Dynamic copula with many exogenous variables meets the Horseshoe

Igor Martins *and Hedibert Freitas Lopes†

May 2023

Abstract

This paper proposes a novel framework that leverages the sparsity-inducing properties of the Horseshoe prior to allow for the inclusion of hundreds of exogenous variables into a dynamic copula model. We provide a Hamiltonian Monte Carlo scheme to simulate from the joint posterior of model parameters and states which is easily implementable and flexible to incorporate diverse types of prior believes. In an empirical application, we model the joint distribution of oil and SP500 returns by combining stochastic volatility models and dynamic copulas with exogenous variables. Our Horseshoe-dynamic Student's t copula approach leads to the highest marginal log-likelihood when compared to traditional static and dynamic copula models. Furthermore, we identify three variables likely to provide useful information for the dependence structure: inflation, unemployment rate and the real price commodity factor.

Key words: Dynamic Copula; Sparsity; Hamiltonian Monte Carlo.

1 Introduction

Appropriate models for dependence structure between series are vital in several fields. For example, in finance, the dependence structure of returns is crucial in assessing portfolio risk, as shown at least since Markowitz [1952], being typically modelled via multivariate normal distribution. Multivariate normality is often employed to being easily tractable it also has potentially several limitations. First, it does not allow for tail dependence which is critical when modelling the dependence stocks, currencies, credit products and across multiple asset classes as show in Ausin and Lopes [2010], Patton [2006], Daul et al. [2003] and Nguyen and Virbickaite [2022], respectively. Second, it does not allow for asymmetries which is present in several asset classes as show in Johnson [2002], Embrechts et al. [2001] and Smith [2007]. Third, multivariate normality for the conditional joint distribution of the returns implies that all marginals are the same and that the dependence structure must be linear which can be to restrictive for modeling the cross-sectional dependence as shown in Embrechts et al. [2002]. Other limitations of multivariate normal are discussed, for example, in Fabozzi et al. [2005].

Copulas are an alternative approach which circumvent the limitations of the multivariate normal discussed above while not necessarily ruling it out. The theoretical basis for copulas is Sklar's theorem which states that any multivariate distribution can be decomposed into its marginal distributions and corresponding copula function. Such decomposition enables a two-step modeling process based first on determining the marginals distribution, which can be different for each series, and then the specification of a well-suited copula to capture stylized facts of the dependence structure.

In addition to allowing for the possibility of tail-dependence and asymmetries, two other features are desired when modelling dependence: time-variation and information via additional variables. Time-variation on the cross-sectional structure is a common feature in subjects ranging from financial asset, survival rates on hemodialysis patients and housing prices as shown in Engle [2002], Kalantar-Zadeh et al. [2006] and Zimmer [2015], respectively. Also, while not necessarily affecting the marginal modelling, additional variables may be used as an additional source of information for the dependence structure. For example, Conrad et al. [2014] use five macroeconomic variables to enhance the performance of their copula model for the dependence between oil and SP500 returns. Hafner and Manner [2012] present a natural way to incorporate time-variation by allowing a transformation of the

*Insper. igorfbm@al.insper.edu.br
†Insper. hedibertfl@insper.edu.br

copula parameter to follow and AR process. Furthermore, we can incorporate additional informational by changing from an AR to ARX process.

This paper focus on this last feature: adding variables as a possible source of information for the evolution of the copula. While conceptually simple, the problem posses econometrics challenges. Even if the temporal evolution of the copula parameters was observable, for a large enough number of exogenous variables added, we would face a high-dimensional problem as the ones discussed in Hastie et al. [2015]. However, the evolution of the copula parameter is unobservable and must be recovered from a non-linear state-space at each point in time adding to the complexity of the problem.

Due to the increasing complexity when the number of covariates are high, the literature considers either directly or indirectly a small amount of predictors for the latent copula parameter incurring in the risk of potentially neglecting useful variables which may be harmful due to the well known textbook problems of misspecification. Maneejuk and Yamaka [2019] is an example of the former. They consider a single variable based on Google trends information to provide additional information for their copula model of contagion between asset markets. Nguyen and Javed [2021] and Nguyen and Virbickaite [2022] provide examples of the later. They start with several macroeconomic variables, divide them into a small number of groups, apply PCA to each of them, keep the first factor, and use it as an additional source of information to model the dependence between the SP500 and oil returns. While starting with multiple variables, this approach relies on a small number of groups being formed, discards portions of the variation in the data by keeping only the first factor of such groups and may eliminate several extreme points which can be key for determining the appropriate copula.

We contribute to the literature by providing a framework based on the Horseshoe prior,for the exogenous variables coupled with a Hamiltonian Monte Carlo scheme which allows for the latent copula parameter, the parameters that guide it's law of motion and the effect of the many exogenous variables to be recovered without having to impose ex-ante zero probability for several potential variables. Also, our framework does not need to discard potentially useful variation from additional variables leading to a more flexible framework.

We apply our proposed model to characterize the dependence structure between West Texas Intermediate (WTI) and SP500 returns. We follow a two-step approach by first modelling the marginals and then moving to the copula function. We consider stochastic volatility (SV) models for the marginals and focus on three main characteristics for the copula function. First, it should allow for possible tail dependency and asymmetries. For example, the Student's t copula allows for upper and lower tail dependence of the same magnitude, while the Clayton copula allow for dependence only on the lower tail. Second, it should allow for the possibility of the dependence structure to vary with time. For example, Zhu et al. [2014] indicates changes in the magnitude of the dependence structure in economic expansions and recessions. Patton [2006], show that time-variation is key when specifying a copula for the standardized residuals of foreign exchange returns. Third, it may allow for exogenous variables as possible sources of information about the copula such as in Maneejuk and Yamaka [2019], Nguyen and Virbickaite [2022] and Gong et al. [2022b]. In our empirical application , our Horseshoe-based method coupled with a Student's t Copula leads to the highest marginal likelihood. Additionally, from the 134 variables, we obtain three which are likely to provide useful information about the dependence structure: unemployment rate, inflation and real price commodity factor.

This paper starts by describing our sampling model and priors on Section 2. Section 3 presents the Hamiltonian Monte Carlo algorithm and describes why it useful for our approach. Section 4 present the WTI and SP500 data, the results for both the marginal and copula modelling and finishes with an analysis on the main exogenous variables capable of providing information about the dependence structure. Section 5 concludes.

2 Model and priors

This section describes the proposed model. We start by defining copula and stating Sklar's theorem which provides the main theoretical result to allow our modelling approach to be disentangle in two steps: marginal modelling and copula function. We present stochastic volatility models for the returns, then we show the baseline static and dynamic copula models. Then, we introduce exogenous variables to the evolution of a transformation of the latent copula parameter. Finally, we conclude with our sparsity inducing approach.

2.1 Copula and Sklar's theorem

Following Czado [2019], a d-dimensional copula C is a multivariate distribution function on the d-dimensional hypercube $[0,1]^d$ with uniformly distributed marginals while the copula density, denoted by c, can be obtained via

$$c(u_1,\ldots,u_d) = \frac{\partial^d}{\partial u_1 \ldots \partial u_d} C(u_1,\ldots,u_d) \forall u \in [0,1]^d$$

The main theoretical results for copulas is due to Sklar [1959] which states that if Y is a d-dimensional random vector with joint distribution function F and marginal distributions F_i for $i = 1, \ldots, d$, then the joint distribution can be written as $F(y_1, \ldots, y_d) = C(F_1(y_1), \ldots, F_d(y_d))$ with associated density $f(y_1, \ldots, y_d) = c(F_1(y_1), \ldots, F_d(y_d))f_1(y_1) \ldots f_d(y_d)$. Also, as shown in Czado [2019], the copula corresponding to a multivariate distribution function F with marginal distribution functions F_i is

$$C(u_1, \ldots, u_d) = F(F_1^{-1}(u_1), \ldots, F_d^{-1}(u_d))$$

or equivalently $c(u_1, \ldots, u_d) = \frac{f(F_1^{-1}(u_1) \ldots F_d^{-1}(u_d))}{f_1(F_1^{-1}(u_1)) \ldots F_d^{-1}(u_d))}$. Additionally, for absolutely continuous distributions, the copula C is unique.

The main point of Sklar's theorem is that it allows to combine arbitrary marginal distributions, $f_i(y_i)$, with copula densities,$c(F_1(y_1), \ldots, F_d(y_d))$, to build new multivariate densities $f(y_1, \ldots, y_d)$. This theorem enables the two-step approach based on first modeling the marginals $f_i(y_i)$ and then the dependence structure via copula function $c(F_1(y_1), \ldots, F_d(y_d))$. In this paper, we focus on the case $d = 2$. However our approach remain valid for higher dimensions. We start by presenting our SV for the marginals and then we move to the static and dynamic copula functions.

2.2 Stochastic volatility models

Let $P_{i,t}$ be the price of a financial asset i with associated log returns as $r_{i,t} = logP_{i,t} - logP_{i,t-1}$. We model the dynamics of the individual asset log returns via stochastic volatility model represented by Equations (1) to (4).

$$r_{i,t} = \mu_r + e^{h_t/2}\varepsilon_t \tag{1}$$

$$h_t = \mu + \phi(h_{t-1} - \mu) + \sigma\eta_t \tag{2}$$

$$\varepsilon_t \sim N(0, 1) \tag{3}$$

$$\eta_t \sim N(0, 1) \tag{4}$$

As discussed in Hafner and Manner [2012], several other models for the marginals are reasonable. For example, Nguyen and Javed [2021] consider GJR-GARCH model with skew-t errors while Hafner and Manner [2012] consider both GARCH and SV models. Due to the good performance show by SV models for oil and the SP500 , exemplified on Baumeister et al. [2022] and Johannes et al. [2009], respectively, we opt for SV marginal modelling in this paper.

Since $\varepsilon_t \sim N(0, 1)$ After estimation, we use $\hat{\varepsilon}_t$ to obtain uniform marginals via inverse pdf which can then be used as inputs for the copulas. Once specified a marginal model, we go on to the copula modelling step.

2.3 Dynamic stochastic copulas with exogenous variables

There are several possible choices of copula functions. In this paper, we consider 4 classes of copulas: Gaussian, Student's t, Clayton and Gumbel. Figure 1 illustrates how the behaviour of the dependence structure for each class of copula with marginals being standard uniforms. For instance, this set of 4 classes provide a different measure of tail dependence replicated from Czado [2019] into Table 2.3. They capture no dependence via Gaussian, same dependence on upper and lower tail via Student's t, and tail dependence only on the lower for the Clayton and upper for the Gumbel cases. Nelsen [2007] provide an analysis of the most common copulas.

Copula class	Upper tail dependence	Lower tail dependence
Gaussian	-	-
Student's t	$2t_{\nu+1}\Big(-\sqrt{\nu+1}\sqrt{\frac{1-\rho}{1+\rho}}\Big)$	$2t_{\nu+1}\Big(-\sqrt{\nu+1}\sqrt{\frac{1-\rho}{1+\rho}}\Big)$
Clayton	-	$2^{-1/\delta}$
Gumbel	$2 - 2^{1/\delta}$	-

Table 1: Tail dependence for the bivariate copula classes in terms of their parameters. Following Czado [2019], the upper and lower tail dependence coefficient are defined as $\lambda^{upper} = \lim_{t\to 1^-} \frac{1-2t+C(t,t)}{1-t}$ and $\lambda_{lower} = \lim_{t\to 0^+} \frac{C(t,t)}{t}$. - indicates no tail dependence. $t_{\nu+1}$ is the univariate Student's t distribution function with $\nu+1$ degrees of freedom. ρ is the Student's t correlation parameter. δ is the parameter for the Clayton and Gumbel copulas.

Traditionally, Gaussian, Stundet's t, Clayton and Gumbel are modelled with static parameters. However, time-variation is key when modeling the cross-sectional dependence for the residuals of SV and GARCH models

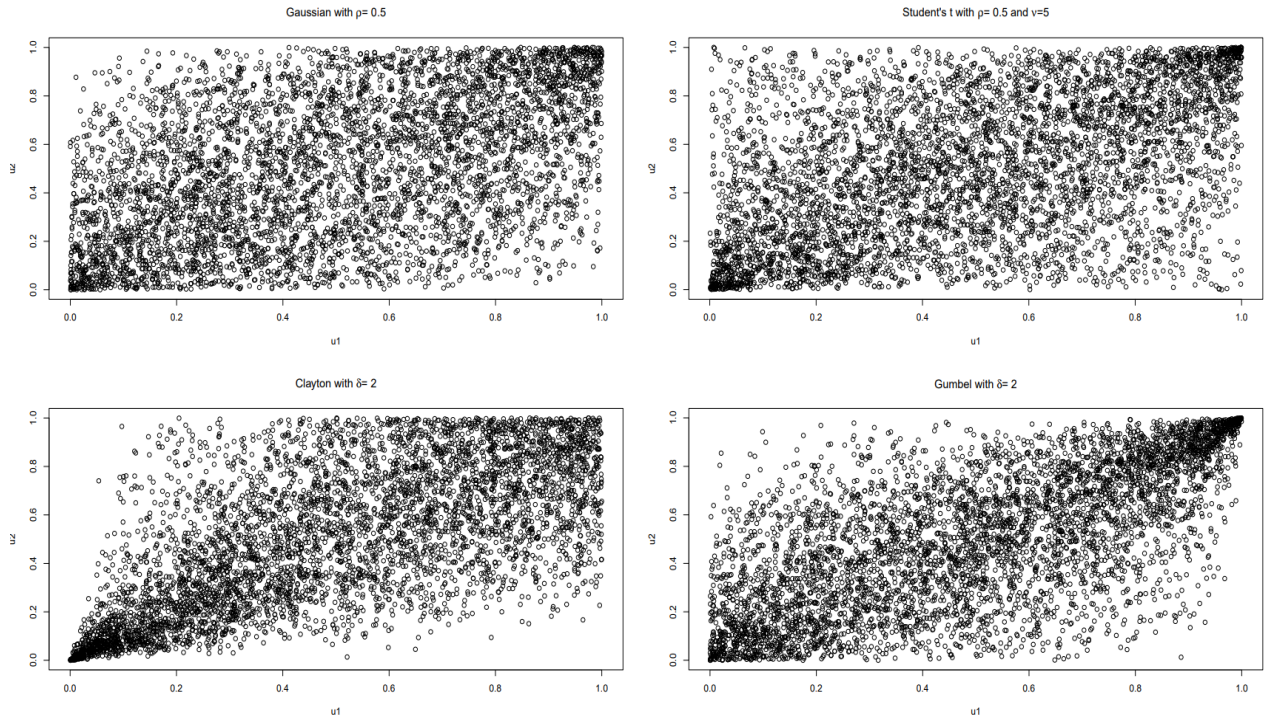

Figure 1: Behaviour of the 4 classes of copulas considered in this paper. While the Gaussian copula (top left) have no tail dependence, the Student's t (top right) have the same dependence on the upper and lower case. Clayton (bottom left) and Gumbel (bottom right) copula have tail dependence only on the lower and upper tail, respectively. The magnitude for each case is given by Table 2.3.

as shown in Patton [2006]. Hafner and Manner [2012] proposed a model in which the dynamics of the copula parameter is obtained via a mean reverting AR1 for a transformation of the parameter as expressed by Equations (5) to (7). Equation (5) represents thet joint distribution of standard uniforms obtained via the marginal modelling are distributed according to a copula c with parameter vector Θ. Equation (6) transform the copula parameters via function $\Lambda(\cdot)$ which is a copula-specific transformation that ensures that Θ is in the appropriate domain. For example, $\Lambda(\cdot) = \frac{exp(2\lambda)-1}{exp(2\lambda)+1}$ for the Gaussian case ensuring $-1 \leq \Theta \leq 1$. A list of the transformations $\Lambda(\cdot)$ for each copula is available at the Appendix A. Finally, Equation (7) models the transformed parameter into a stationary AR1.

$$(u_{1,t}, u_{2,t}) \sim c(u_{1,t}, u_{2,t}; \Theta) \tag{5}$$

$$\Theta = \Lambda(\lambda_t) \tag{6}$$

$$\lambda_t = \alpha(1-\phi_\lambda) + \phi_\lambda \lambda_{t-1} + \sigma_\lambda \varepsilon_t \text{ with } \varepsilon_t \sim N(0,1) \tag{7}$$

Hafner and Manner [2012] is not the only approach to modeling dynamic copulas. For example, Patton [2006] consider a observation driven approach in which a transformation of the copula parameter is modelled based on a GARCH-type dynamics. However, Hafner and Manner's approach performs at least as well as Patton's. Also, the smoothness introduced via the stochastic component in λ_t is useful for our application when compared to the predetermined aspect of Patton [2006] since parameter smoothness is associated with smoother risk estimates leading to smaller portfolio rebalancing costs.

In this paper we add exogenous variables to Equation (7) changing it from an AR1 to an AR1X as presented in Equation (8) where X represents a vector of exogenous variables with associated coefficient β. Therefore, the dynamic copula model considered in this paper is characterized by Equations (5), (6), (7) and (8) which forms a non-linear state-space with, potentially, many lagged variables X_{t-1} which may be informative about λ_t.

$$\alpha = \alpha_0 + X_{t-1}\beta \tag{8}$$

Therefore, after estimating the marginal model, we must recover $\{\alpha_0, \phi_\lambda, \sigma_\lambda, \{\lambda_t\}_1^T, \beta\}|\cdot$. We choose the following priors: α_0 has a normal prior with 0 mean and large variance, ϕ_λ follows a normal distribution truncated between -1 and 1 to ensure stationarity and σ_λ^2 follows and inverse gamma.

As discussed before, even if $\{\lambda_t\}$ was observable, for a large enough number of X variables added, we would face a high-dimensional problem as the ones discussed in Hastie et al. [2015]. However, the evolution of the copula parameter is unobservable and must be recovered from a non-linear state-space at each point in time adding to the complexity of the problem. To deal with the high-dimensional problem, we rely on a sparsity-inducing prior for β. The prior choice for β is describe on subsection . Subsection 3 describes our approach to simulate from the posterior.

2.4 Sparsity-inducing prior for β and remaining priors

In the non-Bayesian literature, sparse problems are typically handled by maximizing the likelihood subject to a penalty function with common approaches being the the LASSO of Tibshirani [1996] or one of its generalizations such as the elastic net discussed in Zou and Hastie [2005]. From a Bayesian point of view, sparsity problems can be represented as a penalization to the log-likelihood via log-prior. For example, Park and Casella [2008] represents the LASSO as a Gaussian linear model penalized via double exponential prior.

Two characteristics are vital when considering sparsity inducing priors. First, it must have a high peak around zero to provide shrinkage towards zero to small elements. Second, it must have heavy tails to not shrink away large enough signals. The horseshoe prior of Carvalho et al. [2009] and Carvalho et al. [2010], represented by Equations (9) and (10), is a popular choice since it satisfies both desired characteristics. Its Cauchy-like tails allow strong signals to remain large a posteriori while its infinitely tall spike at the origin provides severe shrinkage for the zero elements of β. The horseshoe is considered a global-local shrinkage prior. The hyperparameter τ shrinks all parameters towards zero while the half-Cauchy priors for the local hyperparameters λ allow some β_j to remain non-zero.

$$\beta_j|\lambda_j, \tau \sim N(0, \tau^2\lambda_j^2) \tag{9}$$

$$\lambda_j \sim C^+(0,1) \text{ for } j = 1, \ldots, p \tag{10}$$

More formally, as shown in Carvalho et al. [2010] and Piironen and Vehtari [2017], if we consider a linear regression problem as in Equation (11) and consider horseshoe priors for the coefficients, then the conditional posterior for β given the hyperparameters, X and y can be written as

$$y_i = \beta^T x_i + \varepsilon_i \text{ with } \varepsilon_i \sim N(0, \sigma^2) \text{ and } i = 1, \ldots, n \tag{11}$$

$$p(\beta|\lambda, \tau, \sigma^2, X, y) = N(\beta|\bar{\beta}, \Sigma)$$

where

$$\bar{\beta} = \tau^2\Lambda(\tau^2\Lambda + \sigma^2(X^TX)^{-1})^{-1}\beta_{MLE}$$

$$\Sigma = \left(\tau^{-2}\Lambda^{-1} + \frac{1}{\sigma^2}X^TX\right)^{-1}$$

where $\Lambda = diag(\lambda_1^2, \ldots, \lambda_p^2)$ and $\beta_{MLE} = (X^TX)^{-1}X^Ty$ is the MLE solution. If the predictors are uncorrelated, have mean zero and $Var(x_j) = s_j^2$, then $X^TX \approx ndiag(s_1^2, \ldots, s_p^2)$ and we have $\bar{\beta}_j \approx (1 - \kappa_j)\beta_{MLE}$ in which $\kappa_j = \frac{1}{1+n\sigma^{-2}\tau^2 s_j^2\lambda_j^2}$ is known as the shrinkage factor. As shown in Piironen and Vehtari [2017], $\bar{\beta} \to 0$ when $\tau \to 0$ and $\bar{\beta} \to \beta_{MLE}$ as $\tau \to \infty$.

Additionally, Carvalho et al. [2010] and Piironen and Vehtari [2017] show that $\kappa_j = \frac{1}{1+n\sigma^{-2}\tau^2 s_j^2\lambda_j^2}$ holds for any prior that can be written as a scale mixture of gaussians. However, when λ_j has half-Cauchy priors for all λ_j and for fixed τ and σ, the shrinkage factor follows Equation (12)

$$p(\kappa_j|\tau, \sigma) = \frac{1}{\pi}\frac{a_j}{(a_j^2 - 1)\kappa_j + 1}\frac{1}{\sqrt{\kappa_j}\sqrt{1 - \kappa_j}} \tag{12}$$

where $a_j = \tau\sigma^{-1}\sqrt{n}s_j$ which for the special case in which $a_j = 1$, reduces to a Beta(1/2, 1/2) which looks like a Horseshoe, as show in Figure (2), yielding the prior its name. Due to the good properties of the horseshoe prior discussed above, we use the horseshoe prior to induce sparsity on the coefficients of the exogenous variables on the ARX shown in Equation (7).

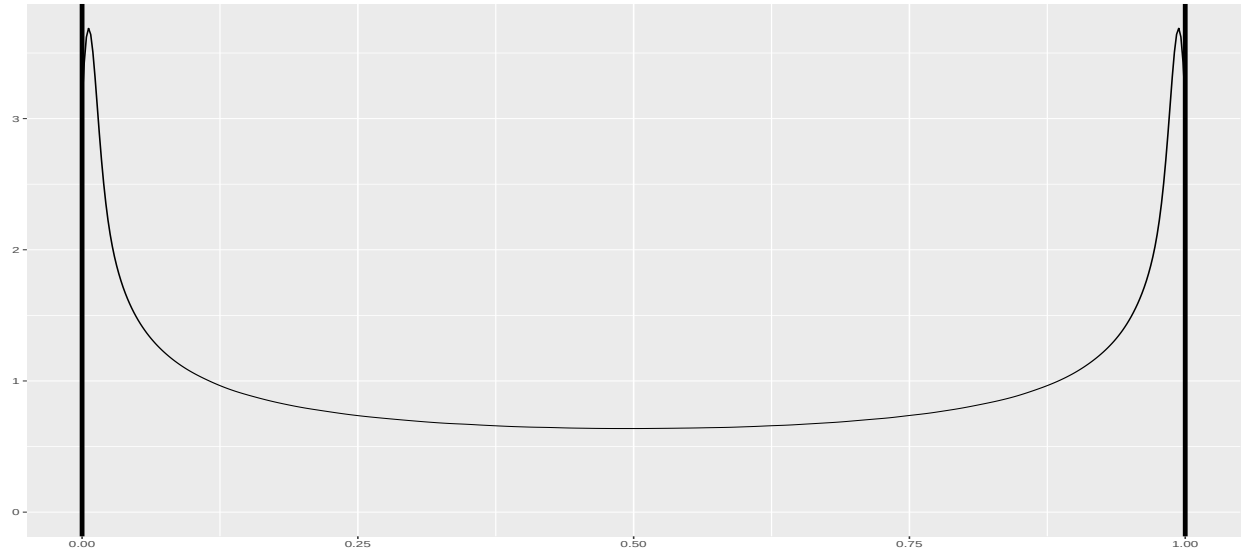

Figure 2: Beta(1/2, 1/2) representing the shrinkage factor, show in Equation (12), for the special case in which $a_j = 1$. Due to the shape of the shrinkage factor, the prior induced by Equations (9) and (10) is named the Horseshoe prior. Vertical bars on 0 and 1 are plotted for reference.

3 Hamiltonian Monte Carlo

Markov Chain Monte Carlo (MCMC) algorithms may be used to simulate samples for approximation of the posterior distribution, presented in Equation (13), where D is the observed data, $f(\theta)$ is the prior for the parameters θ and $f(D|\theta)$ represents the sampling model. For brevity, I will write $f(\theta)$ instead of $f(\theta|D)$ from now on.

$$f(\theta|D) = \frac{f(D|\theta)f(\theta)}{\int f(D|\theta)f(\theta)d\theta} \propto f(D|\theta)f(\theta) \tag{13}$$

This paper uses a particular MCMC method known as Hamiltonian Monte Carlo (HMC) instead of the more traditional Metropolis-Hastings (MH) algorithm. As discussed in Gamerman and Lopes [2006], the MH algorithm generates a sequence of values θ that form a Markov chain. Each $\theta^{(i)}$, where i represents the iteration index going from 0 to a desired number of iterations, is defined in part by a proposal density $q(\theta^{prop}|\theta^{i-1})$ where θ^{prop} is a proposal for the next value in the chain and θ^{i-1} represents the value of θ on the previous iteration. Random walk proposals are one of the most popular choices in the literature receiving the name of Random-Walk Metropolis Hastings. Each proposed value for θ^i may be accepted or rejected based on Equation (14) where acc represents the acceptance ratio.

$$acc = min\left(1, \frac{f(\theta^{prop})q(\theta^{t-1}|\theta^{prop})}{f(\theta^{t-1})q(\theta^{prop}|\theta^{t-1})}\right) \tag{14}$$

As discussed in Betancourt [2017], the limitations of the MH algortihm are mainly computational due to the acceptance ratio being usually low. For example, Gelman et al. [1997] shows well-tuned MH algorithms still accept less than 25% of the proposals often. Additionally, Thomas and Tu [2021] indicates lower acceptance ratios are even more pronounced in high-dimensional problems.

It is in this context that HMC emerges as a preferred alternative. As show in Thomas and Tu [2021], Betancourt [2017] and Gelman et al. [2015], HMC improves the efficiency of MH by employing guided proposals based on the gradient of the log posterior to direct the Markov chain towards regions of higher posterior density while also sampling the tail areas properly. As a result, HMC chain will accept proposals at a much higher rate than the traditional MH algorithm.

As discussed in Duane et al. [1987] and Thomas and Tu [2021], HMC approach is analogous to the movement of a hypothetical object on a frictionless curve, where the object passes and stay near the bottom of the valley while occasionally visiting the higher grounds on both sides. In classical mechanics, such movements are described by the Hamiltonian function $H(\theta, p)$, where the exchanges of kinetic, K(p), and potential energy, $U(\theta)$, dictate the

object's location at any given moment by $H(\theta, p) = U(\theta) + K(p)$, where p and $\theta \in R^K$. This paper is interested in generating θ from a given distribution $f(\theta)$ where $U(\theta) = -logf(\theta)$ and, typically, $p \sim N_k(0, M)$ leading to $H(\theta, p) = -logf(\theta) + \frac{1}{2}p^T M^{-1} p$.

Over time, HMC travels on trajectories that are governed by the Hamiltonian equations:

$$\frac{dp}{dt} = -\frac{\partial H(\theta, p)}{\partial \theta} = -\frac{\partial U(\theta)}{\partial \theta} = \nabla_\theta logf(\theta)$$

$$\frac{d\theta}{dt} = \frac{\partial H(\theta, p)}{\partial p} = \frac{\partial K(p)}{\partial p} = M^{-1}p$$

where $\nabla_\theta logf(\theta)$ is the gradient of the log posterior density.

A solution to the Hamiltonian equations is a function that defines the path of (θ, p) from which specific values of θ could be sampled. As show in Betancourt [2017], for a given momentum vector within an HMC iteration, the path defined by the Hamiltonian equations is deterministic and proposals generated from an exact solution would be accepted with probability 1. However, usually, the system has no analytical solution and relies on numerical methods. The most popular numerical method to solve Hamiltonian equations is the leapfrog method. As show in Thomas and Tu [2021], it modifies Euler's method by using a discrete step size ε individually for p and θ, with a full step ε in θ sandwiched between two half-steps $\varepsilon/2$ for p,

$$p(t + \varepsilon/2) = p(t) + (\varepsilon/2)\nabla_\theta logf(\theta(t))$$

$$\theta(t + \varepsilon) = \theta(t) + \varepsilon M^{-1} p(t + \varepsilon/2)$$

$$p(t + \varepsilon) = p(t + \varepsilon/2) + (\varepsilon/2)\nabla_\theta logf(\theta(t + \varepsilon))$$

However, due to the leapfrog method being an approximation, an acceptance step is added to ensure that proposals does not deviate far from the specified Hamiltonian $H(\theta, p)$. While the acceptance rate of HMC proposals is less than 100%, it is usually higher than that of the MH making it specially suitable for the challenges of this paper. Due to its improvements over the traditional MH and flexibility, we opt for a HMC approach for this problem.

4 Empirical application

4.1 Data

This paper models the joint distribution of daily returns for the WTI and SP500 from February 2, 2005 to December 30, 2022 a sample with 4511 observations displayed on Figure (3). The stock-oil relationship its of key interest since WTI is usually considered to be a hedge to market risk as discussed in Batten et al. [2021] and Sadorsky [1999]. Additionally, both oil and the stock market have been shown to be useful to evaluate future economic conditions such as in Hamilton [1983], Barsky and Kilian [2002] and Billio et al. [2013]. Thus, insights on their relationship can give insights about the future of the US and global economy.

Macroeconomic variables may provide information about the dependence structure between oil and the stock market. For instance, Conrad et al. [2014] show that industrial production, non farm payroll, unemployment rate, National activity index, and the leading index from the Federal Reserve Bank of Philadelphia provide additional information to the dependence structure when compared to a simple AR1 structure. Benefiting of our sparsity inducing approach, we consider 125 macroeconomic variables made available by the Federal Reserve Economic Data (FRED) and discussed at length in McCracken and Ng [2016] which classifies them into eight broad groups. Group 1, Output and Income, includes information on Real Personal Income as well as the Industrial production idex and its components. Group 2 provides Labor market series such as unemployment rate and initial jobless claims. Group 3 focuses on Housing. Group 4 informs on Consumption and inventories. Group 5 brings information about money stocks and credit information. Group 6 provides information on interest rate variables such as FED funds rate as well as exchange rate date. Group 7 represents inflation via producer price index (PPI), consumer price index (CPI) and its components. The final group reports on stock market variables such as VIX and the SP500 dividend yield. Note that among our 115 macroeconomic variable, we include the five series discussed in Conrad et al. [2014]. We expand the dataset with the global economic condition (GECON) index by Baumeister et al.

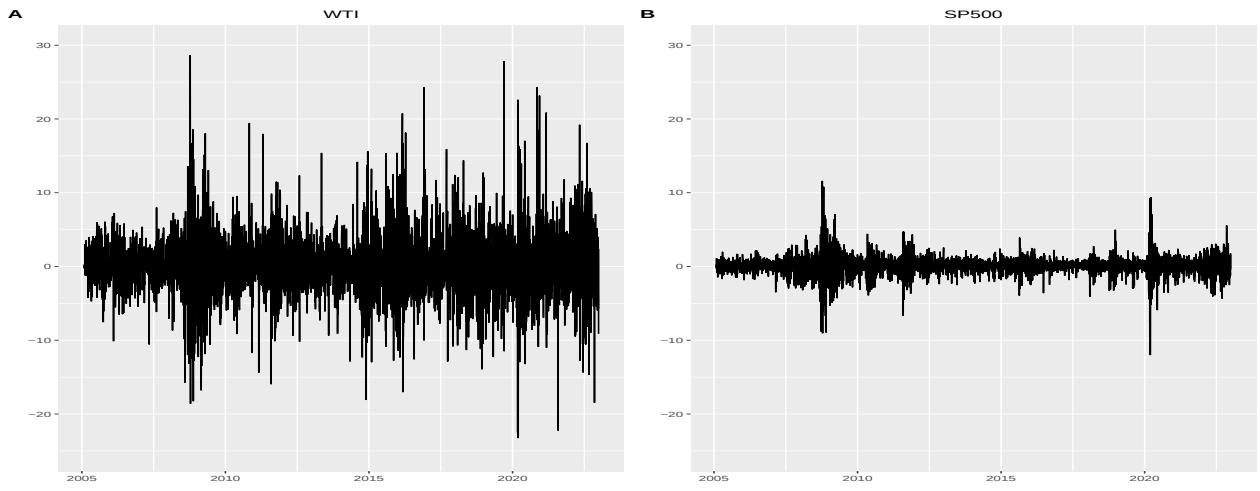

Figure 3: Daily returns for the WTI and SP500 from February 2, 2005 to December 30, 2022.

[2022]. The GECON index is based on 16 variables that cover multiple dimensions of the global economy and shows that is vital for forecasting the real price of Brent and global petroleum consumption.

Not only macroeconomic variables but Gong et al. [2022a] also shows that demand and supply oil variables have affect the relationship between the oil and stock market. They consider changes in global oil production in thousands of barrels per day, made available by the Energy Information Administration, as a supply factor and Kilian [2009] global index of dry cargo single voyage freight rates as a demand factor. In this paper, we not only consider both factor presented in Gong et al. [2022a], but add the market-based demand and supply shocks considered in Baumeister and Hamilton [2019] as well as the real commodity price factor in Baumeister and Guérin [2021] which aims to capture the demand-driven global fluctuations that make a large cross-section of real commodity prices to comove. [1].

Therefore, our dataset consist of WTI and SP500 returns, 125 macroeconomic variables from FRED, the GECON index, changes in global oil production, Killian's Cargo Index, four shocks described in Baumeister and Hamilton [2019] and the real commodity price factor of Baumeister and Guérin [2021]. As far as we known, no other paper have directly considered such a large amount of variables in a dynamic copula with forecasting variables. Appendix B list all time series considered in this paper. We use the sample from February 2, 2015 up to December 31, 2019 for estimation and the remaining observations for future out-of-sample analysis.

Due to macroeconomic variables listed on Appendix B being on monthly frequency, we employ a weighting scheme based on Almon polynomials discussed in Ghysels et al. [2016] in which the monthly series have their impact diluted throughout the month with around 80% of the weight on the first 2/3 of the month and them a fast decay. Other possibilities are available in the literature such as the restricted beta function employed by Nguyen and Virbickaite [2022].

4.2 Marginals

We consider a stochastic volatility model for both WTI and SP500 returns. Table (4.2) shows the parameters for both assets. For WIT, μ indicates an annualized level of $e^{2.33/2}\sqrt{252} = 50.9\%$ while for the SP500 it indicates $e^{-0.48/2}\sqrt{252} = 12.5\%$. Additionally, both series present persistent volatility represented by ϕ with posterior averages of 0.96 and 0.97 for WTI and SP500, respectively. Figures (4) and (5) plots the recovered time-varying volatility highlighting the pattern of clustered high and low volatility for both time series. For example, the volatility of both assets spiked around the 2008 financial crisis while remaining on a lower level on 2005-06.

After estimation, we recover the residual $\hat{\varepsilon}_t$ which should be normal based on our SV modelling. In order to consider the feasibility of this hypothesis, we show QQ plots in Figure (6). While there are some departures from from theoretical quantile values such deviations are not large enough to claim not gaussianity. For example, when performing a Jarque - Bera test, we don't reject the null hypothesis of gaussianity at the 1% significance level for neither series.

[1]The time series for each shock, the GECON index and the real commmodity price factor are available at Bauemeister's website: https://sites.google.com/site/cjsbaumeister/datasets

	μ_r	μ	ϕ	σ
WTI mean	0.05	2.34	0.96	0.25
WTI q05	-0.05	2.15	0.94	0.21
WTI q95	0.11	2.51	0.97	0.30
SP500 mean	0.05	-0.48	0.97	0.25
SP500 q05	0.03	-0.75	0.96	0.22
SP500 q95	0.06	-0.21	0.98	0.29

Table 2: Posterior summaries from the SV model described by Equations (1) to (4). μ_r, μ, ϕ, σ represent the average asset return, log volatility level, persistence and standard deviation, respectively. We simulate from the posterior of the model via R's stochvol package described in Hosszejni and Kastner [2019].

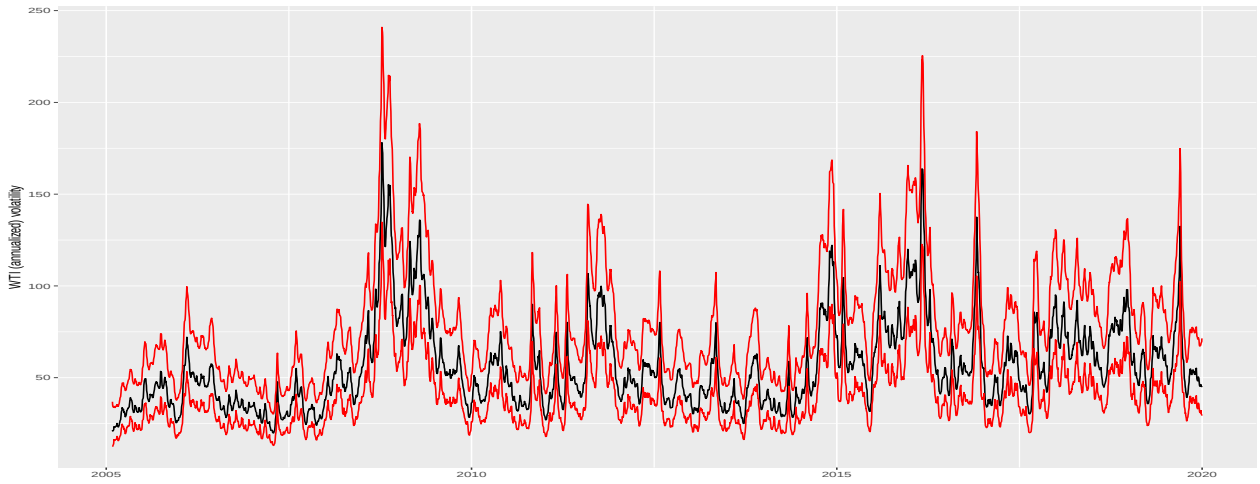

Figure 4: (Annualized) Daily WTI volatility recovered via exp(h/2) from February 2, 2015 to December 31, 2019. Black lines indicates posterior means and red lines form a 90% credible interval by highlighting the 5th and 95th quantiles.

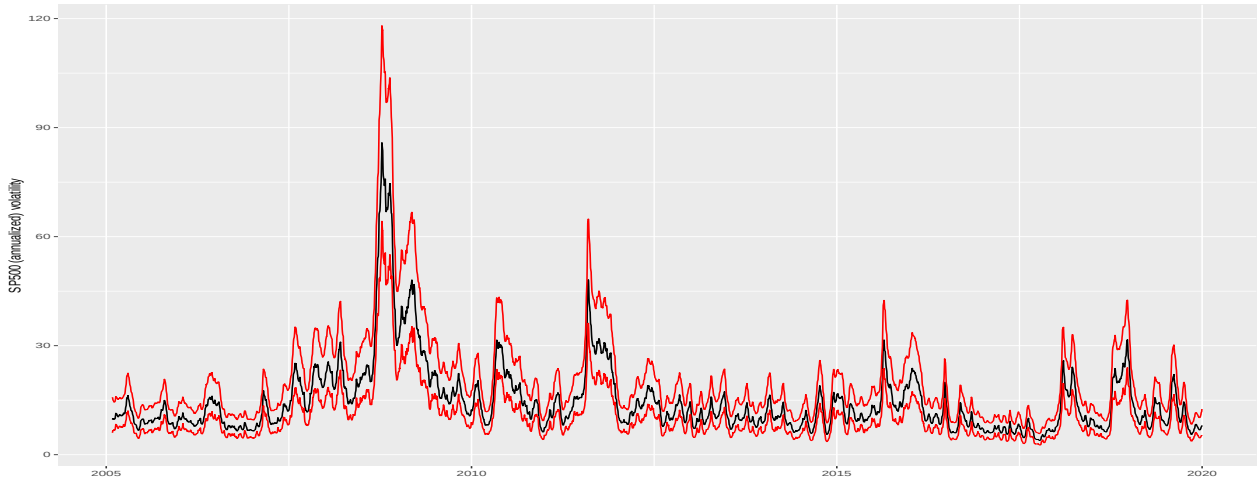

Figure 5: (Annualized) Daily SP500 volatility recovered via exp(h/2) from February 2, 2015 to December 31, 2019. Black lines indicates posterior means and red lines form a 90% credible interval by highlighting the 5th and 95th quantiles.

4.3 Copula

After we recover the parameters for the SV used to model the marginals, we model the cross-section dependence via a copula. We consider 4 types of copulas on both static and dynamic setups: Gaussian, Student's t, Clayton

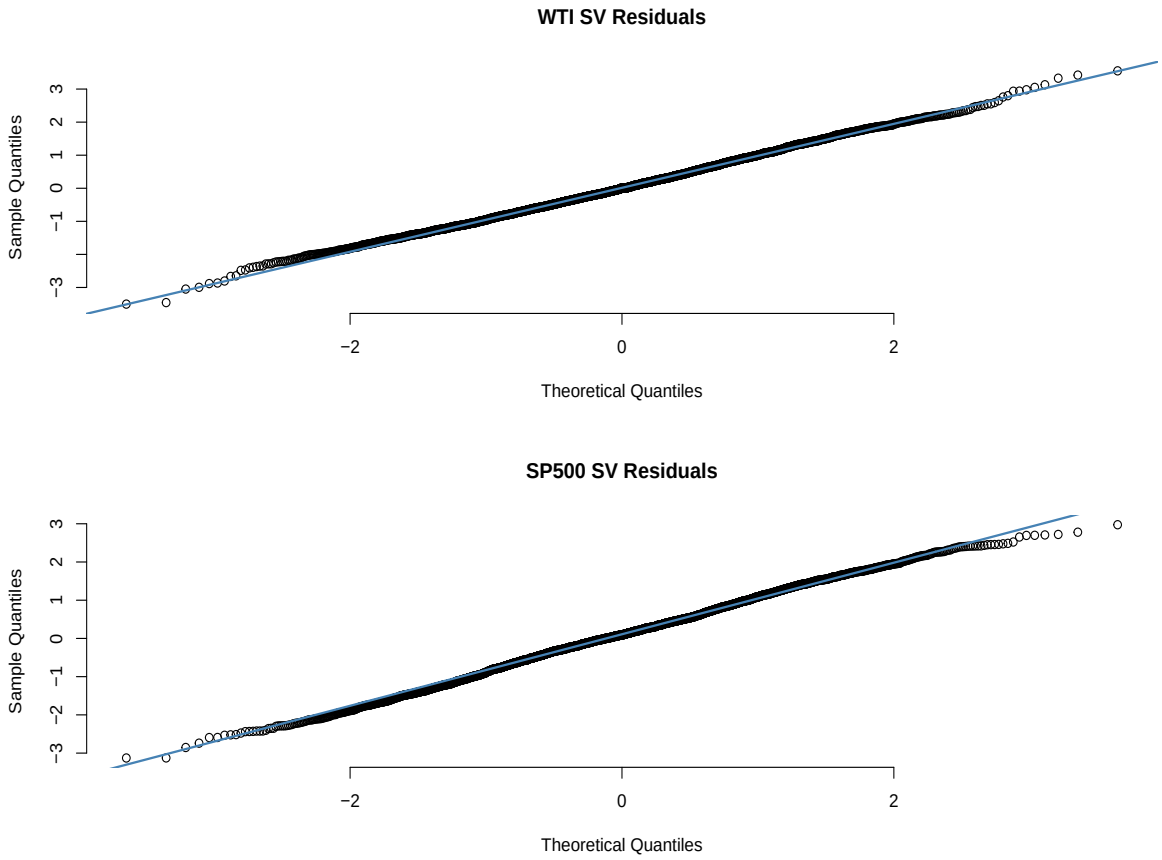

Figure 6: QQ plot for the SV residuals for both WTI and SP500, $\hat{\varepsilon}_{wti,t}$ and $\hat{\varepsilon}_{sp500,t}$, respectively. The residuals are obtained by considering the posterior mean for μ_r and $\{h\}$ via$\hat{\varepsilon}_t = (r_{i,t} - \mu_r)e^{-h_t/2}$.

and Gumbel. As discussed in subsection (2.1), this set of copulas allow for time-variation, tail dependence and asymmetries on the dependence structure.

Table (4.3) reports the log-marginal likelihood (LML) for the static and dynamic versions of the copulas without pepredictor variables. As described in Robert et al. [2007], the marginal likelihood of a model is the average of the likelihood of the data across all possible parameter values given conditional on the model weighted by the prior plausibility of those parameter values. It is the main ingredient to compute, for example, the Bayes factor and posterior model probabilities providing information about the plausibility of the model.

Our results indicates very strong evidence in favor of Student's t copulas when compared to other copula classes due to the the differences in LML. Therefore, the existence of dependence on both the upper and lower tails is a key feature of our sample. Additionally, the posterior mean for the degrees of freedom for the dynamic case with additional variables is 8.6 which also indicate a strong tail dependence. Finally, we identify time variation as being key for modelling the cross-sectional dependence of the SV residuals for the WTI and SP500 returns.

From the hundreds of variables considered in X_t, we identify only three variables with β not including zero on the 95% credible interval: unemployment rate, CPI and Baumeister's real price commodity factor with their respective β presented on Table 4.3. Conrad et al. [2014] also finds unemployment rate as a relevant variable for the dependence structure of Oil and SP500 returns using a DCC-GARCH-MIDAS approach also obtaining and increase in their correlation when unemployment rate rises. Our estimates indicate a diminishing dependence when inflation, as measured by the CPI, rises. According to Chen et al. [1986] the negative relation between stock returns and inflation are caused by the money demand and counter-cyclical money supply effects while Gorton and Rouwenhorst [2006] show that rises in inflation increase the expected return of oil. Therefore, by connecting both results, inflation would diminish the correlation between SP500 and WTI as in our result. Finally, the real price commodity factor capture the demand-driven global fluctuations responsible for large comovements in a cross-section of 23 real commodity prices. Since commodity prices movements are mechanically reflected on the CPI,

	LML
Gaussian	400.52
Student's t	452.80
Clayton	365.87
Gumbel	347.60
Dynamic Gaussian	586.71
Dynamic Student's t	592.32
Dynamic Clayton	427.72
Dynamic Gumbel	412.86
Dynamic Student's t with forecasting variables	595.47

Table 3: Log-marginal likelihood (LML) obtained for each model. Higher values of both LML indicates evidence in favor of the model. The LML results presents more evidence in favor of the Student's t copulas when compared to other classes which indicates tail dependence in both tails. Additionally, copulas allowing for time variation dominate the static counterparts. Finally, LML indicates our proposal as the more reasonable model when compared to others when assuming equal prior model probabilities.

the previous explanation also reflects the sign of the real price commodity factor on the dependece structure and moreover explains why its magnitude is close to the one on the CPI.

	CPI	UNRATE	RPCF
β_{mean}	-1.59	1.14	-2.27
β_{q05}	-3.31	0.02	-3.86
β_{q95}	-0.03	2.23	-0.79

Table 4: Posterior mean, 5th and 95th quantile from the posterior distribution of β when considering a Horseshoe prior for the coefficients of the variables used to provide additional information about the cross-sectional dependence across WTI and SP500 returns. CPI, UNRATE and RPCF refers to the consumer price inflation index, unemployment rate and the real commodity price factor, respectively.

5 Conclusion

Due to the high-dimensional nature of estimating dynamic copula models expanded with forecasting variables, most papers rely on a small number of potential sources of information for the cross-sectional dependence across time-series. By neglecting potentially useful information, the model may omit links between series and risks incurring into textbook problems of misspecification. This paper proposes a Horsehoe-based approach to perform variable selection on a dynamic stochastic copula model with hundreds of exogenous variables coupled with a Hamiltonian Monte Carlo to sample from the parameters and states joint posterior. In an empirical application, we combine stochastic volatility models with a dynamic copula with predictor variables to model the joint distribution of oil and SP500 returns. Our proposal leads to the highest marginal log-likelihood when compared to traditional static and dynamic copula models. We identify two main features for the dependence structure: time-variation and dependence structure on both the upper and lower tail. Additionally, we identify three main variables which are more likely to be a useful source of information for the dependence structure: Inflation, unemployment rate and real price commodity factor.

References

M. C. Ausin and H. F. Lopes. Time-varying joint distribution through copulas. *Computational Statistics & Data Analysis*, 54(11):2383–2399, 2010.

R. B. Barsky and L. Kilian. Oil and the macroeconomy since the 1970s. *Journal of Economic Perspectives*, 18(4): 115–134, 2002.

J. A. Batten, H. Kinateder, P. G. Szilagyi, and N. F. Wagner. Hedging stocks with oil. *Energy Economics*, 93: 104422, 2021.

C. Baumeister and P. Guérin. A comparison of monthly global indicators for forecasting growth. *International Journal of Forecasting*, 37(3):1276–1295, 2021.

C. Baumeister and J. D. Hamilton. Structural interpretation of vector autoregressions with incomplete identification: Revisiting the role of oil supply and demand shocks. *American Economic Review*, 109(5):1873–1910, 2019.

C. Baumeister, D. Korobilis, and T. K. Lee. Energy markets and global economic conditions. *Review of Economics and Statistics*, 104(4):828–844, 2022.

M. Betancourt. A conceptual introduction to hamiltonian monte carlo. *arXiv preprint arXiv:1701.02434*, 2017.

M. Billio, R. Casarin, F. Ravazzolo, and H. K. Van Dijk. Time-varying combinations of predictive densities using nonlinear filtering. *Journal of Econometrics*, 177(2):213–232, 2013.

C. M. Carvalho, N. G. Polson, and J. G. Scott. Handling sparsity via the horseshoe. In *Artificial intelligence and statistics*, pages 73–80. PMLR, 2009.

C. M. Carvalho, N. G. Polson, and J. G. Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2): 465–480, 2010.

N.-F. Chen, R. Roll, and S. A. Ross. Economic forces and the stock market. *Journal of business*, pages 383–403, 1986.

C. Conrad, K. Loch, and D. Rittler. On the macroeconomic determinants of long-term volatilities and correlations in us stock and crude oil markets. *Journal of Empirical Finance*, 29:26–40, 2014.

C. Czado. Analyzing dependent data with vine copulas. *Lecture Notes in Statistics, Springer*, 222, 2019.

S. Daul, E. G. De Giorgi, F. Lindskog, and A. McNeil. The grouped t-copula with an application to credit risk. *Available at SSRN 1358956*, 2003.

S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth. Hybrid monte carlo. *Physics letters B*, 195(2):216–222, 1987.

P. Embrechts, F. Lindskog, and A. McNeil. Modelling dependence with copulas. *Rapport technique, Département de mathématiques, Institut Fédéral de Technologie de Zurich, Zurich*, 14:1–50, 2001.

P. Embrechts, A. McNeil, and D. Straumann. Correlation and dependence in risk management: properties and pitfalls. *Risk management: value at risk and beyond*, 1:176–223, 2002.

R. Engle. Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *Journal of Business & Economic Statistics*, 20(3):339–350, 2002.

F. J. Fabozzi, S. T. Rachev, and C. Menn. *Fat-tailed and skewed asset return distributions: implications for risk management, portfolio selection, and option pricing.* John Wiley & Sons, 2005.

D. Gamerman and H. F. Lopes. *Markov chain Monte Carlo: stochastic simulation for Bayesian inference.* CRC press, 2006.

A. Gelman, W. R. Gilks, and G. O. Roberts. Weak convergence and optimal scaling of random walk metropolis algorithms. *The annals of applied probability*, 7(1):110–120, 1997.

A. Gelman, D. Lee, and J. Guo. Stan: A probabilistic programming language for bayesian inference and optimization. *Journal of Educational and Behavioral Statistics*, 40(5):530–543, 2015.

E. Ghysels, V. Kvedaras, and V. Zemlys. Mixed frequency data sampling regression models: the r package midasr. *Journal of statistical software*, 72:1–35, 2016.

Y. Gong, R. Bu, and Q. Chen. What affects the relationship between oil prices and the us stock market? a mixed-data sampling copula approach. *Journal of Financial Econometrics*, 20(2):253–277, 2022a.

Y. Gong, C. Ma, and Q. Chen. Exchange rate dependence and economic fundamentals: A copula-midas approach. *Journal of International Money and Finance*, 123:102597, 2022b.

G. Gorton and K. G. Rouwenhorst. Facts and fantasies about commodity futures. *Financial Analysts Journal*, 62 (2):47–68, 2006.

C. M. Hafner and H. Manner. Dynamic stochastic copula models: Estimation, inference and applications. *Journal of Applied Econometrics*, 27(2):269–295, 2012.

J. D. Hamilton. Oil and the macroeconomy since world war ii. *Journal of political economy*, 91(2):228–248, 1983.

T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. CRC press, 2015.

D. Hosszejni and G. Kastner. Modeling univariate and multivariate stochastic volatility in r with stochvol and factorstochvol. *arXiv preprint arXiv:1906.12123*, 2019.

M. S. Johannes, N. G. Polson, and J. R. Stroud. Optimal filtering of jump diffusions: Extracting latent states from asset prices. *The Review of Financial Studies*, 22(7):2759–2799, 2009.

T. C. Johnson. Volatility, momentum, and time-varying skewness in foreign exchange returns. *Journal of Business & Economic Statistics*, 20(3):390–411, 2002.

K. Kalantar-Zadeh, N. Kuwae, D. Regidor, C. Kovesdy, R. Kilpatrick, C. Shinaberger, C. McAllister, M. Budoff, I. Salusky, and J. Kopple. Survival predictability of time-varying indicators of bone disease in maintenance hemodialysis patients. *Kidney international*, 70(4):771–780, 2006.

L. Kilian. Not all oil price shocks are alike: Disentangling demand and supply shocks in the crude oil market. *American Economic Review*, 99(3):1053–1069, 2009.

P. Maneejuk and W. Yamaka. Predicting contagion from the us financial crisis to international stock markets using dynamic copula with google trends. *Mathematics*, 7(11):1032, 2019.

H. Markowitz. Portfolio selection. *Journal of finance*, 7(1):77–91, 1952.

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

R. B. Nelsen. *An introduction to copulas*. Springer science & business media, 2007.

H. Nguyen and F. Javed. Dynamic relationship between stock and bond returns: A gas midas copula approach. Technical report, Working Paper, 2021.

H. Nguyen and A. Virbickaite. Modeling stock-oil co-dependence with dynamic stochastic midas copula models. Technical report, Working Paper, 2022.

T. Park and G. Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

A. J. Patton. Modelling asymmetric exchange rate dependence. *International economic review*, 47(2):527–556, 2006.

J. Piironen and A. Vehtari. Sparsity information and regularization in the horseshoe and other shrinkage priors. 2017.

C. P. Robert et al. *The Bayesian choice: from decision-theoretic foundations to computational implementation*, volume 2. Springer, 2007.

P. Sadorsky. Oil price shocks and stock market activity. *Energy economics*, 21(5):449–469, 1999.

M. Sklar. Fonctions de repartition an dimensions et leurs marges. *Publ. inst. statist. univ. Paris*, 8:229–231, 1959.

D. R. Smith. Conditional coskewness and asset pricing. *Journal of Empirical Finance*, 14(1):91–119, 2007.

S. Thomas and W. Tu. Learning hamiltonian monte carlo in r. *The American Statistician*, 75(4):403–413, 2021.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

H.-M. Zhu, R. Li, and S. Li. Modelling dynamic dependence between crude oil prices and asia-pacific stock market returns. *International Review of Economics & Finance*, 29:208–223, 2014.

D. M. Zimmer. Time-varying correlation in housing prices. *The Journal of Real Estate Finance and Economics*, 51:86–100, 2015.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.

Appendix A: Transformations of λ_t

Equation (7) implies that λ_t may assume any real value. Thus it must be converted into the appropriate domain of the copula parameter. We follow Hafner and Manner [2012] by specifying $\Lambda(\cdot)$ as in Equation (15) for the Gaussian and Student's t copulas ensuring that $-1 \geq \Theta \geq 1$. Equation (16) for the Clayton ensuring $\Theta > 0$, and as Equation (17) for the Gumbel copula making $1 \leq \Theta < \infty$.

$$\Lambda(\cdot) = \frac{exp(2\lambda) - 1}{exp(2\lambda) + 1} \tag{15}$$

$$\Lambda(\cdot) = e^{\lambda} \tag{16}$$

$$\Lambda(\cdot) = e^{\lambda} + 1 \tag{17}$$

Appendix B: Time Series considered in this study

The tables bellow list the time series considered in this study.

Group 1 – Outcome and Income
Real Personal Income
Real personal income ex transfer receipts
IP Index
IP: Final Products and Nonindustrial Supplies
IP: Final Products (Market Group)
IP: Consumer Goods
IP: Durable Consumer Goods
IP: Nondurable Consumer Goods
IP: Business Equipment
IP: Materials
IP: Durable Materials
IP: Nondurable Materials
IP: Manufacturing (SIC)
IP: Residential Utilities
IP: Fuels
Capacity Utilization: Manufacturing

Group 2 – Labor Market
Help-Wanted Index for United States
Ratio of Help Wanted/No. Unemployed
Civilian Labor Force
Civilian Employment
Civilian Unemployment Rate
Average Duration of Unemployment (Weeks)
Civilians Unemployed - Less Than 5 Weeks
Civilians Unemployed for 5-14 Weeks
Civilians Unemployed - 15 Weeks & Over
Civilians Unemployed for 15-26 Weeks
Civilians Unemployed for 27 Weeks and Over
Initial Claims
All Employees: Total nonfarm
All Employees: Goods-Producing Industries
All Employees: Mining and Logging: Mining
All Employees: Construction
All Employees: Manufacturing
All Employees: Durable goods
All Employees: Nondurable goods
All Employees: Service-Providing Industries
All Employees: Trade, Transportation & Utilities
All Employees: Wholesale Trade
All Employees: Retail Trade
All Employees: Financial Activities
All Employees: Government
Avg Weekly Hours : Goods-Producing
Avg Weekly Overtime Hours : Manufacturing
Avg Weekly Hours : Manufacturing
Avg Hourly Earnings : Goods-Producing
Avg Hourly Earnings : Construction
Avg Hourly Earnings : Manufacturing

Group 3 – Housing
Housing Starts: Total New Privately Owned
Housing Starts, Northeast
Housing Starts, Midwest
Housing Starts, South
Housing Starts, West
New Private Housing Permits (SAAR)
New Private Housing Permits, Northeast (SAAR)
New Private Housing Permits, Midwest (SAAR)
New Private Housing Permits, South (SAAR)
New Private Housing Permits, West (SAAR)

Group 4 - Consumption, orders, and inventories
Real personal consumption expenditures
Real Manu. and Trade Industries Sales
Retail and Food Services Sales
New Orders for Consumer Goods
New Orders for Durable Goods
New Orders for Nondefense Capital Goods
Unfilled Orders for Durable Goods
Total Business Inventories
Total Business: Inventories to Sales Ratio
Consumer Sentiment Index

Group 5 – Money and Credit
M1 Money Stock
M2 Money Stock
Real M2 Money Stock
Monetary Base
Total Reserves of Depository Institutions
Reserves Of Depository Institutions
Commercial and Industrial Loans
Real Estate Loans at All Commercial Banks
Total Nonrevolving Credit
Nonrevolving consumer credit to Personal Income
Consumer Motor Vehicle Loans Outstanding
Total Consumer Loans and Leases Outstanding
Securities in Bank Credit at All Commercial Banks

Group 6 – Interest and exchange rates
Effective Federal Funds Rate
3-Month AA Financial Commercial Paper Rate
3-Month Treasury Bill:
6-Month Treasury Bill:
1-Year Treasury Rate
5-Year Treasury Rate
10-Year Treasury Rate
Moody's Seasoned Aaa Corporate Bond Yield
Moody's Seasoned Baa Corporate Bond Yield
3-Month Commercial Paper Minus FEDFUNDS
3-Month Treasury C Minus FEDFUNDS
6-Month Treasury C Minus FEDFUNDS
1-Year Treasury C Minus FEDFUNDS
5-Year Treasury C Minus FEDFUNDS
10-Year Treasury C Minus FEDFUNDS
Moody's Aaa Corporate Bond Minus FEDFUNDS
Moody's Baa Corporate Bond Minus FEDFUNDS
Trade Weighted U.S. Dollar Index
Switzerland / U.S. Foreign Exchange Rate
Japan / U.S. Foreign Exchange Rate
U.S. / U.K. Foreign Exchange Rate
Canada / U.S. Foreign Exchange Rate

Group 7 – Prices
PPI: Finished Goods
PPI: Finished Consumer Goods
PPI: Intermediate Materials
PPI: Crude Materials
PPI: Metals and metal products:
CPI : All Items
CPI : Apparel
CPI : Transportation
CPI : Medical Care
CPI : Commodities
CPI : Durables
CPI : Services
CPI : All Items Less Food
CPI : All items less shelter
CPI : All items less medical care
Personal Cons. Expend.: Chain Index
Personal Cons. Exp: Durable goods
Personal Cons. Exp: Nondurable goods
Personal Cons. Exp: Services

Group 8 – Stock Market
S&P's Common Stock Price Index: Industrials
S&P's Composite Common Stock: Dividend Yield
S&P's Composite Common Stock: Price-Earnings Ratio
VIX

Group 9 - Additional series
Baumeister's GECON index
Killian's dry cargo single voyage freight rates
Changes in thousands of barrels per day
Baumeister's Oil supply shocks
Baumeister's Oil consumption demand shocks
Baumeister's Oil inventory demand shocks
Baumeister's Economic activity shocks
Baumeister's Real Commodity Price Factor