}_t = \mathbf{A}^{-1}\mathbf{e}_t = \mathbf{A}^{-1}\left(\left(\mathbf{W}_t - \overline{\mathbf{W}}\right)\boldsymbol{\gamma} + \mathbf{W}_t^{1/2}\mathbf{H}_t^{1/2}\boldsymbol{\epsilon}_t\right), \tag{5}$$

where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_k)'$ is a k-dimensional vector of skewness parameters, the mixing matrix $\mathbf{W}_t = \text{diag}(\boldsymbol{\xi}_t) = \text{diag}(\xi_{1t}, \ldots, \xi_{kt})$ is a $k \times k$ diagonal matrix, ξ_{it} follows inverse Gamma

[2] The triangular form of $\mathbf{A}$ is here mainly a device for partitioning the likelihood and is not necessary for identifying the shocks in models with non-Gaussian innovations and/or stochastic volatility, see e.g. Lanne et al. (2017), Carriero et al. (2021a) and Lewis (2021) on identification. While convenient, the triangular form of $\mathbf{A}$ comes with the drawback that it introduces dependence on the order of the variables in the reduced form innovations, $\mathbf{u}_t$, with stochastic volatility and/or the multivariate distributions we develop below. If order dependence is an issue, $\mathbf{A}$ can be modelled as an unrestricted matrix using the approach of Chan et al. (2021). We abstain from this as we want to include the Gaussian non-SV VAR, where an unrestricted $\mathbf{A}$ would be unidentified, in our model comparison.

distributions with shape parameter $\nu_i/2$ and scale parameter $\nu_i/2$ independent of $\boldsymbol{\epsilon}_t$, i.e. $\xi_{it} \sim \mathcal{IG}(\frac{\nu_i}{2}, \frac{\nu_i}{2})$. For later reference we collect the degree of freedom parameters in the vector $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_k)'$. Finally $\overline{\mathbf{W}} = E(\mathbf{W}_t) = \text{diag}(\mu_{\xi,1}, \ldots, \mu_{\xi,k}) = \text{diag}(\frac{\nu_1}{\nu_1 - 2}, \ldots, \frac{\nu_k}{\nu_k - 2})$ centers the innovations to have mean zero. Note that the mixing variables ξ_{it} are specific to the "structural" innovations e_{it} and that the reduced form innovations, u_{it}, (except for u_{1t}) do not have a generalized hyperbolic skew t distribution even if the degrees of freedom are the same across equations. By setting $\boldsymbol{\gamma}$ to zero we obtain the symmetric and orthogonal t distribution (OT) used by Cúrdia et al. (2014), Clark and Ravazzolo (2015) and Chiu et al. (2017). As usual, by letting the degree of freedom $\nu_i \to \infty$ for $i = 1, \ldots, p$, the VAR with an OT innovation becomes a VAR with Gaussian innovations.

Chiu et al. (2017) interprets the mixing matrix $\mathbf{W}_t$ as capturing the high-frequency shocks in mean and volatility while the stochastic volatility accounts for the low-frequency shocks. The data will determine whether the extreme time variation comes from the volatility shift or from the idiosyncratic heavy tail shocks.

2.3 A VAR Model with Multi-Skew-t-SV innovations

The VAR with an OST distribution builds the distribution of the innovation terms from the ground up in terms of the structural form innovations. This makes for a straightforward structural interpretation but also means that the model is sensitive to the (over) identifying assumptions, in this case the triangular structure of $\mathbf{A}$ and the ordering of the variables. To partially overcome this and link the skewness and heavy tailed properties to the reduced form innovations rather than the structural shocks we can model the reduced form innovations directly as a correlated vector of univariate skew-t distributions. We propose a class of VAR models with multi skew-t innovations (MST) by defining the reduced form innovations $\mathbf{u}_t$ as

$$\mathbf{u}_t = (\mathbf{W}_t - \overline{\mathbf{W}})\boldsymbol{\gamma} + \mathbf{W}_t^{1/2}\mathbf{A}^{-1}\mathbf{H}_t^{1/2}\boldsymbol{\epsilon}_t. \tag{6}$$

To aid in interpretation, note that $\widetilde{\boldsymbol{\epsilon}}_t = \mathbf{A}^{-1}\mathbf{H}_t^{1/2}\boldsymbol{\epsilon}_t \sim \mathcal{N}_k(\mathbf{0}, \mathbf{A}^{-1}\mathbf{H}_t\mathbf{A}^{-1\prime})$.[3] We then apply individual variance-mean mixtures to each element of the vector $\widetilde{\boldsymbol{\epsilon}}_t$ to allow for the different tail behaviour of the reduced form innovations, $\mathbf{u}_t$, which sets this apart from the usual (skew) multivariate t distributions. The marginal distribution of u_{it} is thus a GHSkew-t distribution for $i = 1, \ldots, k$ and $t = 1, \ldots, T$. Restricting the mixing variables to be equal for the different equations, $\xi_{1t} = \ldots = \xi_{kt}$, induces a common tail behaviour and the conditional distribution of $\mathbf{u}_t$ is a multivariate generalized hyperbolic skew Student t (Skew-t) distribution (McNeil et al., 2015). If we in addition set $\gamma_1 = \ldots = \gamma_k = 0$, a multivariate Student t (Student-t) distribution is obtained. The last special case we consider sets $\gamma_i = 0$ for symmetry but retains the equation specific variance mixtures for a multi Student's t (MT) distribution. As usual, if $\nu_i \to \infty$, the MT model becomes a VAR with Gaussian stochastic volatility in spirit of Cogley and Sargent (2005) and Primiceri (2005).

2.4 Comparison of the model implied distributions

To illustrate the properties of the MST and OST distributions we focus on the bivariate vector of innovations $\mathbf{u}_t = (u_{1t}, u_{2t})'$ given as follows

$$\mathbf{MST} : \mathbf{u}_t = (\mathbf{W}_t - \overline{\mathbf{W}})\boldsymbol{\gamma} + \mathbf{W}_t^{1/2}\mathbf{A}^{-1}\mathbf{H}_t^{1/2}\boldsymbol{\epsilon}_t,$$

$$u_{1t} = (\xi_{1t} - \mu_{\xi,1})\gamma_1 + \sqrt{\xi_{1t}h_{1t}}\epsilon_{1t}, \tag{7}$$

$$u_{2t} = (\xi_{2t} - \mu_{\xi,2})\gamma_2 + \sqrt{\xi_{2t}}(\rho\sqrt{h_{1t}}\epsilon_{1t} + \sqrt{h_{2t}}\epsilon_{2t}); \tag{8}$$

$$\mathbf{OST} : \mathbf{u}_t = \mathbf{A}^{-1}(\mathbf{W}_t - \overline{\mathbf{W}})\gamma + \mathbf{A}^{-1}\mathbf{W}_t^{1/2}\mathbf{H}_t^{1/2}\boldsymbol{\epsilon}_t,$$

$$u_{1t} = (\xi_{1t} - \mu_{\xi,1})\gamma_1 + \sqrt{\xi_{1t}h_{1t}}\epsilon_{1t}, \tag{9}$$

$$\begin{aligned} u_{2t} &= \rho u_{1t} + (\xi_{2t} - \mu_{\xi,2})\gamma_2 + \sqrt{\xi_{2t}h_{2t}}\epsilon_{2t} \\ &= (\xi_{1t} - \mu_{\xi,1})\rho\gamma_1 + (\xi_{2t} - \mu_{\xi,2})\gamma_2 + \rho\sqrt{\xi_{1t}h_{1t}}\epsilon_{1t} + \sqrt{\xi_{2t}h_{2t}}\epsilon_{2t}, \end{aligned} \tag{10}$$

[3] There is order dependence in the term $\mathbf{A}^{-1}\mathbf{H}_t^{1/2}\boldsymbol{\epsilon}_t$ as the stochastic volatility is affected by the order of the variables. While this might be seen as a problem we note that the specification is standard practice in VAR models with stochastic volatility.

where ρ is the free parameter in the $\mathbf{A}^{-1}$ matrix. For both the MST and OST distributions, u_{1t} follows a GHSkew-t distribution and the third moment of u_1 follows from (1),

$$E\left(u_{1t}^3\right) = \frac{2\gamma_1\nu_1^2}{(\nu_1-2)^2(\nu_1-4)}\left(\frac{8\gamma_1^2\nu_1}{(\nu_1-2)(\nu_1-6)} + 3h_{1t}\right).$$

The distinguishing features of the MST and OST distributions can be seen by comparing (8) and (10). In the MST distribution, u_{2t} also follows a GHSkew-t distribution and is correlated with u_{1t}, however, they do not share the same mixing variable. The third moment of u_{2t} and cross-moments of u_{1t} and u_{2t} are given by,

$$\begin{aligned}
E\left(u_{1t}u_{2t}^2\right) &= \frac{1}{2}\rho\gamma_2 h_{1t}\frac{\nu_1^{1/2}\nu_2^{3/2}}{\nu_2-2}\frac{\Gamma(\frac{\nu_1-1}{2})\Gamma(\frac{\nu_2-3}{2})}{\Gamma(\frac{\nu_1}{2})\Gamma(\frac{\nu_2}{2})}, \\
E\left(u_{1t}^2u_{2t}\right) &= \frac{1}{2}\rho\gamma_1 h_{1t}\frac{\nu_2^{1/2}\nu_1^{3/2}}{\nu_1-2}\frac{\Gamma(\frac{\nu_2-1}{2})\Gamma(\frac{\nu_1-3}{2})}{\Gamma(\frac{\nu_1}{2})\Gamma(\frac{\nu_2}{2})}, \\
E\left(u_{2t}^3\right) &= \frac{2\gamma_2\nu_2^2}{(\nu_2-2)^2(\nu_2-4)}\left(\frac{8\gamma_2^2\nu_2}{(\nu_2-2)(\nu_2-6)} + 3(\rho^2h_{1t}+h_{2t})\right).
\end{aligned} \tag{11}$$

In the OST distribution on the other hand, u_2 is a linear combination of two GHSkew-t distributions. The cross-moments of u_{1t} and u_{2t} and the third moment of u_{2t} can be derived as,

$$\begin{aligned}
E\left(u_{1t}u_{2t}^2\right) &= \frac{2\gamma_1\rho^2\nu_1^2}{(\nu_1-2)^2(\nu_1-4)}\left(\frac{8\gamma_1^2\nu_1}{(\nu_1-2)(\nu_1-6)} + 3h_{1t}\right), \\
E\left(u_{1t}^2u_{2t}\right) &= \frac{2\gamma_1\rho\nu_1^2}{(\nu_1-2)^2(\nu_1-4)}\left(\frac{8\gamma_1^2\nu_1}{(\nu_1-2)(\nu_1-6)} + 3h_{1t}\right), \\
E\left(u_{2t}^3\right) &= \rho^3E\left(u_{1t}^3\right) + \frac{2\gamma_2\nu_2^2}{(\nu_2-2)^2(\nu_2-4)}\left(\frac{8\gamma_2^2\nu_2}{(\nu_2-2)(\nu_2-6)} + 3h_{2t}\right).
\end{aligned} \tag{12}$$

The tail dependence between u_{1t} and u_{2t} can be measured by the coskewness, the normalized cross-moment. Intuitively, the coskewness measures how the variable is related to the magnitude of the other variable. For u_{1t} and u_{2t} we have

$$coskew_1 = \frac{E\left(u_{1t}u_{2t}^2\right)}{\sqrt{V\left(u_{1t}\right)}V\left(u_{2t}\right)}, coskew_2 = \frac{E\left(u_{1t}^2u_{2t}\right)}{V\left(u_{1t}\right)\sqrt{V\left(u_{2t}\right)}}.$$

From (11) and (12) we see that the sign of the coskewness measures depends on the γ parameter for the squared variable while the sign of both measures only depends on γ_1 for the OST. In the MST distribution, $coskew_1$ (or $coskew_2$) increases in the product of ρ and γ_2 (or γ_1) and the sign depends on ρ but it is quite insensitive to the changes in the volatilities h_{1t} and h_{2t}. In the OST distribution, $coskew_1$ and $coskew_2$ increases in the skewness γ_1, the volatilities h_{1t} and h_{2t} and the absolute value of ρ with the sign of $coskew_2$ depending on ρ. The effect of the different parameters on the skewness and coskewness is illustrated in Figure 6 in Appendix A.

Figure 1 shows the shape of the MST and OST distributions and illustrates the effect of the volatilites on the joint distributions of u_{1t} and u_{2t}. For $h_{1t} = 1$ and $h_{2t} = 1$, we observe that the MST only induces skewness and heavy tails in each marginal distribution and the joint distribution reveals no tail dependence, while the OST shows stronger tail dependence. A decrease in h_{1t} increases the skewness of u_1 in both the MST and OST distributions. In addition, a decrease in h_{1t} will also increase the skewness of u_2, although to a smaller degree. Changes to h_{2t} only affects u_2 in both the MST and OST distributions, see Figure 6 in Appendix A. The VAR models with MST-SV or OST-SV innovations can thus have time-varying skewness.

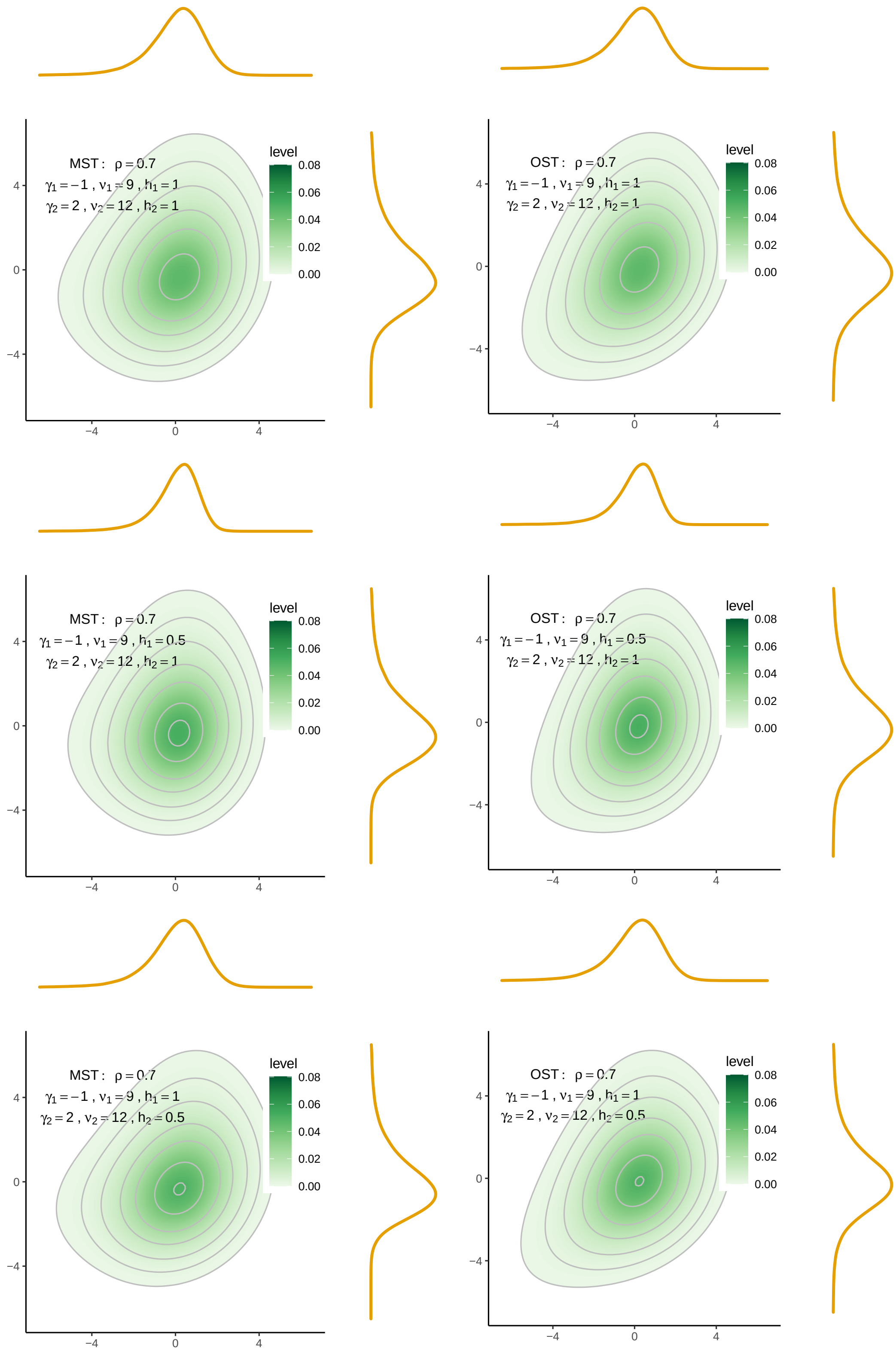

Figure 1: Emprirical contour and density plots based on 100,000 draws from the MST and OST distributions with different dispersion parameters h_{1t} and h_{2t} and $\rho = 0.7$, $\boldsymbol{\gamma} = (-1, 2)$, $\boldsymbol{\nu} = (9, 12)$.

3 Bayesian Inference

To conserve space, prior distributions, procedures for posterior inference using a Gibbs sampler and model selection based on marginal likelihoods are only given for the MST-SV specification of the VAR. In most cases the modifications (simplifications) needed for the other specifications (Gaussian, Student-t, Skew-t, orthogonal Student's t (OT), multi Student's t (MT) and orthogonal skew Student's t (OST)) are straightforward with details given in the Online Appendix A.

3.1 Prior Distribution

Denote the set of the VAR-MST-SV model parameters and latent variables by $\boldsymbol{\theta} = \{\mathbf{B}, \mathbf{a}, \boldsymbol{\gamma}, \boldsymbol{\nu}, \boldsymbol{\sigma}^2, \boldsymbol{\xi}_{1:T}, \mathbf{h}_{0:T}\}$, where $\mathbf{a} = (a_{2,1}, a_{3,1}, a_{3,2}, \ldots, a_{k,k-1})'$ is the set of elements of the lower triangular matrix $\mathbf{A}$; the latent variables are $\boldsymbol{\xi}_{1:T} = (\xi_{11}, \ldots, \xi_{kT})$ and $\mathbf{h}_{0:T} = (h_{10}, \ldots, h_{kT})$ with $\mathbf{h}_0 = (h_{10}, \ldots, h_{k0})$ as the vector of initial values for the stochastic volatilities. We employ the Minnesota prior for the prior distributions of $\mathbf{B}$ with overall shrinkage $l_1 = 0.2$ and cross-variable shrinkage $l_2 = 0.5$, see Koop and Korobilis (2010), and vague prior distributions for other parameters. In details, the Minnesota-type prior assume a Gaussian prior for vec($\mathbf{B}$), i.e. $\text{vec}(\mathbf{B}) \sim \mathcal{N}(\mathbf{b}_0, \mathbf{V}_{\mathbf{b}_0})$, that shrinks the regression coefficients towards univariate random walks with a tighter prior around zero for longer lags. The prior for $\mathbf{a}$ is also Gaussian, $\mathbf{a} \sim \mathcal{N}_{0.5k(k-1)}(0, 10\mathbf{I})$, which implies a weak assumption of no interaction among endogenous variables. The parameters which account for the heavy tails are endowed with Gamma priors, $\nu_i \sim \mathcal{G}(2, 0.1)$ truncated to the range (4,100) to ensure finite second moments for $i = 1, \ldots, k$ and the skewness parameters are given a normal prior, $\boldsymbol{\gamma} \sim \mathcal{N}_k(\mathbf{0}, \mathbf{I})$. That is the prior mean of the degrees of freedom of the t-distribution is 20 and the skewness has zero prior mean. These priors are uninformative and allow the data to choose between the Gaussian distribution and the skew and/or heavy-tailed GHSkew-t distribution. Finally, the prior for the variance of the shock to the volatility is $\sigma_i^2 \sim \mathcal{G}(\frac{1}{2}, \frac{1}{2V_\sigma})$ where $V_\sigma = 1$

which is equivalent to $\pm\sqrt{\sigma_i^2} \sim \mathcal{N}(0, V_\sigma)$, see Kastner and Frühwirth-Schnatter (2014), this prior is less influential in comparison to the conjugate inverse Gamma prior especially when the true value is small. In all cases of VAR models with and without stochastic volatility $\log h_{i0} \sim \mathcal{N}\left(\log \widehat{\Sigma}_{i,OLS}, 4\right)$ where $\widehat{\Sigma}_{i,OLS}$ is the estimated residual variance of a univariate AR(p) model using the ordinary least square method, see Clark and Ravazzolo (2015).

3.2 Posterior Inference

Let $\mathbf{\Psi}$ be all the parameters and latent variables in $\boldsymbol{\theta}$ except the ones we sample from in a given step of the MCMC procedure. The MCMC algorithm for the VAR with MST-SV can then be outlined as follows (see Appendix B for details).

1. Sample $(\boldsymbol{b}|\mathbf{\Psi})$ where $\boldsymbol{b} = (\text{vec}(\mathbf{B})', \boldsymbol{\gamma}')'$ from the full conditional normal posterior.

2. Sample $(\mathbf{a}|\mathbf{\Psi})$ from the full conditional normal posterior as in Cogley and Sargent (2005).

3. Sample $(\mathbf{h}_{0:T}|\mathbf{\Psi})$ using the forward filter backward smoothing algorithm in Carter and Kohn (1994).

4. Sample $(\boldsymbol{\sigma}^2|\mathbf{\Psi})$ from the full conditional generalized inverse Gamma (GIG) posterior.

5. Sample $(\nu_i|\mathbf{\Psi}) \propto \mathcal{G}(\nu_i; 2, 0.1) \prod_{t=1}^{T} \mathcal{IG}\left(\xi_{it}; \frac{\nu_i}{2}, \frac{\nu_i}{2}\right)$ for $i = 1, \ldots, k$ using a Metropolis-Hastings step with a random walk proposal.

6. Sample $(\boldsymbol{\xi}_t|\mathbf{\Psi})$ for $t = 1, \ldots, T$, using a Metropolis-Hasting step with an independence proposal.

3.3 Model Selection

The marginal likelihoods of the VAR models with GHSkew-t-SV innovation require the high-dimensional integration

$$p(\mathbf{y}_{1:T}) = \int p(\mathbf{y}_{1:T}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}, \tag{13}$$

where the main issue is the need to integrate out the latent variables $\boldsymbol{\xi}_{1:T}$ and $\mathbf{h}_{0:T}$. $\boldsymbol{\xi}_{1:T}$ can, with exception for the MST and MST-SV models, be integrated out analytically. For $\mathbf{h}_{0:T}$ (and when needed $\boldsymbol{\xi}_{1:T}$) we use importance sampling as in Chan and Eisenstat (2018). For the integral over the static parameters $\boldsymbol{\theta}_1 = \{\mathbf{B}, \mathbf{a}, \boldsymbol{\gamma}, \boldsymbol{\nu}, \boldsymbol{\sigma}^2\}$ we use the cross-entropy method following Chan and Eisenstat (2018) as well as the Chib and Jeliazkov (2001) method and find that both give reliable estimates of the marginal likelihood.

Integrating over the latent states $\boldsymbol{\theta}_2 = \{\boldsymbol{\xi}_{1:T}, \mathbf{h}_{0:T}\}$ yields an estimate of the integrated likelihood, $\widehat{p}(\mathbf{y}_{1:T}|\boldsymbol{\theta}_1)$. See Appendix C.1 for details on the integration. The cross-entropy method is also based on importance sampling for $\boldsymbol{\theta}_1$ and we use an approximation of the ideal importance function $\pi(\boldsymbol{\theta}_1|\mathbf{y}_{1:T})$ as follows:

1. Obtain the posterior samples $\boldsymbol{\theta}_1^{(1)}, \ldots, \boldsymbol{\theta}_1^{(R)}$ from the posterior density $\pi(\boldsymbol{\theta}|\mathbf{y}_{1:T})$.

2. Given the parametric family $f(\boldsymbol{\theta}_1; \lambda)$ parameterized by parameter λ find the maximum likelihood estimate
$$\lambda^* = \arg\max_{\lambda} \frac{1}{R}\sum_{r=1}^{R} \log f(\boldsymbol{\theta}_1^{(r)}; \lambda).$$

3. Obtain new samples $\boldsymbol{\theta}_1^{(1)}, \ldots, \boldsymbol{\theta}_1^{(N)}$ from $f(\boldsymbol{\theta}_1; \lambda^*)$. For each new value $\boldsymbol{\theta}_1^{(n)}$, the integrated likelihood $\widehat{p}(\mathbf{y}_{1:T}|\boldsymbol{\theta}_1^{(n)})$ is estimated using an inner importance sampling loop. Then the marginal likelihood is calculated via importance sampling
$$\widehat{p}_{IS}(\mathbf{y}_{1:T}) = \frac{1}{N}\sum_{n=1}^{N} \frac{\widehat{p}(\mathbf{y}_{1:T}|\boldsymbol{\theta}_1^{(n)})p(\boldsymbol{\theta}_1^{(n)})}{f(\boldsymbol{\theta}_1^{(n)}|\lambda^*)}.$$

The number of samples N is chosen such that the variance of the estimated quantity using important sampling is less than one. The parametric families of $f(\boldsymbol{\theta}_1; \lambda)$ are the multivariate Gaussian distribution for $(\mathbf{B}, \mathbf{a}, \boldsymbol{\gamma})$, independent Gamma distributions for ν_i and independent Gamma distribution for σ_i^2.

In the final importance sampling step we observe a large sample variance of the ratios $\widehat{p}(\mathbf{y}_{1:T}|\boldsymbol{\theta}_1^{(n)})p(\boldsymbol{\theta}_1^{(n)})/f(\boldsymbol{\theta}_1^{(n)}|\lambda^*)$ for the more complicated models and we cannot rule out that the true variance is infinite. As an alternative and a check on the reliability we also use the Chib and Jeliazkov (2001) algorithm for the integration over $\boldsymbol{\theta}_1$. The method is based on the marginal likelihood identity

$$p(\mathbf{y}_{1:T}) = \frac{p(\mathbf{y}_{1:T}|\mathbf{B}^*, \boldsymbol{\gamma}^*, \mathbf{A}^*, \boldsymbol{\sigma}^{2*}, \boldsymbol{\nu}^*)p(\mathbf{B}^*, \boldsymbol{\gamma}^*, \mathbf{A}^*, \boldsymbol{\sigma}^{2*}, \boldsymbol{\nu}^*)}{p(\mathbf{B}^*, \boldsymbol{\gamma}^*, \mathbf{A}^*, \boldsymbol{\sigma}^{2*}, \boldsymbol{\nu}^*|\mathbf{y}_{1:T})},$$

where $\mathbf{B}^*, \boldsymbol{\gamma}^*, \mathbf{A}^*, \boldsymbol{\sigma}^{2*}, \boldsymbol{\nu}^*$ are the posterior means of the parameters. The prior is available in closed form and the posterior density ordinate in the denominator is estimated using a sequence of reduced MCMC samplers. See Appendix C.2 for details.

4 What can we learn about skewness and heavy tails in the data?

To investigate the extent of skewness in macroeconomic data and the ability of our models to capture this we estimate a four-variable VAR with the growth rate of industrial production, inflation rate, unemployment rate, Chicago board options exchange's volatility index (VIX). We use monthly data for the period 01/1970 to 12/2019 from the Federal Reserve Bank of St. Louis, see McCracken and Ng (2016). Industrial production is included as a growth rate (first difference of the logarithm of the index), the inflation rate is calculated as the first difference of the log of the CPI and the logarithm of the VIX is used. The variables enter with $p = 4$ lags.

To assess the importance of different features of the distribution of the innovations We compare 14 different specifications: Gaussian, Student-t, Skew-t, orthogonal Student's t (OT), multi Student's t (MT), orthogonal skew Student's t (OST), multi skew Student's t (MST). All with and without stochastic volatility.

We first estimate the 14 VAR models with and without stochastic volatility using the full data set in order to provide in-sample evidence. Then we perform an out-of-sample forecasting exercise to measure the forecast accuracy of each VAR model. Details on the numerical performance and convergence diagnostics of the MCMC algorithms are provided in the Online Appendix B.

4.1 In-sample Analysis

The left-hand side of Figure 2 shows the growth rate of industrial production, inflation rate, unemployment rate and the VIX. Extreme values of the variables are often observed during recession periods based on the NBER indicators. Industrial production growth decreased by more than 4% during the financial crisis in 2008, while the unemployment rate peaked at 10% and the VIX reached as high as 4.2. As indicated by the statistics reported in Table 1 this skewed behaviour does not necessarily imply that the distribution of the innovations is skewed. Heavy tail, on the other hand, seems to be a robust feature of the innovations. The observed heavy tails can be due both to heteroskedasticity and an unconditional heavy tailed distribution. The extent to which these are present in the data is illustrated in the right hand side of Figure 2 where we plot the posterior means of the volatilities for the Gaussian-SV and OST-SV models. The estimated volatility tends to be larger for the Gaussian-SV indicating that the volatility might be overestimated when not allowing for a heavy-tailed distribution, an observation in line with the findings of, among others, Cúrdia et al. (2014) and Chiu et al. (2017). The Gaussian-SV also gives much more erratic estimates of the volatilities for the VIX while the estimates from the OST-SV are quite smooth. Here we see the effect of

Table 2: Log marginal likelihood for VAR models with and without stochastic volatility

		Gaussian	Student-t	Skew-t	OT	MT	OST	MST
Non SV	LML	-217.627	-128.804	-137.211	-128.478	-125.879	-128.137	-125.268
	sd	(0.003)	(0.005)	(0.006)	(0.063)	(0.009)	(0.021)	(0.048)
SV	LML	-46.919	-28.884	-27.223	-30.293	-27.403	-20.583	-18.427
	sd	(0.031)	(0.023)	(0.245)	(0.042)	(0.044)	(0.129)	(0.150)

We compare the LMLs of 14 VAR models with/without SV. We use the cross entropy methods by Chan and Eisenstat (2018) to calculate the LMLs. We first sample 100,000 draws from the conditional posterior distributions with 10,000 draws as burn-in. Then, all LMLs estimated using 100,000 draws from the proposal distributions, see details in Section 3.3. The standard errors of the estimation using the batch means method (10 batches) are reported in the brackets. Estimates of the log marginal likelihoods using the Chib and Jeliazkov (2001) method are reported in Table 5 in Appendix C.2.

not allowing for a skewed distribution where the Gaussian-SV interprets the occasional large positive outlier as an increase in the volatility.

Table 2 provides formal evidence on how well the different specifications captures the features in the data in the form of log marginal likelihoods. In the class of VAR models without SV, allowing for heavy tails leads to a substantial improvement in the marginal likelihood while the addition of skewness is less useful. Allowing for stochastic volatility leads to a dramatic improvement in the marginal likelihood for all seven specifications. Allowing for heavy tails improves on the Gaussian-SV and allowing for skewness now makes a difference with the more flexible OST and MST specifications of skewness performing best with a log Bayes factor of 9.0 (MST) and 6.8 (OST) against the best Student-t specification. It is interesting that skewness plays a more important role in the VAR model with SV than in the VAR model without SV. The flexibility of the tail behaviour in the OST and MST is important as evidenced by the relatively poor performance of the skew-t VAR models where only one mixing variable is used to model the heavy tails.

Next, we take a closer look at the posterior distributions of the skewness and heavy tail parameters in the best fitting models with and without stochastic volatility. Figure 3 shows the posterior distribution of the skewness parameters and the degree of freedom parameters in the VAR models with MST and MST-SV. Consistent with stochastic volatility inducing heavier tails in the marginal distribution of the innovations the left column shows higher

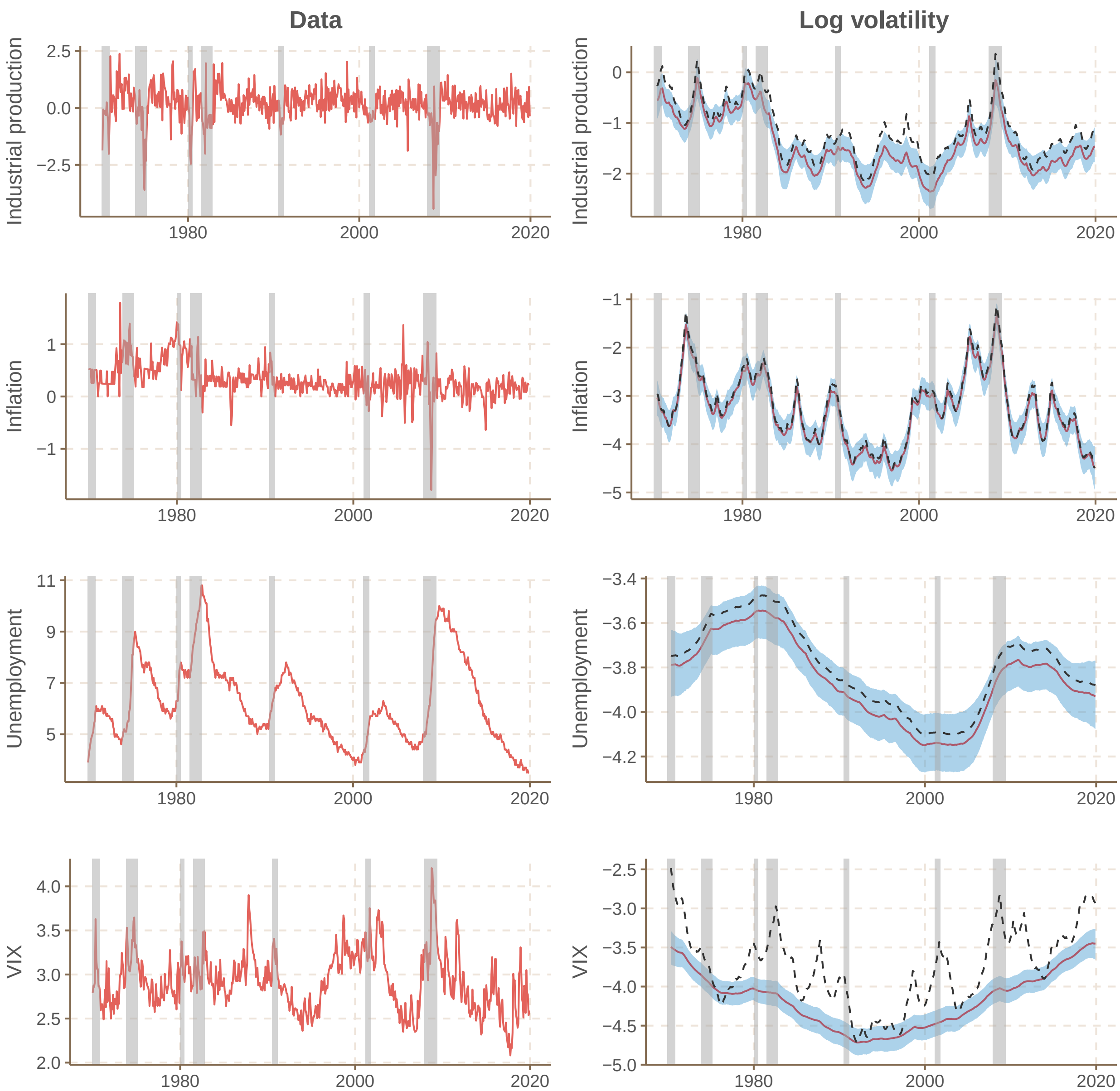

Figure 2: Data and estimated volatilities.

The figures on the left-hand side show the variables while the figures on the right-hand side draw the estimated mean log volatility of the OST-SV model using a solid line (red) with their 50% credible interval. The dashed line shows the estimated mean log volatility from the Gaussian-SV model. The shaded areas highlight the recession periods based on the NBER indicators.

degrees of freedom for industrial production and inflation with the SV specification. The posterior distribution of ν_i for unemployment and the VIX is, on the other hand, barely affected by the addition of stochastic volatility. For industrial production and the VIX there is clear evidence of heavy tails in the distribution and less so for inflation and unemployment. Turning to the skewness parameters, γ_i, in the right column we observe a relatively large shift to the right in the posterior distribution for industrial production when we allow for stochastic volatility (the posterior probability of a positive γ_1 increases from 0.71 to 0.85) and a small shift for the VIX. It thus seems that stochastic volatility helps in unmasking some of the underlying skewness in the data. For inflation and unemployment there is little evidence of skewness. The difference in the posterior distributions between the variables highlights the importance of a flexible specification of the distribution of the innovations. Focusing on the model with stochastic volatility we have a heavy tailed distribution with some skewness for the growth rate in industrial production while inflation and industrial production appear to have normal and symmetric distributions. The VIX, finally, is heavy tailed with a clearly positive skewness. The results are similar for the OST specifications with and without stochastic volatility, see Figures 1 and 2 in Online Appendix C.

As the OST and MST specifications can be sensitive to the ordering of the variables we also analyse the data with an alternative ordering for these specifications. The alternative ordering we consider is Inflation, the VIX, Industrial production and Unemployment. It turns out that the result is largely unaffected by the change in ordering, partly because the $\mathbf{A}$ matrix is close to diagonal in our application. This is illustrated in Figures 3 and 4 in Online Appendix D which compares the posterior distribution of the skewness and heavy tail parameters γ_i and ν_i for the two orderings of the variables.

As a complement to the evidence on skewness in the data provided by the marginal likelihoods and the posterior distribution of the skewness parameters Figure 4 shows the time-varying skewness of the innovations for the MST-SV specification. Consistent with the results in Figure 3 there is little evidence of skewness in inflation and unemployment while

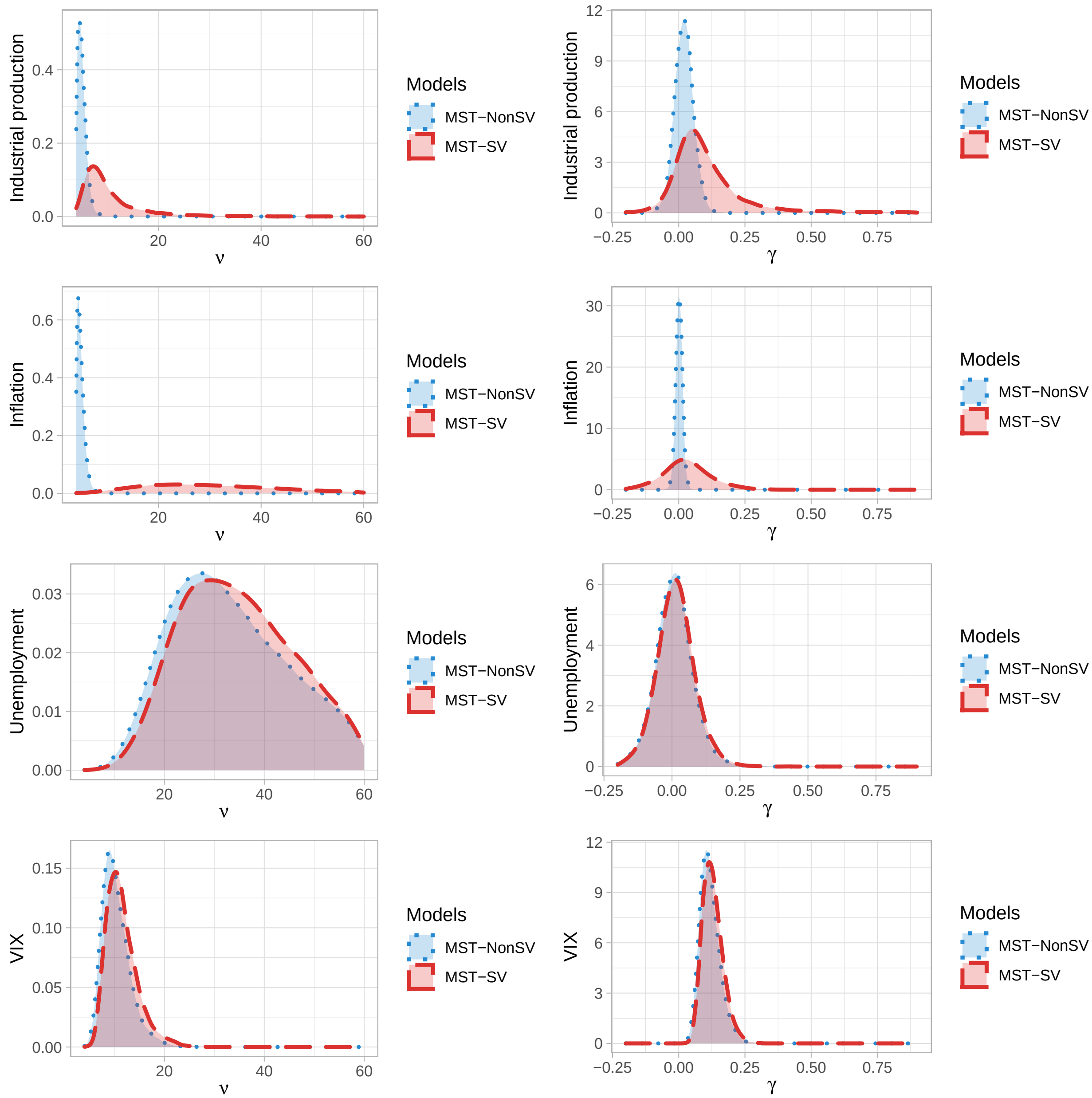

Figure 3: Posterior distribution of the heavy tail (left column) and skewness (right column) parameters of the VAR models with MST and MST-SV.

the posterior distribution of the skewness is separated from zero for industrial production and the VIX. For the VIX we also observe substantial time variation in the skewness and it is clear how the skewness is affected by the time varying volatility. Even if industrial production and the VIX show significant degrees of skewness, the coskewness is very small due to the small correlations between them. Hence, a large magnitude shock to industrial production (or VIX) does not necessary induce a extreme shock to the other variable.

4.2 Out-of-sample forecasts

To assess the out-of-sample predictive accuracy of the different specifications, we conduct a recursive forecast exercise using the 01/2000 to 12/2019 period as our evaluation sample. We calculate the mean square forecast error (MSFE) to evaluate the point forecasts, and the log predicitive density (LP) and continuous rank probability score (CRPS) to evaluate the density forecasts. Details on the forecast metrics are given in Online Appendix E. As the VAR models can be nested based on the distributional assumptions, they are divided into two model groups without and with stochastic volatility for ease of comparison. Using the Gaussian VAR as a benchmark in each group, we test for equal forecast accuracy using the two-sided Diebold and Mariano (1995) test where the standard errors of the test statistics are computed with the Newey–West estimator. We also compare the models with skew distributions to the alternative Student-t distributions and highlight the effect of allowing for skewness in the VAR models.

Table 3 reports on the performance of the point forecasts from the different specifications and show the improvements in MSFE over the Gaussian VAR models. Each panel reports the ratio of the MSFE for each variable to the MSFE of the Gaussian VAR model with (and without) stochastic volatility. Entries less than 1 indicate that the given model is better than the corresponding Gaussian model. In the non-stochastic volatility VAR group, skewness and heavy tail models improve the point forecast of the growth rate of Industrial

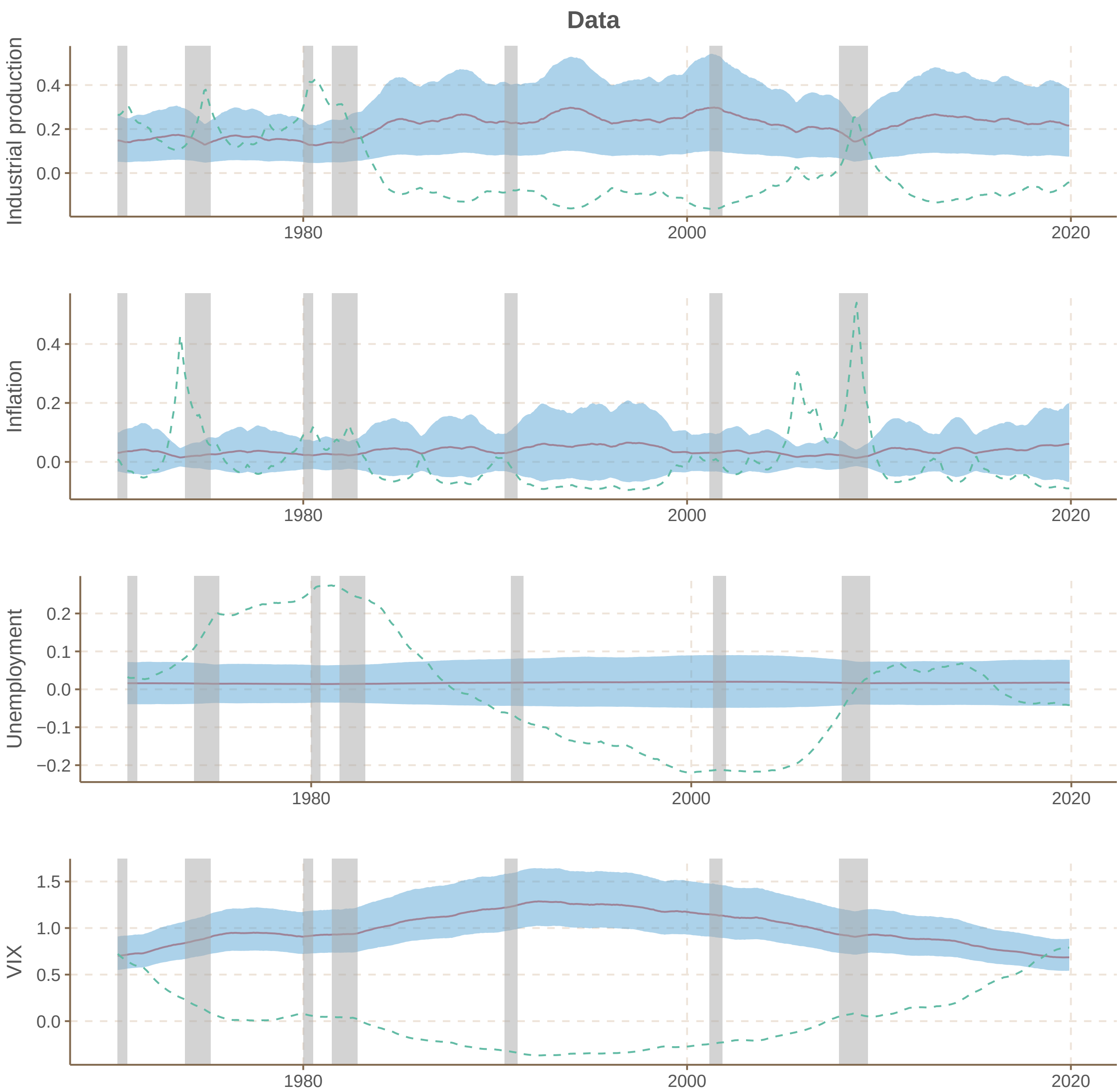

Figure 4: The time-varying skewness of the conditional distribution of the innovations with their 50% credible interval in the VAR model with a MST-SV. The dashed lines illustrate the scaled values of the time-varying volatility of the variables.

production up to 12 months ahead but the improvement is only statistically significant up to 6 months ahead. Heavy tails and skewness improves the forecasts of inflation and the unemployment rate, but not significantly. For the VIX we note a slight (insignificant) deterioration of the forecast performance with skewness and heavy tails. In the stochastic volatility VAR group, the advantage of skewness and heavy tail models over the Gaussian model diminishes. For inflation and the VIX a few specifications with skewness and/or heavy tails produce significantly worse forecasts. For unemployment there is an improvement overall when allowing for heavy tails and or skewness, but not significantly. The VAR models with skewed distributions are better in the long term point forecast for industrial production than their symmetric counterparts and/or the Gaussian VAR. The OST specification is significantly better than the symmetric OT specification for the 3 and 12 month forecasts of inflation.

Table 4 reports on the density forecasts using the relative improvements in LP over the Gaussian VAR models as the criterion. Here entries greater than 0 indicate that the given model is better than the Gaussian non-SV/SV model. As a result of the more careful modelling of the distribution of the innovations the heavy tailed and skewed specifications improve more on the density forecasts than the point forecasts. For the non-SV class of models heavy tails and skewness improve significantly on the Gaussian model forecasts of industrial production for all lead times and the 1 month forecasts of inflation while the forecasts for longer lead times are significantly worse. The forecasts of unemployment tend to be worse. For the VIX the improvement is significant for the 1 month forecasts and all lead times for the OST and MST specification while the forecasts from the symmetric specifications are significantly worse for longer lead times. In addition, we note a pattern where the OST and MST specifications improve significantly on their symmetric counterparts. Turning to the models with stochastic volatility we note a substantial improvement just by allowing for time varying variances. Comparing the SV models, the models with heavy tails and skewness improve significantly on the Gaussian model for the longer horizon

Table 3: Relative improvements in MSFE over the Gaussian VAR models

	1M	3M	6M	12M	1M	3M	6M	12M
	(a) Industrial Production				(b) Inflation			
Gaussian	0.393	0.413	0.465	0.471	0.080	0.116	0.114	0.115
Student-t	0.964*	0.967*	0.979	1.013	0.980	0.995	0.970	0.955*
Skew-t	0.973†	0.965*	0.960*†	0.981†	0.986	0.996	0.990†	1.006†
OT	0.954*	0.955*	0.975	1.002	0.988	1.019	0.998	0.972
MT	0.962*	0.961*	0.975	1.002	0.992	1.014	1.000	0.968
OST	0.961*†	0.954*	0.967†	0.988†	0.996†	1.022	1.003	0.976
MST	0.963*	0.968*†	0.979	0.998	0.994	0.996	1.011	0.989†
Gaussian-SV	0.373*	0.390*	0.448	0.477	0.080	0.116	0.108*	0.102*
Student-t-SV	1.004	1.004	1.001	1.003	0.996	0.997	0.992	0.984*
Skew-t-SV	1.007	1.007	1.001	0.994†	0.994	0.994	0.993	0.997
OT-SV	1.002	1.000	1.000	0.997	1.005	1.008*	1.005	1.002
MT-SV	1.004	1.004	1.003	0.997	1.006	1.006*	1.004	0.998
OST-SV	1.005	1.001	0.999	0.987*†	1.008*†	1.005†	1.001	0.993†
MST-SV	1.006	1.003	1.000	0.988†	1.010*	1.004	1.001	0.995
	(c) Unemployment rate				(d) VIX			
Gaussian	0.021	0.071	0.218	0.796	0.033	0.079	0.106	0.142
Student-t	0.986	0.973	0.971	1.017	1.014	1.014	1.009	1.003
Skew-t	0.986	0.962	0.944	0.980	1.019	1.022	1.016	1.034
OT	1.002	0.976	0.963	1.000	1.014	1.018	1.005	0.989
MT	0.996	0.983	0.980	1.013	1.015	1.017	1.007	0.988
OST	0.994†	0.974	0.955	0.988	1.027*	1.016	1.006	0.996
MST	0.998	0.981	0.976	1.008	1.019	1.001	1.034	1.001
Gaussian-SV	0.021	0.070	0.214	0.798	0.032	0.078	0.105	0.138
Student-t-SV	0.992	0.991	0.996	1.013	1.004	1.010	1.013	1.020
Skew-t-SV	0.990	0.984†	0.982	0.994	1.015	1.012	1.014	1.026
OT-SV	0.998	0.993	0.990	1.002	1.008	1.014	1.016	1.015
MT-SV	0.994	0.997	0.996	1.007	1.007	1.013	1.014	1.012
OST-SV	0.997	0.993	0.989	0.996	1.022*	1.013	1.010	1.015
MST-SV	0.997	0.999	0.994	1.001	1.021*	1.012	1.008	1.025

Each panel reports the MSFE of the models relative to the Gaussian VAR model with (and without) stochastic volatility. The relative improvements over the Gaussian models are computed as the ratio of the MSFE of alternative specifications over the Gaussian models during 2000-2019. We perform a two-sided Diebold and Mariano (1995) test. * denotes that the corresponding model is significantly different from the Gaussian VAR at the 10% level. † denotes that the skew Student model significantly different from the corresponding Student at the 10% level. The entries less than 1 indicate that the given model is better.

forecasts of industrial production. For inflation the improvement is small and insignificant while we observe a small and insignificant deterioration for the unemployment rate. The 1 month forecast of the VIX improves significantly for all heavy tailed and skew specifications as well as the 3 month forecasts for the OST and MST which also improves significantly on their symmetric counterparts. Overall we see cases with both better and worse forecasts than the Gaussian for models with heavy tails and skewness. Most of the improvement occurs for the forecasts of industrial production and the VIX where the in sample analysis shows clear signs of skewness while there is little or no improvement for inflation and unemployment where there are no signs of skewness. The density forecasts using the CRPS also report similar results, see details in Online Appendix F.

While there is no clear advantage for the more flexible specifications that allow for skewness and heavy tails in terms of the point forecasts once stochastic volatility is allowed for, the results are different for the density forecasts. Here the results are largely in-line with the in-sample evidence. As can be expected we see little or no improvement over the Gaussian-SV model for inflation and the unemployment rate where the in-sample analysis suggests that the innovation distribution is close to being Gaussian. For industrial production and the VIX the in-sample analysis shows that innovations have a fat-tailed and skew distribution and for these variables we find a notable improvement of the density forecasts with the OST-SV and MST-SV models.

Next, we concentrate on the effect of skewness parameters in VAR models with stochastic volatility. Figure 5 shows the cumulative log Bayes factors of the predictive density for the 3-month forecast horizon between the OT-SV and OST-SV models, see the computational details in Geweke and Amisano (2010). Positive values (red) mean that OST-SV predicts better than the OT-SV. A common feature across the variables is that the OST-SV performs better than or roughly on par with the OT-SV during recessions and crises. The OST-SV performs better for industrial production during the 2001 recession and for unemployment during the 2008 recession. As expected, skewness does not help with the short term forecast

Table 4: Improvement in LP over the Gaussian VAR models

	1M	3M	6M	12M	1M	3M	6M	12M
	(a) Industrial Production				(b) Inflation			
Gaussian	-1.005	-1.078	-1.153	-1.187	-0.378	-0.562	-0.626	-0.673
Student-*t*	0.044*	0.040*	0.040*	0.036*	0.026*	0.004	-0.016*	-0.029*
Skew-*t*	0.031†	0.032*†	0.036*	0.040*	0.011*†	-0.006†	-0.023*	-0.033*
OT	0.052*	0.041*	0.036*	0.031*	0.038*	0.007	-0.028*	-0.061*
MT	0.049*	0.034*	0.031	0.026*	0.038*	0.008	-0.027*	-0.057*
OST	0.051*	0.043*	0.043*†	0.043*†	0.038*	0.013†	-0.019*†	-0.050*†
MST	0.034*	0.033*	0.033*	0.036*†	0.032*†	0.019	-0.020*†	-0.046*†
Gaussian-SV	-0.850*	-0.883*	-0.984*	-1.024*	-0.031*	-0.242*	-0.243*	-0.247*
Student-*t*-SV	0.002	0.012	0.025*	0.031*	-0.001	0.007	0.016*	0.019*
Skew-*t*-SV	0.002	0.006	0.017*	0.030*	-0.007	-0.001	0.003†	0.006†
OT-SV	0.006	0.016	0.035*	0.035*	-0.001	-0.000	0.005	0.004
MT-SV	0.004	0.016	0.038*	0.041*	-0.002	0.001	0.005	0.003
OST-SV	0.006	0.018	0.021*	0.027*	-0.005†	0.001	0.005	0.007
MST-SV	0.001	0.012	0.028*	0.038*	-0.005	0.000	0.004	0.003
	(c) Unemployment rate				(d) VIX			
Gaussian	-0.066	-0.600	-1.079	-1.599	-0.112	-0.484	-0.646	-0.774
Student-*t*	-0.010*	-0.047*	-0.049*	-0.049*	0.020*	-0.034*	-0.047*	-0.055*
Skew-*t*	-0.011*	-0.037*†	-0.030*†	-0.017*†	0.019*	-0.011*†	-0.017*†	-0.027*
OT	-0.002*	-0.015*	-0.011*	-0.021*	0.026*	-0.023*	-0.037*	-0.047*
MT	0.001*	-0.018*	-0.017*	-0.026*	0.027*	-0.022*	-0.035*	-0.043*
OST	-0.001†	-0.009*†	-0.002†	-0.001†	0.053*†	0.041*†	0.039*†	0.029*†
MST	0.002*†	-0.012*†	-0.007*†	-0.007†	0.039*†	0.030*†	0.027*†	0.020*†
Gaussian-SV	0.522*	-0.016*	-0.528*	-1.257	0.327*	-0.135*	-0.309*	-0.472*
Student-*t*-SV	-0.006*	-0.022*	-0.040	-0.037	0.023*	0.002	0.000	-0.003
Skew-*t*-SV	-0.005*	-0.018	-0.031	-0.043	0.056*†	0.024	0.014	-0.010
OT-SV	-0.002	-0.018	-0.031	-0.021	0.028*	0.000	-0.003	-0.006
MT-SV	-0.004	-0.025	-0.037	-0.008	0.028*	0.001	-0.003	-0.003
OST-SV	-0.002	-0.012	-0.022	-0.029	0.071*†	0.035†	0.017	-0.013
MST-SV	-0.003	-0.025	-0.043	-0.004	0.065*†	0.034†	0.012	-0.018

Each panel reports the LP of the models relative to the Gaussian VAR model with (and without) stochastic volatility. The relative improvements over the Gaussian models are computed as the difference between the LP of alternative specifications and the Gaussian models during 2000-2019. We perform a two-sided Diebold and Mariano (1995) test. * denotes that the corresponding model is significantly different from the Gaussian VAR at the 10% level. † denotes that the skew Student model significantly different from the corresponding Student at the 10% level. The entries greater than 0 indicate that the given model is better.

of inflation. For the VIX the OST-SV also improves its performance during the expansion and does significantly better overall. Hence, skewness of the distribution is a value-added feature to the VAR model with heavy tails.

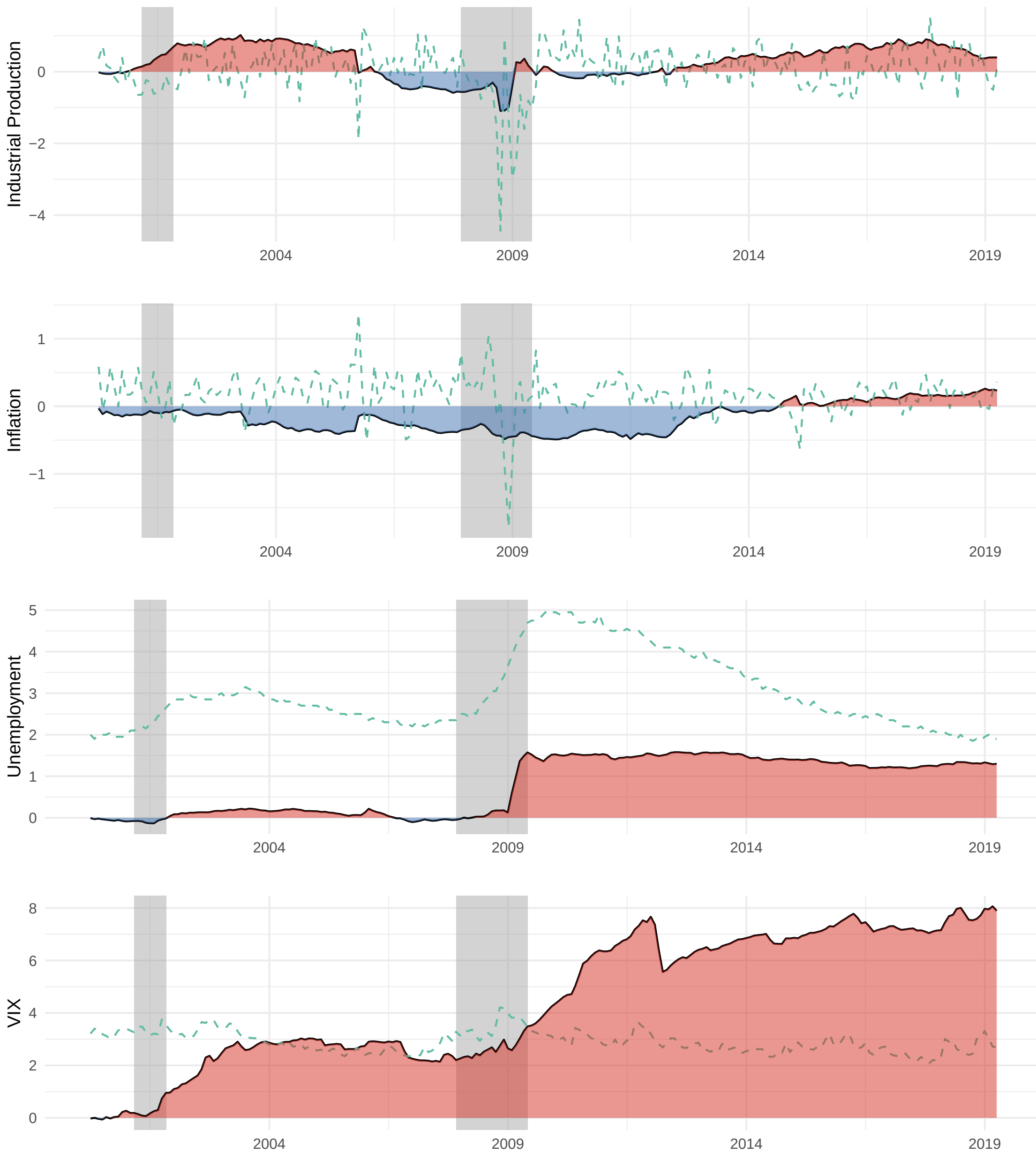

Figure 5: Cumulative log Bayes factors of the predictive density for the 3-month ahead forecast between the OT-SV and OST-SV models.

Positive values (red) means OST-SV predicts better and negative values (blue) means that OT-SV model does better. The dashed lines illustrate the scaled values of the original variables. See Geweke and Amisano (2010) for details.

5 Conclusion

Skewness and heavy tails are empirically relevant features in many application areas – not only the macroeconomic and financial application we consider in this paper. While these features to some extent can be accommodated or masked by time-varying heteroskedasticity modelled as GARCH-type or stochastic volatility processes there is a need for models that explicitly account for skewness and heavy tails in the data. We contribute to this by proposing flexible skew and heavy tailed distributions with the symmetric normal distribution as a special case. Specifically, we introduce a general class of Generalized Hyperbolic Skew Student's t distributions with stochastic volatility for VAR models. The stochastic representation of the GHSkew-t can be written in term of a variance-mean mixture which leads to a straightforward implementation of a Gibbs sampler for posterior inference. We show how model comparison and choice can be conducted using the cross entropy methods of Chan and Eisenstat (2018) or the Chib and Jeliazkov (2001) method to calculate the model marginal likelihood and compare the in-sample fit among different specifications. In an application to US data we find support for VAR models with skewness and heavy tails. The VAR models with skewness and heavy tails gives better point forecasts and density forecasts compared to Gaussian VAR models for many, but not all, variables we model. Crucially, in sample measures such as the marginal likelihood or the posterior distribution of the skewness parameters and degrees of freedom are informative about for which variables the forecasts can benefit from allowing for skewness and/or heavy tails. We recommend that skewness should be taken into account for improving forecasting performance.

Acknowledgment

We thank Pär Österholm, the editor and two anonymous referees for helpful comments. The authors acknowledge financial support from the project "Models for macro and financial economics after the financial crisis" (Dnr: P18-0201, BV18-0018, BFV22-0005) funded by the Jan Wallander and Tom Hedelius Foundation. Stepan Mazur also acknowledges financial support from the internal research grants of Örebro University. The computations were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC) at HPC2N partially funded by the Swedish Research Council through grant agreement no. 2018-05973.

Appendix

A Properties of the bivariate MST and OST distributions

Consider the bivariate vector of innovations $\mathbf{u}_t = (u_{1t}, u_{2t})'$ given in Section 2.4. We analyse how the the skewness and coskewness of bivariate MST and OST distributions depends on the distributional parameters. Following the discussion in Section 2.4, Figure 6 shows the skewness and coskewness of the bivariate MST and OST distributions with $\rho = 0.7$, $\boldsymbol{\gamma} = (-1, 2)$, $\boldsymbol{\nu} = (9, 12)$, $\mathbf{h}_t = (1, 1)$. In each row we let the parameter of interest vary around these values. In both distributions, the skewness of u_{1t} is insensitive to the parameters of the second variable, u_{2t} as well as the correlation. The MST and OST distributions differ in the distribution of u_{2t} where the first variable can induce skewness of u_{2t} in the OST distribution even when $\gamma_2 = 0$. In the MST distribution, the $coskew_1$ and $coskew_2$ oscillate mildly close to zero while those of the OST distribution vary in a greater extent.

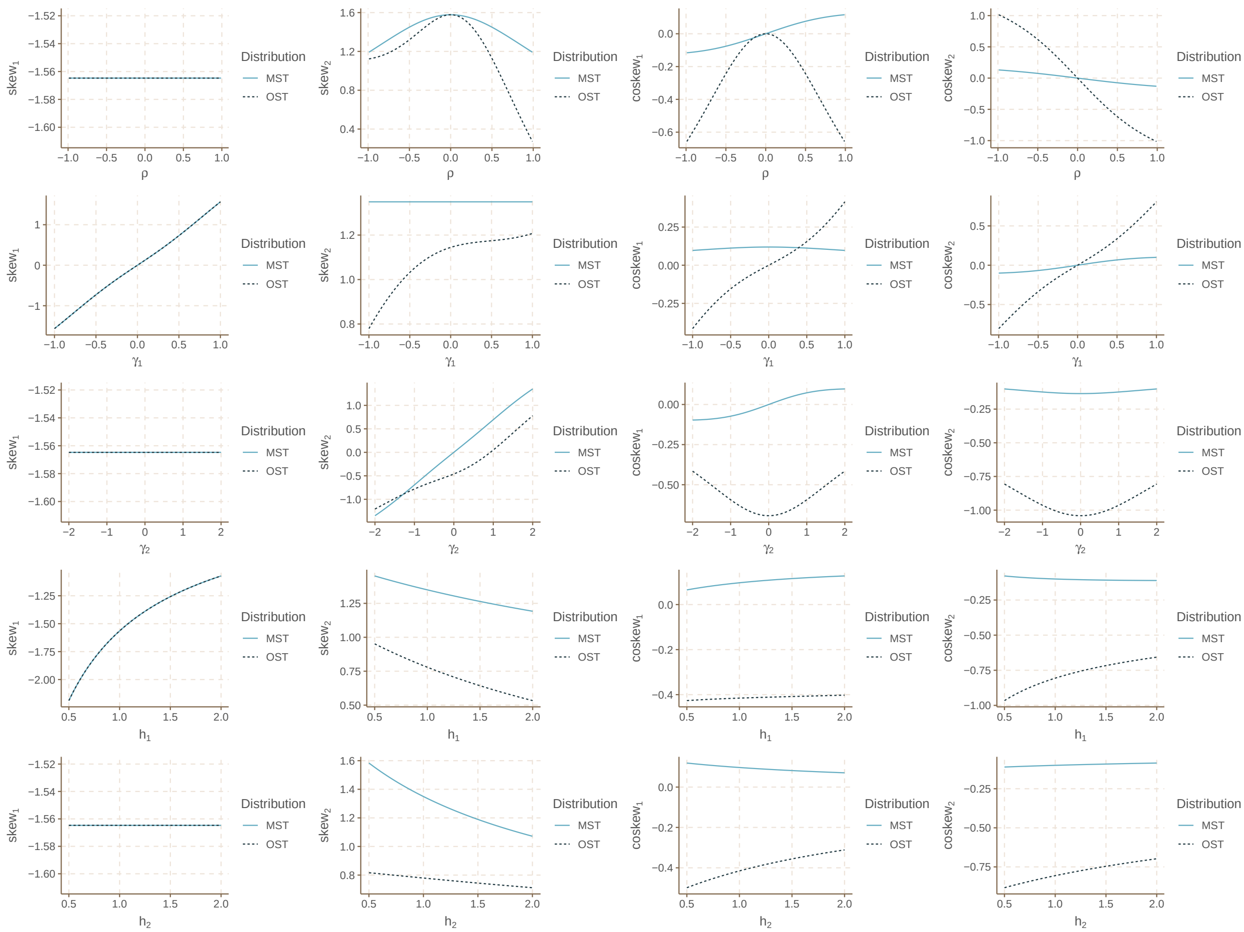

Figure 6: Skewness and coskewness of the bivariate MST and OST distributions.

B Estimation Procedure

Given the latent variables $\boldsymbol{\xi}_{1:T}$ and the skewness parameters $\boldsymbol{\gamma}$, the conditional posterior distributions of the remaining parameters in the VAR-MST-SV model are similar to those in the VAR-Gaussian-SV model. Hence, the MST model can be estimated using a six-step Metropolis-within-Gibbs Markov chain Monte Carlo (MCMC) algorithm. Let $\boldsymbol{\Psi}$ be all the parameters and latent variables in $\boldsymbol{\theta}$ except the ones we sample from in a given step of the MCMC procedure.

1. In order to sample from $\pi(\boldsymbol{b}|\boldsymbol{\Psi})$ where $\boldsymbol{b} = (\text{vec}(\mathbf{B})', \boldsymbol{\gamma}')'$, we rewrite the MST-SV VAR as a multivariate linear regression,

$$\begin{aligned}
\mathbf{y}_t &= \mathbf{B}\mathbf{x}_t + (\mathbf{W}_t - \overline{\mathbf{W}})\boldsymbol{\gamma} + \mathbf{W}_t^{1/2}\mathbf{A}^{-1}\mathbf{H}_t^{1/2}\boldsymbol{\epsilon}_t, \\
\mathbf{A}\mathbf{W}_t^{-1/2}\mathbf{y}_t &= \mathbf{A}\mathbf{W}_t^{-1/2}\mathbf{B}\mathbf{x}_t + \mathbf{A}\mathbf{W}_t^{-1/2}(\mathbf{W}_t - \overline{\mathbf{W}})\boldsymbol{\gamma} + \mathbf{H}_t^{1/2}\boldsymbol{\epsilon}_t \\
&= \mathbf{x}_t' \otimes \mathbf{A}\mathbf{W}_t^{-1/2}\,\text{vec}(\mathbf{B}) + \mathbf{A}\mathbf{W}_t^{-1/2}(\mathbf{W}_t - \overline{\mathbf{W}})\boldsymbol{\gamma} + \mathbf{H}_t^{1/2}\boldsymbol{\epsilon}_t \\
&= (\mathbf{x}_t' \otimes \mathbf{A}\mathbf{W}_t^{-1/2} \quad \mathbf{A}\mathbf{W}_t^{-1/2}(\mathbf{W}_t - \overline{\mathbf{W}}))\boldsymbol{b} + \mathbf{H}_t^{1/2}\boldsymbol{\epsilon}_t, \\
\widetilde{\mathbf{y}}_t &= \widetilde{\mathbf{X}}_t\boldsymbol{b} + \mathbf{H}_t^{1/2}\boldsymbol{\epsilon}_t,
\end{aligned}$$

   where $\widetilde{\mathbf{y}}_t = \mathbf{A}\mathbf{W}_t^{-1/2}\mathbf{y}_t$ and $\widetilde{\mathbf{X}}_t = (\mathbf{x}_t' \otimes \mathbf{A}\mathbf{W}_t^{-1/2} \quad \mathbf{A}\mathbf{W}_t^{-1/2}(\mathbf{W}_t - \overline{\mathbf{W}}))$. Then the conditional posterior distribution of $\boldsymbol{b}$ is a conjugate Gaussian distribution

$$\pi(\boldsymbol{b}|\boldsymbol{\Psi}) \sim \mathcal{N}(\mathbf{b}^*, \mathbf{V}_{\mathbf{b}}^*),$$

   where

$$\begin{aligned}
\mathbf{V}_{\mathbf{b}}^{*-1} =& \mathbf{V}_{\mathbf{b}_0}^{-1} + \sum_{t=1}^{T} \widetilde{\mathbf{X}}_t'\mathbf{H}_t^{-1}\widetilde{\mathbf{X}}_t, \\
\mathbf{b}^* =& \mathbf{V}_{\mathbf{b}}^* \left[\mathbf{V}_{\mathbf{b}_0}^{-1}\mathbf{b}_0 + \sum_{t=1}^{T} \widetilde{\mathbf{X}}_t'\mathbf{H}_t^{-1}\widetilde{\mathbf{y}}_t\right].
\end{aligned}$$

2. In order to sample from $\pi(\mathbf{a}|\boldsymbol{\Psi})$, we follow Cogley and Sargent (2005) and use that (6)

is a triangular model for the reduced form residuals,

$$\mathbf{A}\widetilde{\mathbf{u}}_t = \mathbf{H}_t^{1/2}\boldsymbol{\epsilon}_t,$$

where $\widetilde{\mathbf{u}}_t = \mathbf{W}_t^{-1/2}(\mathbf{y}_t - \mathbf{B}\mathbf{x}_t - (\mathbf{W}_t - \overline{\mathbf{W}})\boldsymbol{\gamma})$. This reduces to a system of linear equations with equation i that has $\widetilde{u}_{it}$ as a dependent variable and $-\widetilde{u}_{jt}$ as independent variables with coefficients a_{ij} for $i = 2, \ldots, k$ and $j = 1, \ldots, i-1$. By multiplying both sides of the equations with $h_{it}^{-1/2}$, we can eliminate the effect of heteroscedasticity. Then, draws from the conditional posterior of a_{ij} can be taken equation by equation using the conditionally Gaussian posterior distribution (Cogley and Sargent, 2005).

3. In order to sample from $\pi(\mathbf{h}_{0:T}|\boldsymbol{\Psi})$, we follow Kim et al. (1998); Primiceri (2005); Del Negro and Primiceri (2015). Let $\widetilde{\widetilde{\mathbf{u}}}_t = \mathbf{A}\widetilde{\mathbf{u}}_t$, for each series $i = 1, \ldots, k$, we have that $\log \widetilde{\widetilde{u}}_{it}^2 = \log h_{it} + \log \epsilon_t^2$. Kim et al. (1998) approximated the distribution of $\log(\epsilon_t^2)$ as $\log(\chi^2)$ using a mixture of 7 Gaussian components. Then using forward filter backward smoothing algorithm in Carter and Kohn (1994), we sample $\log h_{it}$ from its smoothing Gaussian distribution.

4. In order to sample from $\pi(\boldsymbol{\sigma}^2|\boldsymbol{\Psi})$, Equation (4) describes a random walk in the logarithm of the volatility. The conditional posterior $\pi(\sigma_i^2|\boldsymbol{\Psi})$ is generalized inverse Gaussian (GIG) and given by

$$\pi(\sigma_i^2|\boldsymbol{\Psi}) \propto \left(\sigma_i^2\right)^{-\frac{T}{2}} \exp\left(-\frac{\sum_{t=1}^T (\log h_{it} - \log h_{it-1})^2}{2\sigma_i^2}\right) \left(\sigma_i^2\right)^{-\frac{1}{2}} \exp\left(-\frac{\sigma_i^2}{2V_\sigma}\right).$$

We sample $\sigma_i^2 \sim GIG(\lambda, \psi, \chi)$ where $\lambda = -0.5(T-1)$, $\chi = \sum_{t=1}^{T} (\log h_{i,t} - \log h_{i,t-1})^2$ and $\psi = 1/V_\sigma$, see Hörmann and Leydold (2014) for more details.

5. In order to sample from $\pi(\nu_i|\boldsymbol{\Psi}) \propto \mathcal{G}(\nu_i; 2, 0.1) \prod_{t=1}^{T} \mathcal{IG}\left(\xi_{it}; \frac{\nu_i}{2}, \frac{\nu_i}{2}\right)$ for $i = 1, \ldots, k$, we use an adaptive random walk Metropolis-Hastings algorithm to accept/reject the draw

$\nu_i^{(*)} = \nu_i + \eta_i \exp(c_i)$, where $\eta_i \sim \mathcal{N}(0,1)$ and the adaptive variance c_i is adjusted automatically such that the acceptance rate is around 0.25 (Roberts and Rosenthal, 2009).

6. In order to sample $\pi(\xi_t|\mathbf{\Psi})$ for $t = 1,\dots,T$, we apply the independent Metropolis-Hastings algorithm to draw $\xi_{it}^{(*)} \sim \mathcal{IG}(\alpha_{it}, \beta_{it})$ for $i = 1,\dots,k$ and accept with the probability

$$\min\left\{1, \frac{\pi(\mathbf{W}_t^{(*)}|\mathbf{\Psi})\prod_{i=1}^{k}\mathcal{IG}(\xi_{it};\alpha_{it},\beta_{it})}{\pi(\mathbf{W}_t|\mathbf{\Psi})\prod_{i=1}^{k}\mathcal{IG}(\xi_{it}^{(*)};\alpha_{it},\beta_{it})}\right\}$$

where

$$\pi(\mathbf{W}_t|\mathbf{\Psi}) \propto \prod_{i=1}^{k}\xi_{it}^{-1/2}\exp\left(-\frac{1}{2}(\mathbf{y}_t - \mathbf{B}\mathbf{x}_t - \mathbf{W}_t\boldsymbol{\gamma} + \overline{\mathbf{W}}\boldsymbol{\gamma})'\mathbf{\Omega}_t^{-1}(\mathbf{y}_t - \mathbf{B}\mathbf{x}_t - \mathbf{W}_t\boldsymbol{\gamma} + \overline{\mathbf{W}}\boldsymbol{\gamma})\right)$$
$$\mathcal{IG}\left(\xi_{it};\frac{\nu_i}{2},\frac{\nu_i}{2}\right).$$

where $\mathbf{\Omega}_t = \mathbf{W}_t^{1/2}\mathbf{A}^{-1}\mathbf{H}_t\mathbf{A}^{-1\prime}\mathbf{W}_t^{1/2}$. The proposal distribution $\mathcal{IG}(\alpha_{it}, \beta_{it})$ is taken from Chiu et al. (2017) with $\alpha_{it} = \frac{c}{2}(\nu_i + 1)$ and $\beta_{it} = \frac{c}{2}\left(\nu_i + \frac{\tilde{\tilde{u}}_{it}^2}{h_{it}}\right)$ where the constant $c = 0.75$ is adjusted so that the acceptance rate range from 20% to 80%.

C Details of the marginal likelihood estimation

C.1 The integrated likelihood

The integrated likelihood $p(\mathbf{y}_{1:T}|\boldsymbol{\theta}_1)$ with $\boldsymbol{\theta}_1 = \{\mathbf{B}, \mathbf{a}, \boldsymbol{\gamma}, \boldsymbol{\nu}, \boldsymbol{\sigma}^2\}$ require a high dimensional integral over the latent states $\boldsymbol{\theta}_2 = \{\boldsymbol{\xi}_{1:T}, \mathbf{h}_{0:T}\}$,

$$p(\mathbf{y}_{1:T}|\boldsymbol{\theta}_1) = \int\int p(\mathbf{y}_{1:T}|\boldsymbol{\theta}_1, \boldsymbol{\xi}_{1:T}, \mathbf{h}_{0:T})p(\boldsymbol{\xi}_{1:T}, \mathbf{h}_{0:T}|\boldsymbol{\theta}_1)d\boldsymbol{\xi}_{1:T}d\mathbf{h}_{0:T}.$$

In the VAR models with Gaussian, Student-t, Skew-t, OT, OST innovation, the conditional likelihood, $p(\mathbf{y}_{1:T}|\boldsymbol{\theta}_1, \mathbf{h}_{0:T})$, can be derived in a closed form. The integrated likelihood becomes,

$$p(\mathbf{y}_{1:T}|\boldsymbol{\theta}_1) = \int p(\mathbf{y}_{1:T}|\boldsymbol{\theta}_1, \mathbf{h}_{0:T}) p(\mathbf{h}_{0:T}|\boldsymbol{\theta}_1) d\mathbf{h}_{0:T}.$$

We propose an importance sampling algorithm to approximate the integrated likelihood following Chan and Eisenstat (2018).

$$p(\mathbf{y}_{1:T}|\boldsymbol{\theta}_1) \approx \sum_{l=1}^{L} \frac{1}{L} \frac{p(\mathbf{y}_{1:T}|\boldsymbol{\theta}_1, \mathbf{h}_{0:T}^{(l)}) p(\mathbf{h}_{0:T}^{(l)}|\boldsymbol{\theta}_1)}{f(\mathbf{h}_{0:T}^{(l)}|\boldsymbol{\lambda}_H)} = \sum_{l=1}^{L} \frac{1}{L} \frac{p(\mathbf{y}_{1:T}|\boldsymbol{\theta}_1, \widetilde{\mathbf{h}}_{0:T}^{(l)}) p(\widetilde{\mathbf{h}}_{0:T}^{(l)}|\boldsymbol{\theta}_1)}{f(\widetilde{\mathbf{h}}_{0:T}^{(l)}|\boldsymbol{\lambda}_H)},$$

where $\widetilde{\mathbf{h}}_{0:T}^{(l)} = \log \mathbf{h}_{0:T}^{(l)}$ is sampled from the importance sampling distribution $f(\widetilde{\mathbf{h}}_{0:T}|\boldsymbol{\lambda}_H)$ for $l = 1, \ldots, L$, and $L = 100$. We take the importance function $f(\widetilde{\mathbf{h}}_{0:T}|\boldsymbol{\lambda}_H)$ to be a multivariate normal distribution with mean and precision matrix $\boldsymbol{\lambda}_H = \{\widehat{\mathbf{h}}_{1:T}, \widehat{\boldsymbol{\Sigma}}_H^{-1}\}$ chosen as

$$\begin{aligned}
\widehat{\mathbf{h}}_{1:T} &= \arg\max_{\widetilde{\mathbf{h}}_{0:T}} \log p(\mathbf{y}_{1:T}|\boldsymbol{\theta}_1, \widetilde{\mathbf{h}}_{0:T}) p(\widetilde{\mathbf{h}}_{0:T}|\boldsymbol{\theta}_1), \\
\widehat{\boldsymbol{\Sigma}}_H^{-1} &= - \frac{\partial^2 \log p(\mathbf{y}_{1:T}|\boldsymbol{\theta}_1, \widetilde{\mathbf{h}}_{0:T}) p(\widetilde{\mathbf{h}}_{0:T}|\boldsymbol{\theta}_1)}{\partial \widetilde{\mathbf{h}}_{0:T}^2} \Bigg|_{\widetilde{\mathbf{h}}_{0:T} = \widehat{\mathbf{h}}_{1:T}}.
\end{aligned}$$

The VAR models with MST and MST-SV distributions do not have a closed form expression for $p(\mathbf{y}_{1:T}|\boldsymbol{\theta}_1, \mathbf{h}_{0:T})$, instead we approximate the integrated likelihood using a two-stage importance sampling algorithm,

$$\begin{aligned}
p(\mathbf{y}_{1:T}|\boldsymbol{\theta}_1) &= \int \left[\int p(\mathbf{y}_{1:T}|\boldsymbol{\theta}_1, \boldsymbol{\xi}_{1:T}, \mathbf{h}_{0:T}) p(\boldsymbol{\xi}_{1:T}|\mathbf{h}_{0:T}, \boldsymbol{\theta}_1) d\boldsymbol{\xi}_{1:T} \right] p(\mathbf{h}_{0:T}|\boldsymbol{\theta}_1) d\mathbf{h}_{0:T}, \\
&\approx \sum_{l=1}^{L} \frac{1}{L} \frac{\left[\sum_{m=1}^{M} \frac{1}{M} \frac{p(\mathbf{y}_{1:T}|\boldsymbol{\theta}_1, \mathbf{h}_{0:T}^{(l)}, \boldsymbol{\xi}_{1:T}^{(m)}) p(\boldsymbol{\xi}_{1:T}^{(m)}|\boldsymbol{\theta}_1)}{f(\boldsymbol{\xi}_{1:T}^{(m)}|\boldsymbol{\lambda}_W)} \right] p(\widetilde{\mathbf{h}}_{0:T}^{(l)}|\boldsymbol{\theta}_1)}{f(\widetilde{\mathbf{h}}_{0:T}^{(l)}|\boldsymbol{\lambda}_H)}
\end{aligned}$$

where the importance function $f(\boldsymbol{\xi}_{1:T}^{(m)}|\boldsymbol{\lambda}_W)$ is the Metropolis-Hasting proposal from step 6 of the MCMC scheme in Appendix B with $M = 100$ and the importance function $f(\widetilde{\mathbf{h}}_{0:T}^{(l)}|\boldsymbol{\lambda}_H)$

is a multivariate normal distribution with mean and precision matrix $\boldsymbol{\lambda}_H = \{\widehat{\mathbf{h}}_{1:T}, \widehat{\boldsymbol{\Sigma}}_H^{-1}\}$ selected in a similar way to the other models. As the conditional likelihood $p(\mathbf{y}_{1:T}|\boldsymbol{\theta}_1, \mathbf{h}_{0:T})$ is not available in closed form we approximate it using a product of skew-t distribution for the purpose of finding $\boldsymbol{\lambda}_H$,

$$p(\mathbf{y}_{1:T}|\widetilde{\mathbf{h}}_{0:T}, \boldsymbol{\theta}_1) = p(\mathbf{u}_{1:T}|\widetilde{\mathbf{h}}_{0:T}, \boldsymbol{\theta}_1) = \prod_{t=1}^{T} p(\mathbf{u}_t|\widetilde{\mathbf{h}}_t, \boldsymbol{\theta}_1) = \prod_{t=1}^{T} p(\mathbf{A}\mathbf{u}_t|\widetilde{\mathbf{h}}_t, \boldsymbol{\theta}_1)/|\mathbf{A}|,$$

$$\approx \prod_{t=1}^{T}\prod_{i=1}^{k} p_{skew-t}(\widetilde{e}_{it}|\mu = 0, \sigma^2 = h_{it}, \nu = \nu_i, \gamma = \gamma_i) = \widetilde{p}(\mathbf{y}_{1:T}|\widetilde{\mathbf{h}}_{0:T}, \boldsymbol{\theta}_1),$$

as the marginal distribution of the elements of $\mathbf{u}_t$ is GHskew-t and the transformation with $\mathbf{A}$ brings them to being close to uncorrelated. The parameters of the multivariate normal importance function are thus chosen as

$$\widehat{\mathbf{h}}_{1:T} = \arg\max_{\widetilde{\mathbf{h}}_{0:T}} \log \widetilde{p}(\mathbf{y}_{1:T}|\boldsymbol{\theta}_1, \widetilde{\mathbf{h}}_{0:T}) p(\widetilde{\mathbf{h}}_{0:T}|\boldsymbol{\theta}_1),$$

$$\widehat{\boldsymbol{\Sigma}}_H^{-1} = -\left.\frac{\partial^2 \log \widetilde{p}(\mathbf{y}_{1:T}|\boldsymbol{\theta}_1, \widetilde{\mathbf{h}}_{0:T}) p(\widetilde{\mathbf{h}}_{0:T}|\boldsymbol{\theta}_1)}{\partial \widetilde{\mathbf{h}}_{0:T}^2}\right|_{\widetilde{\mathbf{h}}_{0:T} = \widehat{\mathbf{h}}_{1:T}}.$$

C.2 Chib and Jeliazkov method

The Chib and Jeliazkov (2001) method is based on the basic marginal likelihood identity

$$p(\mathbf{y}_{1:T}) = \frac{p(\mathbf{y}_{1:T}|\mathbf{B}^*, \boldsymbol{\gamma}^*, \mathbf{A}^*, \boldsymbol{\sigma}^{2*}, \boldsymbol{\nu}^*) p(\mathbf{B}^*, \boldsymbol{\gamma}^*, \mathbf{A}^*, \boldsymbol{\sigma}^{2*}, \boldsymbol{\nu}^*)}{p(\mathbf{B}^*, \boldsymbol{\gamma}^*, \mathbf{A}^*, \boldsymbol{\sigma}^{2*}, \boldsymbol{\nu}^*|\mathbf{y}_{1:T})}$$

where $B^*, \gamma^*, A^*, \sigma^{2*}, \nu^*$ are the posterior means of the corresponding parameters.

The prior $p(\mathbf{B}^*, \boldsymbol{\gamma}^*, \mathbf{A}^*, \boldsymbol{\sigma}^{2*}, \boldsymbol{\nu}^*)$ is available in closed form. The rest of the algorithm works by first running a complete MCMC chain. This is used to estimate the integrated likelihood, $p(\mathbf{y}_{1:T}|\mathbf{B}^*, \boldsymbol{\gamma}^*, \mathbf{A}^*, \boldsymbol{\sigma}^{2*}, \boldsymbol{\nu}^*)$, using the technique outlined in Appendix C.1. The posterior is then decomposed into a sequence of conditional densities,

$$p(\mathbf{B}^*, \boldsymbol{\gamma}^*, \mathbf{A}^*, \boldsymbol{\sigma}^{2*}, \boldsymbol{\nu}^*|\mathbf{y}_{1:T}) = p(\boldsymbol{\nu}^*|\mathbf{y}_{1:T})p(\boldsymbol{\sigma}^{2*}|\boldsymbol{\nu}^*, \mathbf{y}_{1:T})p(\mathbf{A}^*|\boldsymbol{\sigma}^{2*}, \boldsymbol{\nu}^*, \mathbf{y}_{1:T})$$
$$\times\, p(\mathbf{B}^*, \boldsymbol{\gamma}^*|\mathbf{A}^*, \boldsymbol{\sigma}^{2*}, \boldsymbol{\nu}^*, \mathbf{y}_{1:T})$$

which are evaluated in turn using reduced MCMC runs fixing the parameters at the posterior means and thereby returning draws from conditional posteriors.

The posterior distribution of $\boldsymbol{\nu}$ is not available in closed form and the elements, ν_i are sampled in separate M-H steps. To evaluate $p(\boldsymbol{\nu}^*|\mathbf{y}_{1:T})$ we further decompose this into $p(\boldsymbol{\nu}^*|\mathbf{y}_{1:T}) = \prod_{i=1}^{k} p(\nu_i^*|\boldsymbol{\nu}_{i-1}^*, \mathbf{y}_{1:T})$ where $\boldsymbol{\nu}_j = (\nu_1, \ldots, \nu_j)$. Following Chib and Jeliazkov (2001) we can then express the posterior ordinates as

$$p(\nu_i^*|\boldsymbol{\nu}_{i-1}^*, \mathbf{y}_{1:T}) = \frac{E_1[\alpha(\nu_i, \nu_i^*|\boldsymbol{\nu}_{i-1}^*, \mathbf{y}_{1:T})q(\nu_i, \nu_i^*|\boldsymbol{\nu}_{i-1}^*, \mathbf{y}_{1:T})]}{E_2[\alpha(\nu_i^*, \nu_i|\boldsymbol{\nu}_{i-1}^*, \mathbf{y}_{1:T})]}$$

where $\alpha(\nu_i, \nu_i^*|\boldsymbol{\nu}_{i-1}^*, \mathbf{y}_{1:T})$ is the acceptance ratio and $q(\nu_i, \nu_i^*|\boldsymbol{\nu}_{i-1}^*, \mathbf{y}_{1:T})$ the proposal distribution from the M-H step for a move from ν_i to ν_i^* in step 6 of section B, the expectation E_1 is with respect to the conditional posterior $p(\mathbf{B}, \boldsymbol{\gamma}, \mathbf{A}, \boldsymbol{\sigma}^2, \boldsymbol{\nu}^i|\boldsymbol{\nu}_{i-1}^*, \mathbf{y}_{1:T})$ for $\boldsymbol{\nu}^j = (\nu_j, \ldots, \nu_k)$ and the expectation E_2 with respect to the distribution $q(\nu_i^*, \nu_i|\boldsymbol{\nu}_{i-1}^*, \mathbf{y}_{1:T})p(\mathbf{B}, \boldsymbol{\gamma}, \mathbf{A}, \boldsymbol{\sigma}^2, \boldsymbol{\nu}^{i+1}|\boldsymbol{\nu}_i^*, \mathbf{y}_{1:T})$. Draws from $p(\mathbf{B}, \boldsymbol{\gamma}, \mathbf{A}, \boldsymbol{\sigma}^2, \boldsymbol{\nu}^i|\boldsymbol{\nu}_{i-1}^*, \mathbf{y}_{1:T})$ are obtained by running the MCMC chain with $\boldsymbol{\nu}_{i-1}$ fixed at $\boldsymbol{\nu}_{i-1}^*$ and draws from $q(\nu_i^*, \nu_i|\boldsymbol{\nu}_{i-1}^*, \mathbf{y}_{1:T})p(\mathbf{B}, \boldsymbol{\gamma}, \mathbf{A}, \boldsymbol{\sigma}^2, \boldsymbol{\nu}^{i+1}|\boldsymbol{\nu}_i^*, \mathbf{y}_{1:T})$ are obtained by running the chain with $\boldsymbol{\nu}_i$ fixed at $\boldsymbol{\nu}_i^*$ and generating a proposal ν_i from $q(\nu_i^*, \nu_i|\boldsymbol{\nu}_i^*, \mathbf{y}_{1:T})$ for each draw from the chain. $p(\nu_i^*|\boldsymbol{\nu}_{i-1}^*, \mathbf{y}_{1:T})$ is then estimated as

$$\widehat{p}(\nu_i^*|\boldsymbol{\nu}_{i-1}^*, \mathbf{y}_{1:T}) = \frac{\frac{1}{R}\sum_{l=1}^{R} \alpha(\nu_i^{(l)}, \nu_i^*|\boldsymbol{\nu}_{i-1}^*, \mathbf{y}_{1:T})q(\nu_i^{(l)}, \nu_i^*|\boldsymbol{\nu}_{i-1}^*, \mathbf{y}_{1:T})}{\frac{1}{R}\sum_{j=1}^{R} \alpha(\nu_i^*, \nu_i^{(j)}|\boldsymbol{\nu}_{i-1}^*, \mathbf{y}_{1:T})}$$

Table 5: LML for VAR models with and without SV, Chib and Jeliazkov method

		Gaussian	Student-t	Skew-t	OT	MT	OST	MST
Non SV	LML	-217.634	-128.773	-137.202	-128.583	-125.844	-128.200	-125.207
	sd	(0.004)	(0.053)	(0.018)	(0.200)	(0.314)	(0.223)	(0.308)
SV	LML	-46.783	-28.029	-26.053	-30.223	-27.992	-21.267	-19.515
	sd	(0.988)	(2.218)	(1.530)	(1.152)	(1.379)	(1.274)	(0.799)

We compare the LMLs of 14 VAR models with and without SV. In the Chib and Jeliazkov method, we calculate the LMLs using 5 runs. In each run, we estimate the models with 100,000 samples. Then we estimate the LLP with 1,000 samples, the $P1$-$P5$ with 20,000 samples and 10,000 burn-in.

To estimate $p(\boldsymbol{\sigma}^{2*}|\boldsymbol{\nu}^*, \mathbf{y}_{1:T})$ run the MCMC chain with $\boldsymbol{\nu}$ fixed at $\boldsymbol{\nu}^*$ and calculate

$$\widehat{p}(\boldsymbol{\sigma}^{2*}|\boldsymbol{\nu}^*, \mathbf{y}_{1:T}) = \frac{1}{R}\sum_{i=1}^{R} p(\boldsymbol{\sigma}^{2*}, |\mathbf{B}^{(i)}, \boldsymbol{\gamma}^{(i)}, \mathbf{A}^{(i)}, \boldsymbol{\nu}^*, \mathbf{y}_{1:T}).$$

Similarly $p(\mathbf{A}^*|\boldsymbol{\sigma}^{2*}, \boldsymbol{\nu}^*, \mathbf{y}_{1:T})$ is estimated by additionally fixing $\boldsymbol{\sigma}^2$ at $\boldsymbol{\sigma}^{2*}$ and averaging the full conditional posterior evaluated at $\mathbf{A}^*$ over the MCMC draws of the reduced chain. Likewise for $p(\mathbf{B}^*, \boldsymbol{\gamma}|\mathbf{A}^*, \boldsymbol{\sigma}^{2*}, \boldsymbol{\nu}^*, \mathbf{y}_{1:T})$.

Table 5 shows estimated log marginal likelihoods using the Chib and Jeliazkov method. While they differ slightly from the estimates using the cross-entropy method reported in Table 2 the ranking of the models is the same and the log Bayes factors are very close. The methods thus produce consistent results.

References

Aas, K. and Haff, I. H. (2006). The generalized hyperbolic skew Student's t-distribution. *Journal of Financial Econometrics*, 4(2):275–309.

Acemoglu, D., Ozdaglar, A., and Tahbaz-Salehi, A. (2017). Microeconomic origins of macroeconomic tail risks. *American Economic Review*, 107(1):54–108.

Carriero, A., Clark, T. E., and Marcellino, M. (2020). Capturing macroeconomic tail risks with Bayesian vector autoregressions. Working Papers 20-02R, Federal Reserve Bank of Cleveland.

Carriero, A., Clark, T. E., and Marcellino, M. (2021a). Using time-varying volatility for identification in vector autoregressions: An application to endogenous uncertainty. *Journal of Econometrics*, 225(1):47–73. Themed Issue: Vector Autoregressions.

Carriero, A., Clark, T. E., Marcellino, M., and Mertens, E. (2021b). Addressing COVID-19 Outliers in BVARs with Stochastic Volatility. Working Papers 21-02, Federal Reserve Bank of Cleveland.

Carter, C. K. and Kohn, R. (1994). On gibbs sampling for state space models. *Biometrika*, 81(3):541–553.

Chan, J. C. and Eisenstat, E. (2018). Bayesian model comparison for time-varying parameter VARs with stochastic volatility. *Journal of Applied Econometrics*, 33(4):509–532.

Chan, J. C., Koop, G., and Yu, X. (2021). Large order-invariant Bayesian VARS with stochastic volatility. *arXiv preprint arXiv:2111.07225*.

Chib, S. and Jeliazkov, I. (2001). Marginal likelihood from the metropolis–hastings output. *Journal of the American Statistical Association*, 96(453):270–281.

Chib, S. and Ramamurthy, S. (2014). DSGE Models with Student-t errors. *Econometric Reviews*, 33(1-4):152–171.

Chiu, C.-W. J., Mumtaz, H., and Pinter, G. (2017). Forecasting with VAR models: Fat tails and stochastic volatility. *International Journal of Forecasting*, 33(4):1124–1143.

Christiano, L. J. (2007). Comment [On the Fit of New Keynesian Models]. *Journal of Business & Economic Statistics*, 25(2):143–151.

Clark, T. E. (2011). Real-time density forecasts from Bayesian vector autoregressions with stochastic volatility. *Journal of Business & Economic Statistics*, 29(3):327–341.

Clark, T. E. and Ravazzolo, F. (2015). Macroeconomic forecasting performance under alternative specifications of time-varying volatility. *Journal of Applied Econometrics*, 30(4):551–575.

Cogley, T. and Sargent, T. J. (2005). Drifts and volatilities: Monetary policies and outcomes in the post WWII US. *Review of Economic Dynamics*, 8(2):262–302.

Cross, J. and Poon, A. (2016). Forecasting structural change and fat-tailed events in Australian macroeconomic variables. *Economic Modelling*, 58:34–51.

Cúrdia, V., Del Negro, M., and Greenwald, D. L. (2014). Rare shocks, great recessions. *Journal of Applied Econometrics*, 29(7):1031–1052.

Del Negro, M. and Primiceri, G. E. (2015). Time varying structural vector autoregressions and monetary policy: A corrigendum. *The Review of Economic Studies*, 82(4):1342–1345.

Delle Monache, D., De Polis, A., and Petrella, I. (2021). Modeling and forecasting macroeconomic downside risk. *Bank of Italy Temi di Discussione (Working Paper) No*, 1324.

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):134–144.

Fagiolo, G., Napoletano, M., and Roventini, A. (2008). Are output growth-rate distributions fat-tailed? some evidence from OECD countries. *Journal of Applied Econometrics*, 23(5):639–669.

Ferreira, J. T. and Steel, M. F. (2007). A new class of skewed multivariate distributions with applications to regression analysis. *Statistica Sinica*, pages 505–529.

Geweke, J. and Amisano, G. (2010). Comparing and evaluating Bayesian predictive distributions of asset returns. *International Journal of Forecasting*, 26(2):216–230.

Hörmann, W. and Leydold, J. (2014). Generating generalized inverse Gaussian random variates. *Statistics and Computing*, 24(4):547–557.

Karlsson, S. (2013). Forecasting with Bayesian vector autoregression. In Elliott, G. and Timmermann, A., editors, *Handbook of economic forecasting*, volume 2, pages 791–897. Elsevier.

Karlsson, S. and Mazur, S. (2020). Flexible fat-tailed vector autoregression. Working Papers 2020:5, Örebro University, School of Business.

Kastner, G. and Frühwirth-Schnatter, S. (2014). Ancillarity-sufficiency interweaving strategy (asis) for boosting mcmc estimation of stochastic volatility models. *Computational Statistics & Data Analysis*, 76:408–423.

Kim, S., Shephard, N., and Chib, S. (1998). Stochastic volatility: Likelihood inference and comparison with ARCH models. *The Review of Economic Studies*, 65(3):361–393.

Koop, G. and Korobilis, D. (2010). *Bayesian multivariate time series methods for empirical macroeconomics*. Now Publishers Inc.

Lanne, M., Meitz, M., and Saikkonen, P. (2017). Identification and estimation of non-gaussian structural vector autoregressions. *Journal of Econometrics*, 196(2):288–304.

Lewis, D. J. (2021). Identifying Shocks via Time-Varying Volatility. *The Review of Economic Studies*, 88(6):3086–3124.

Liu, X. (2019). On tail fatness of macroeconomic dynamics. *Journal of Macroeconomics*, 62:103154.

McCracken, M. W. and Ng, S. (2016). FRED-MD: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, 34(4):574–589.

McNeil, A. J., Frey, R., and Embrechts, P. (2015). *Quantitative risk management: Concepts, Techniques and Tools*. Princeton university press.

Montes-Galdón, C. and Ortega, E. (2022). Skewed SVARS: Tracking the structural sources of macroeconomic tail risks.

Nakajima, J. and Omori, Y. (2012). Stochastic volatility model with leverage and asymmetrically heavy-tailed error using gh skew student's t-distribution. *Computational Statistics & Data Analysis*, 56(11):3690–3704. 1st issue of the Annals of Computational and Financial Econometrics Sixth Special Issue on Computational Econometrics.

Ni, S. and Sun, D. (2005). Bayesian estimates for vector autoregressive models. *Journal of Business & Economic Statistics*, 23(1):105–117.

Panagiotelis, A. and Smith, M. (2008). Bayesian density forecasting of intraday electricity prices using multivariate skew-t distributions. *International Journal of Forecasting*, 24(4):710–727.

Primiceri, G. E. (2005). Time varying structural vector autoregressions and monetary policy. *The Review of Economic Studies*, 72(3):821–852.

Roberts, G. O. and Rosenthal, J. S. (2009). Examples of adaptive mcmc. *Journal of Computational and Graphical Statistics*, 18(2):349–367.

Sahu, S. K., Dey, D. K., and Branco, M. D. (2003). A new class of multivariate skew distributions with applications to Bayesian regression models. *Canadian Journal of Statistics*, 31(2):129–150.

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica*, 48(1):1–48.

Uhlig, H. (1997). Bayesian vector autoregressions with stochastic volatility. *Econometrica*, pages 59–73.