Bayesian Inference for Generalized Autoregressive Score Models

By Robin Niesert (344760)

Master Thesis in Quantitative Finance
Erasmus School of Economics
Erasmus University Rotterdam

Supervisor: Rutger-Jan Lange
Second assessor: Bart Keijsers

August 10, 2017

Abstract

In this thesis I explore the benefits of adopting a Bayesian methodology when doing inference for generalized autoregressive score (GAS) models. Although analytical results regarding the form of the posterior or its conditional will generally not be available for this class of models, I show that for most simple GAS models several novel Markov chain Monte Carlo methods can be applied to enable accurate Bayesian inference in very reasonable time frames. I consider three illustrative empirical applications of GAS models where particular emphasize is placed on contrasting Bayesian inferences with those stemming from the traditional approach of estimating GAS models using the Maximum Likelihood (ML) method. I argue that there are certain complexities intrinsic to models in the GAS framework that can be dealt with far more naturally under a Bayesian methodology, such as (i) the non-nestedness of comparable models that arises as a consequence of the freedom of choice in scaling matrices and parametrization of GAS models and (ii) the "curse of dimensionality" problem that occurs primarily for multivariate GAS models. The logical Bayesian solution to the former is to apply Bayesian model comparison techniques - which I explore in the context of dynamic intensity factor models applied to credit rating data - whereas the later can be addressed by imposing additional structure on the parameter space using hierarchical prior setups - which I illustrate on a time-varying covariance GAS Student-t model. Additionally, I demonstrate how the typically high degree of non-linearity with which parameters enter the likelihood for GAS models cause slow convergence to the normal distribution for the parameters - as is highlighted for the Beta-Gen-t-EGARCH volatility model. Implying that considerable sample sizes are necessary to allow for valid appeals to the asymptotic convergence arguments used in ML estimation.

Keywords: Generalized autoregressive score (GAS) models, Bayesian inference, Markov chain Monte Carlo, Bayesian model comparison, hierarchical multivariate GAS-t model

Contents

1 Introduction

In financial econometrics the modeling of time series variables is often central to the research objective. Many of the models that prove most effective at describing financial time series utilize time-varying specifications for one or more of the model parameters. Recently, Creal et al. (2013) proposed a generic class of observation-driven, time-varying parameter models, dubbed Generalized Autoregressive Score (GAS) models. GAS models are characterized by an update of the time-varying parameters that is driven by the gradient of the log-likelihood with respect to these parameters; a quantity known as the score in the statistics literature.

GAS models encompass many of financial econometrics' most familiar time-varying parameter models such as the generalized autoregressive conditional heteroskedastic (GARCH) model by Bollerslev (1986) and Engle & Bollerslev (1986), autoregressive conditional duration (ACD) model due to Russell & Engle (1998) and multiple time-varying parameter models such as the dynamic conditional correlation (DCC) model and the autoregressive conditional multinomial (ACM) model of Engle (2002) and Engle & Russell (1998) respectively. These models however, constitute a mere subset of the wide-variety of useful model specification that the GAS framework allows for. The original working paper by Creal et al. (2011b) illustrates the versatility of the GAS framework.

In this thesis I apply Bayesian methods to do inference on models that fall within the GAS framework - as opposed to the usual approach of estimating GAS models with the method of Maximum Likelihood (ML). To my knowledge no preceding work has applied Bayesian methods to GAS models, other than for the previously mentioned familiar time-varying parameter models which the GAS framework encompasses. The arguments in favor of Bayesian methods over ML that are identified in the literature for GARCH models (Ardia & Hoogerheide, 2010, Virbickaite et al., 2015), directly translate and arguably apply even more convincingly for the more general class of GAS models.

First, although ML estimation of GAS models is relatively straightforward, the validity of standard asymptotic properties of ML estimators has thus far only been established for certain limited classes of GAS models (see e.g. Blasques et al. (2014, 2016)). The challenges with generalizing asymptotic properties are due to the in general highly nonlinear way in which the dependent variable enters the update equation for the time-varying parameters. Moreover, even when asymptotic properties of ML estimators apply, the empirically often high persistence of time-varying parameters coupled with constraints to enforce stationarity or non-negativity, are likely to induce finite-sample bias in time-varying parameter models (Hwang & Valls Pereira, 2006). Bayesian methods inherently make no appeal to asymptotic convergence arguments and hence provide a logical alternative to ML estimation.

Secondly, in practical applications of time-varying parameter models we are often interested in nonlinear functions of the estimated parameters. Performing inference on such nonlinear functions of the parameters is complicated using the ML method. Whereas using Bayesian methods, a nonlinear transformation of the draws from the posterior can in most cases straightforwardly be interpreted as draws from the posterior of the transformed quantities and can readily be used for inference. Consider for instance the case of the Beta-Gen-t-EGARCH model of Harvey & Lange (2017), which falls under the GAS framework. Its general application is to model the volatility or variance of financial instruments, yet the second order central moment is a highly nonlinear function of the time-varying scale parameter and the shape parameters. In disciplines such as risk management, predictions of such volatility are highly relevant but dangerous to interpret without an indication of the associated

uncertainty, as evidenced by the analysis in Section 4.1.3, which reveals a substantially long right tail for the posterior of the volatility predicted by the Beta-Gen-t-EGARCH model.

Thirdly, the GAS framework contains several degrees of freedom in its model specification. This can lead to a variety of models designed to describe the same phenomena, but for which standard likelihood based model comparison tests can not be applied due to the non-nestedness of the models. The popular Beta-t-EGARCH and t-GAS model by Harvey & Chakravarty (2008) and Creal et al. (2011a) respectively, are for example both volatility models based on the assumption of a Student-t distributed dependent variable, but the different link functions from time-varying parameter to scale parameter limit formal model comparison based on the likelihood. Similarly, the appropriate scaling of the score is still an open question and it is hard to determine based on model fit when ML is used for estimation. Currently, such model comparisons are usually informal and based on quantities such as the mode of the log-likelihood or information criteria such as the Bayesian information criterion (BIC) that use standard penalties for the number of parameters. Bayesian model comparison using Bayes factors allows such model specification choices to be formalized. Unlike the traditional ML methods, Bayes factors take full account of the parameter uncertainty in the models being compared. As illustrated in Section 4.2, comparison in terms of Bayes factors can therefore lead to different conclusions as likelihood or information criterion based comparison.

Finally, as GAS models become more complex and the number of time-varying parameters increases, the number of autoregressive parameters - for fully parameterized GAS models - increases quadratically. Traditionally the approach to maintaining parsimonious models for which the likelihood optimization converges, is to impose parameter restrictions and factor structures on the time-varying parameters. In a Bayesian framework a natural alternative approach to enforce parsimony is by means of hierarchical priors. In Section 4.3 I apply a hierarchical prior setup to the multivariate Student-t covariance model of Creal et al. (2011a), resulting in significant reductions in parameter uncertainty while retaining most of the flexibility of an unrestricted model. The resulting hierarchical model outperforms both restricted and unrestricted versions of the Student-t covariance model in terms of Bayesian model probabilities.

Like for GARCH models, Bayesian inference on GAS models will in general be challenging due to the recursive specification of the time-varying parameters which convolute the way the model parameters interact with the dependent variable. Consequently, known forms for neither the full or marginal posteriors are obtainable such that we need to rely on Markov chain Monte Carlo (MCMC) methods that work on generic distributions. In addition the highly nonlinear way in which parameters enter the likelihood can cause irregularities in the posterior such as skewness, fat-tails and nonlinear dependencies, which might challenge the effectiveness of standard MCMC methods. As will be argued in Section 3, there are several promising choices among the existing MCMC methods for time-varying parameter models. Section 4.1 illustrates that the Hamiltonian Monte Carlo (HMC) method proves particularly well suited to cope with the challenges posed by a typical GAS model posterior.

I proceed by introducing the GAS model in its generic form along with the specific modeling choices that are typical for GAS models in Section 2. Section 3 presents three MCMC algorithms - the Griddy Gibbs of Ritter & Tanner (1992), AdMit-MH by Hoogerheide (2006) and HMC due to Duane et al. (1987). All three have been successfully applied to time-varying parameter models in the literature and can be applied more generally to GAS models. Section 3 also introduces Bayesian model comparison and hierarchical modeling and discusses how these techniques enable inference

generally unavailable in a frequentist setting. Section 4 discusses multiple illustrative empirical applications of GAS models. In Section 4.1 the Beta-Gen-t-EGARCH model is analyzed and serves as a comparative example for the three MCMC methods. In Section 4.2 the dynamic pooled marked point process models of Creal et al. (2013) with different factor specifications and a variety of scaling matrices are compared by means of both Bayes factors and informal non-nested model comparison tools such as the BIC. Section 4.3 demonstrates how Bayesian hierarchical modeling can be used in the GAS-t covariance model of Creal et al. (2011a) to provide a more natural and effective way to cope with the "the curse of dimensionality" problem typically associated with time-varying covariance models, than the common approach of enforcing parameter restrictions. Section 5 concludes with a review of the most important findings and a discussion of promising future applications of Bayesian methods for GAS models.

2 The GAS Model

Following Creal et al. (2011b, 2013), I assume that the dynamics of a $k \times 1$ vector of dependent variables $\boldsymbol{y}_t$ are governed by a probability distribution, which conditions on the set of preceding values of the dependent variables $\boldsymbol{Y}_{t-1} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_{t-1}\}$, the set of contemporaneous and preceding time-varying parameters $\boldsymbol{F}_t = \{\boldsymbol{f}_1, \boldsymbol{f}_2, ..., \boldsymbol{f}_t\}$ and a $d \times 1$ vector of static parameters denoted by $\boldsymbol{\theta}$. Let this distribution be specified as

$$p(\boldsymbol{y}_t | \boldsymbol{Y}_{t-1}, \boldsymbol{F}_t, \boldsymbol{\theta}), \tag{1}$$

for $t = 1, 2, ..., T$. The update equation of the $n \times 1$ time-varying parameter vector $\boldsymbol{f}_t$ is defined as

$$\boldsymbol{f}_t = \boldsymbol{\omega} + \boldsymbol{A}\boldsymbol{s}_{t-1} + \boldsymbol{B}\boldsymbol{f}_{t-1}, \tag{2}$$

for $t = 2, 3, ..., T$, where $\boldsymbol{\omega}$, $\boldsymbol{A}$ and $\boldsymbol{B}$ are the autoregressive coefficients that are part of the set of static parameters $\boldsymbol{\theta}$. The parameter matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ can be dense, but are often restricted to diagonal matrices. The process is initialized with $\boldsymbol{f}_1$ set to some fixed value usually inspired by sample moments of the dependent variable. In several instances the time-varying parameter process will be reparameterized as

$$\boldsymbol{f}_t = (\boldsymbol{I}_n - \boldsymbol{B})\tilde{\boldsymbol{\omega}} + \boldsymbol{A}\boldsymbol{s}_{t-1} + \boldsymbol{B}\boldsymbol{f}_{t-1}, \tag{3}$$

where $\boldsymbol{I}_n$ denotes the n-dimensional identity matrix. Doing so decorrelates the parameters $\boldsymbol{\omega}$ and $\boldsymbol{B}$, which greatly improves the performance for certain MCMC methods.

The vector $\boldsymbol{s}_t$ is defined by

$$\boldsymbol{s}_t = \boldsymbol{S}_t \boldsymbol{\nabla}_t,$$

where $\boldsymbol{\nabla}_t = \partial \ell_t / \partial \boldsymbol{f}_t$ is the score of the time-varying parameters and $\boldsymbol{S}_t$ is a scaling matrix. Here $\ell_t = \log(p(\boldsymbol{y}_t | \boldsymbol{Y}_{t-1}, \boldsymbol{F}_t, \boldsymbol{\theta}))$ is used to denote the log-likelihood for a single observation $\boldsymbol{y}_t$. The scaling matrix matrix is usually set equal to a power of the inverse Fisher information matrix for a single observation,

$$\boldsymbol{S}_t = \boldsymbol{\mathcal{I}}_t^{-a}, \tag{4}$$

for $a = 0, 1/2, 1$ and

$$\boldsymbol{\mathcal{I}}_t = -\operatorname{E}\left(\frac{\partial^2 \ell_t}{\partial \boldsymbol{f}_t \partial \boldsymbol{f}_t'}\right). \tag{5}$$

The specifications of $a = 0$ or $a = 1$ have the benefit of the convenient interpretations as a gradient ascent or a Newton-Raphson type update respectively for ℓ_t. In the literature on GAS models the

choice of scaling matrix centers around its implications for proving the stationarity and ergodicity conditions of the time-varying parameter process (2) (Blasques et al., 2014). In this thesis I instead consider how the choice of $\boldsymbol{S}_t$ affects model fit in terms of Bayesian posterior model probabilities. Nelson (1996) for instance proves analytically the optimal filter properties for GARCH models when $a = 1/2$. Intuitively it also seems advantageous to include second order information in the update.

The process (2) is in general covariance-stationary if the variance of $\boldsymbol{s}_t$ is finite and the eigenvalues of the matrix of autoregressive coefficients $\boldsymbol{B}$ are less than one in modulus. In case the scaling matrix is of the form (4), the finite variance of $\boldsymbol{s}_t$ is guaranteed if $a = 1/2$ and follows for $a = 0$ or $a = 1$ if the Fisher information matrix (5) is bounded (Creal et al., 2011b). For all models presented in Section 4, the constraints on the eigenvalues of $\boldsymbol{B}$ are enforced during the estimation procedure.

Apart from the choice of scaling matrix and probability distribution, many variations of the GAS model are obtained by the choice of parameterization of the model (1). Creal et al. (2011b) describe the use of a link function to obtain more convenient and easier to estimate models. Since the process (2) allows $\boldsymbol{f}_t$ to range over the entirety of $\mathbb{R}^n$, the link function is particularly useful if $\boldsymbol{f}_t$ needs to be constrained to a certain range. For example, exponential GARCH (EGARCH) models specify $f_t = \log(\sigma_t^2)$, where σ_t is a time-varying scale parameter (Harvey, 2010). The logarithmic link function naturally ensures that the variance process remains positive.

Alternatively, the link function is commonly used for imposing a factor structure on the time-varying parameters (see e.g Bartels & Ziegelmann (2016) or Creal et al. (2014)). This simplifies estimation by reducing the number of time-varying parameters and in many cases it is also reasonable to assume that the dynamics of a group of parameters is driven by a much smaller set of time-varying factors. Since different factor specifications typically result in non-nested models, determining the optimal number of factors is in most cases not straightforward using frequentist methods. Bayesian model comparison does offer such a formal approach to comparing different factor specifications, as will be illustrated in Section 4.2 on dynamic pooled marked point process factor models.

Besides the use of factor structures, direct restrictions on the autoregressive coefficients such as imposing $\boldsymbol{A}$ and $\boldsymbol{B}$ to be diagonal, is another common method for achieving more parsimonious parameterizations. Such parameter restrictions are however a rather crude approach and might significantly limit the models capacity to capture the dynamics of $\boldsymbol{f}_t$ (see e.g. Burda & Maheu (2013)). The hierarchical modeling approach explored in Section 4.3 offers a more intuitive alternative way to induce parsimony while sacrificing considerably less in terms of flexibility.

The full specification of a GAS model thus involves four generic choices: 1.) the conditional probability distribution of the dependent variables $p(\cdot|\cdot)$, 2.) the scaling matrix $\boldsymbol{S}_t$, 3.) the link function and 4.) the number of free parameters in $\boldsymbol{\theta}$. Ordinarily, there will be many different viable combinations to describe one particular time-varying parameter process. The resulting models are often non-nested and traditional frequentist methods therefore typically fail to provide a coherent evidence based approach to support such modeling decisions. In this thesis I focus extensively on how Bayesian methods can improve how we navigate the four modeling choices inherent to the GAS framework, either by means of Bayesian model comparison or by hierarchical prior specifications that allow a subset of these choices (particularly the degree of parametric restriction) to be partially incorporated into the model as lower level hyperparameters.

3 Bayesian Inference

In a Bayesian setting the central object of interest is the posterior distribution of the parameters

$$p(\boldsymbol{\theta}|\boldsymbol{Y}_T) \propto p(\boldsymbol{Y}_T|\boldsymbol{\theta})p(\boldsymbol{\theta}), \tag{6}$$

which is the product of the likelihood $p(\boldsymbol{Y}_T|\boldsymbol{\theta})$ and the prior $p(\boldsymbol{\theta})$, which reflects prior beliefs about the parameters. The likelihood for GAS models can be further decomposed as

$$p(\boldsymbol{Y}_T|\boldsymbol{\theta}) = \Pi_{i=1}^{T} p(\boldsymbol{y}_t|\boldsymbol{Y}_{t-1}, \boldsymbol{F}_t, \boldsymbol{\theta}).$$

Bayesian inference typically involves the computation of expectations of some function of the parameters $g(\boldsymbol{\theta})$ with respect to the posterior distribution

$$\mathrm{E}_{\boldsymbol{\theta}|\boldsymbol{Y}_T}(g(\boldsymbol{\theta})) = \int_{\mathbb{R}^d} g(\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{Y}_T)\mathrm{d}\boldsymbol{\theta}. \tag{7}$$

Nearly all quantities of interest, such as estimates of the parameters, but also model probabilities, or highest posterior density intervals, all can be expressed as expectations of the form (7). Taking such expectation however implies computing an integral. For GAS models the set of parameters $\boldsymbol{\theta}$ includes the autoregressive coefficients of the time-varying parameter process whose relations to the dependent variables $\boldsymbol{y}_t$ are highly convoluted. The implication is that the likelihood for GAS models will generally be of a form that renders analytical solutions to the integral in (7) unobtainable. Numerical integration strategies are also infeasible for more complex GAS models since the computational burden quickly turns restrictive as $\boldsymbol{\theta}$ increases in dimension.

In order to compute these integrals efficiently, even when $\boldsymbol{\theta}$ is of high dimension, therefore requires the ability to simulate from the posterior. Using Monte Carlo integration the resulting draws can then be converted to the desired expectations (Geweke, 1989, Hammersley et al., 1965). For most empirically relevant GAS models, direct sampling from the posterior is not possible due to the fact that the right hand-side of (6) will not be reducible to a distribution for $\boldsymbol{\theta}$ that belongs to a family of closed-form distributions. Luckily the class of algorithms known as Markov Chain Monte Carlo (MCMC) enables sampling from such difficult posteriors. Below, I briefly describe the principle of MCMC and discuss the methods applied in Section 4 in more detail.

3.1 MCMC methods

Given an initial state $\boldsymbol{\theta}^{(1)}$, MCMC methods generate a Markov chain $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, ..., \boldsymbol{\theta}^{(M)}$ whose distribution converges to a target distribution; that is the posterior of $\boldsymbol{\theta}$ in this case. Markov chains are constructed by the sequential application of a Markov transition kernel, which is defined as a random map from a given state $\boldsymbol{\theta}^{(m-1)}$ to a new state $\boldsymbol{\theta}^{(m)}$. Alternatively it can be thought of as a conditional probability distribution $p(\boldsymbol{\theta}^{(m)}|\boldsymbol{\theta}^{(m-1)})$. For MCMC methods the Markov transition kernel needs to be carefully constructed so that it has the target distribution as its invariant distribution, meaning that if we have a draw from the target distribution and sequentially applied the transition kernel to generate a sample, this sample should be distributed according to the target distribution. Due to the dependence on the previous state, the sample need however not consist of independent draws and Markov chains in fact commonly display strong autocorrelation reducing their effective sample size.

Among the valid choices of transition kernels for MCMC methods there is significant variation in the effectiveness with which the target distribution is explored; where I consider effectiveness as a product of not just the autocorrelation in the Markov chain, but also the computational cost to obtain a draw. The degree of autocorrelation in the chain, typically depends on how much information of the target distribution a transition kernel can incorporate in its transitions. The popular Gibbs sampler is for instance very effective as it exploits the information in the conditional posteriors, which is especially effective if the parameters are largely independent since that would imply the conditionals capture most of the full posterior distribution. On the opposite end of the spectrum, a random walk Metropolis-Hastings (RW-MH) sampler uses a transition kernel that incorporates very little information with regards to the target distribution. As the name suggest the result is a random walk like Markov chain; meaning high autocorrelations in the draws and ineffective exploration of the target distribution.

In applying MCMC to GAS model posteriors we are limited in our choices of methods by the fact that we know very little about the posterior distributions. For instance, analytical expressions for the conditionals are usually not available for observation driven time-varying parameter models (see e.g. Bauwens & Lubrano (1998)) making the Gibbs sampler inapplicable. Also, as a result of the likelihood being expensive to evaluate, algorithms such as the the RW-MH that require extremely long chains - and thereby many function evaluations - to compensate for the high correlation among consecutive draws, are likely to be inefficient. Effective MCMC methods for GAS models therefore, devote part of their computational resources to extract information about the target distribution, which then informs the Markov transitions. In doing so these methods balance a trade-off between effectiveness in terms of low autocorrelation in the Markov chain and effectiveness in terms of the computational cost per draw. The three methods that I focus on in this section and in the comparative analysis of Section 4.1, the Griddy Gibbs sampler by Ritter & Tanner (1992), the Adaptive Mixture of Student-t distributions - Metropolis-Hastings algorithm by (Hoogerheide, 2006) and Hamiltonian Monte Carlo due to Duane et al. (1987), rely exactly on such a strategy of first gathering information about the target distribution prior to producing a draw. All three have been successfully applied to univariate GARCH models (see Ardia & Hoogerheide (2010), Bauwens & Lubrano (1998) and Takaishi (2007)) and are naturally suitable for more general form GAS models.

3.1.1 Griddy Gibbs Sampler

The Griddy Gibbs method resolves the limitation with regards to not knowing the conditional posteriors of $\boldsymbol{\theta}$, using numerical integration. Rather then applying numerical integration directly to the full posterior (6) - which would require constructing a d-dimensional grid, where d denotes the dimension of $\boldsymbol{\theta}$ - the Griddy Gibbs sampler employs the more efficient strategy of numerically integrating d one-dimensional integrals.[1] Following Bauwens & Lubrano (1998) I use a trapezoidal integration rule combined with linear interpolation to convert the numerical integration of the conditionals to an approximation of the inverse conditional CDFs. The inverse CDF method then enables sampling from the approximated conditional posteriors. Using the standard Gibbs transition kernel, for $m = 2, 3, ..., M$, we draw the individual parameters $\theta_i^{(m)}$ for all $i = 1, ..., d$, separately as

[1] Assuming we require 20 grid points per dimension, numerical integration requires 20^d posterior evaluations as opposed to $20d$ for Griddy Gibbs.

follows:

$$\begin{aligned}
\theta_1^m &\sim p(\theta_1^m|\theta_2^{m-1}, \theta_3^{m-1}, ..., \theta_d^{m-1}), \\
\theta_2^m &\sim p(\theta_2^m|\theta_1^m, \theta_3^{m-1}, ..., \theta_d^{m-1}), \\
&\vdots \\
\theta_d^m &\sim p(\theta_d^m|\theta_1^m, \theta_2^m, ..., \theta_{d-1}^m).
\end{aligned}$$

Jointly these draws constitute a draw $\boldsymbol{\theta}^{(m)}$ from the full posterior (6). More details on the implementation of Griddy Gibbs can be found in Bauwens & Lubrano (1998) and the original paper by Ritter & Tanner (1992).

The Griddy Gibbs sampler is powerful in that it is effective regardless of the complexity of the posterior. It can handle asymmetries, skewness, fat tails and even multi-modalities, making it particularly attractive for GAS models. Strong correlations can stifle Griddy Gibbs, drastically increasing the required number of grid points and inducing high autocorrelations in the chain. Reparameterization such as in (3) can however often sufficiently decorrelate the parameters. Also, in order to limit the number of grid points required it is necessary to restrict the grid to a range with reasonable probability mass. For GAS models we usually have a reasonable idea as to a plausible range for the parameters, but it might still be challenging to correctly tune the range of the grid. A more serious limitation of Griddy Gibbs is that, although the algorithm is far more efficient than full numerical integration, the computational cost per draw still far outweigh the costs per draw for most other MCMC methods, as demonstrated in a comparative analysis by Asai (2006) and confirmed by the analysis in Section 4.1. Also computational costs typically scale quadratically with the dimension of $\boldsymbol{\theta}$, contrary to for instance HMC. For GAS models with multiple time-varying parameters, which logically have more static parameters, Griddy Gibbs is thus unsuitable.

3.1.2 Metropolis-Hastings samplers and the AdMit method

The Adaptive Mixture of Student-t densities (AdMit) is an algorithm developed by Hoogerheide (2006) that fits a mixture of Student-t distributions to a target density. The fitted mixture is used as a candidate density for either importance sampling (IS) or an independence chain Metropolis-Hastings (MH) sampler. I consider just the MH variant. Like the Griddy Gibbs algorithm, AdMit-MH is able to deal with challenging posteriors with properties such as skewness, fat tails and multimodality (Ardia et al., 2009). Since a comparative review of MCMC methods for GARCH(1,1) models proves a simpler version of the algorithm - with a proposal density based on a single fitted Student-t distribution - to be the most efficient (Asai, 2006), the AdMit-MH algorithm is likely among the more effective general purpose MCMC methods for GARCH and by extension GAS models.

First, I briefly introduce the generic MH algorithm by Metropolis et al. (1953) and Hastings (1970) and its two most popular variants: the random walk and independence chain sampler (see Chib & Greenberg (1995) for a general reference on MH). MH consists of a proposal and an accept or reject step. Let $\boldsymbol{\theta}^*$ denote the proposed new state, which is generated by a draw from a candidate density $q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(m-1)})$. The proposal is accepted, meaning we set $\boldsymbol{\theta}^{(m)} = \boldsymbol{\theta}^*$, with probability

$$\alpha(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(m-1)}) = \min\left(1, \frac{k(\boldsymbol{\theta}^*)/q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(m-1)})}{k(\boldsymbol{\theta}^{(m-1)})/q(\boldsymbol{\theta}^{(m-1)}|\boldsymbol{\theta}^*)}\right),$$

where $k(\boldsymbol{\theta})$ is a kernel of the posterior distribution. On the other hand, if we reject we set $\boldsymbol{\theta}^{(m)} = \boldsymbol{\theta}^{(m-1)}$.

Both the random walk and the independence chain rely on special cases of the candidate density. For the random walk sampler the candidate density is symmetric in the preceding draw such that $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(m-1)}) = q(\boldsymbol{\theta}^{(m-1)}|\boldsymbol{\theta}^*)$. Proposals can therefore straightforwardly be generated as

$$\boldsymbol{\theta}^* = \boldsymbol{\theta}^{(m-1)} + \epsilon\boldsymbol{\varrho}, \tag{8}$$

where ϵ is a tuning parameter for the step size and $\boldsymbol{\varrho}$ is a random variate drawn from a symmetric distribution with zero mean vector and a scale matrix that is preferably set to the inverse Hessian of the log kernel evaluated at its mode. Alternatively, one can generate an initial posterior sample from a warm up run with an identity scale matrix and then reset the scale matrix to the sample covariance matrix and discarding the warm up run. This procedure can also be repeated several times until reasonable convergence of the sample covariance estimator is achieved. Proposals are accepted or rejected based on the simplified acceptance probability $\alpha(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(m-1)}) = \min(1, k(\boldsymbol{\theta}^*)/k(\boldsymbol{\theta}^{(m-1)}))$. The step size parameter should be tuned to target acceptance rates of around 0.5 for models with few parameters and 0.25 for moderate to higher dimensional parameter spaces (Chib & Greenberg, 1995).

Given the description of the random walk sampler provided above, the sampler thus only utilizes information regarding the covariance of the posterior to inform its Markov transitions. As a result, the Markov chains of the random walk sampler typically display notoriously high autocorrelations (Neal, 2011). For GAS models with computationally expensive likelihoods, the random walk sampler is therefore likely to be inefficient relative to the independence chain algorithm described next. Therefore I include only an independence chain algorithm in the comparison presented in Section 4.1. The random walk sampler is however applied in Section 4.2 because of the samplers ease of implementation and the particular model having a somewhat less complex and expensive likelihood as common for GAS models.

As the name suggests, the independence chain sampler uses a candidate density for which the proposal density is independent of the previous draw. Under these conditions, the acceptance probability (8) simplifies to $\alpha(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(m-1)}) = \min((1, k(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}^{(m-1)})/k(\boldsymbol{\theta}^{(m-1)})q(\boldsymbol{\theta}^*))$. The effectiveness of the transition kernel for the independence chain sampler is determined entirely by how well the candidate density $q(\boldsymbol{\theta})$ fits the target density $p(\boldsymbol{\theta}|\boldsymbol{Y}_T)$. This is where AdMit comes in, as it provides an automated method for abstracting information from the target density and applying it in the construction of an effective candidate density.

The mixture of Student-t candidate density takes the form

$$q(\boldsymbol{\theta}) = \textstyle\sum_{s=1}^{S} \pi_s t_d(\boldsymbol{\theta}|\boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s, \nu),$$

where S denotes the number of mixtures, $t_d(\cdot|\cdot)$ the d dimensional Student-t density, π_s the mixing probabilities and $\boldsymbol{\mu}_s$ and $\boldsymbol{\Sigma}_s$ the mean and scale matrix of the s-th Student-t mixture. The degrees of freedom parameter ν is usually fixed at 1 for all components of the mixture distribution. This ensures that the mixture is fatter tailed as the target distribution, which is vital to the success of independence chain MH.

The mixtures are fitted using a series of optimizations. The first mixture is obtained as $\boldsymbol{\mu}_1 = \text{argmax}_{\boldsymbol{\theta}} \log k(\boldsymbol{\theta})$ and $\boldsymbol{\Sigma}_1 = -[\partial^2 \log k(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}']^{-1}_{\boldsymbol{\theta}=\boldsymbol{\mu}_1}$. Initially the candidate density $q(\boldsymbol{\theta})$ is set equal to the first mixture. Additional components are added to $q(\boldsymbol{\theta})$ by a series of optimizations of the log of an importance sampling weights function $\log w(\boldsymbol{\theta}) = \log k(\boldsymbol{\theta}) - \log q(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. At each optimization step the optimum and negative inverse Hessians of the weights function are used as the mean and covariance matrix of the component that is added. The adding of components stops when a statistic known as the coefficient of variation (CV) - defined as the standard deviation over the mean - for the importance sampling weights ($w(\boldsymbol{\theta}) = k(\boldsymbol{\theta})/q(\boldsymbol{\theta})$) no longer improves by more than a certain percentage by adding an additional component. I use the default value of 10% for this percentage (Ardia et al., 2009). To compute the CV requires that we draw a sample from $q(\boldsymbol{\theta})$ to estimate the mean and standard deviation of the importance sampling weights, each time we decide whether to add a component to the candidate density. Additionally, after each time a component is added the mixing probabilities π_s need to be optimized, which requires another sample from $q(\boldsymbol{\theta})$. The mixing probabilities are optimized with respect to the squared sample CV. For the reader interested in the implementation, Ardia & Hoogerheide (2010) and Ardia et al. (2009) provide detailed descriptions of AdMit and also motivate the choice of the stopping criteria based on the CV.

All in all, AdMit-MH will have substantial upfront computational cost, but low incremental cost to obtain the draws once the mixture is fitted. A drawback is the fact that as with ML, we do need to optimize the log-likelihood. In spite of the popularity of ML, the actual execution of the numerical optimization is not always straightforward for GAS models (Ardia et al., 2016). The score function can induce numerical instabilities in the likelihood function and even if the optimization is numerically stable, solutions are often close to the boundaries imposed by the stationarity condition resulting in poor convergence and leaving the Hessian non-positive definite. In practice I find this challenges the universal applicability of the AdMit method for GAS models. Furthermore, little is known about how the algorithm scales with dimensionality. In general, as the dimension of $\boldsymbol{\theta}$ increases, the majority of probability mass will rapidly center away from the mode (Betancourt, 2017) and if we combine this effect with lots of variation in curvature, the number of mixtures required to properly fit the posterior might quickly become unmanageable. Hence, like Griddy Gibbs, AdMit-MH is probably not best suited for high dimensional GAS models such as the covariance matrix model analyzed in Section 4.3.

3.1.3 Hamiltonian Monte Carlo

The methods described thus far are both likely to struggle in higher dimensions; a limitation that the methods share with the traditional ML method. But currently higher dimensional time-varying processes are receiving much research interest in financial econometrics (Bauwens et al., 2006). As we are trying to advance our understanding of how financial instruments interact with each other, the natural solution is to model groups of financial instruments jointly, explaining the rising interest in multivariate GARCH models, copula models and multivariate intensity and duration models. Models of multiple dependent variables logically come with increased dimension of the parameter space. In contrast to the Griddy Gibbs and AdMit-MH sampler, Hamiltonian Monte Carlo (HMC) is proving greatly successful in high dimensions.[2] Moreover, like Griddy Gibbs and AdMit-MH, HMC can be applied to any posterior, provided we can take the derivative of the log of its kernel - where the functional form of the kernel is usually just the right-hand side of (6). This will generally

[2] Neal (2011) shows that under fairly general assumptions, the amount of computation time for HMC will typically grow in proportion to $d^{5/4}$, whereas for RW-MH it grows in proportion to d^2. These costs assume linear scaling of the computational time for function and gradient evaluations w.r.t d.

not be a restriction for GAS models however. HMC should therefore be uniquely positioned to enable Bayesian inference in multivariate time-varying parameter processes, as also evidenced by several recent studies (e.g. Burda & Maheu (2013) and Burda (2015)).

Similar as for the Griddy Gibbs and Admit-MH, I will focus on describing the mechanics of HMC[3], how it's implemented in practice - with particular focus on how to tune the algorithm - and the expected efficiency of the HMC transition kernel relative to other MCMC methods. For readers unfamiliar with the method, but interested in gaining more intuition and insight for how and why the method works so well I recommend a relatively recent introduction to HMC by Betancourt (2017), where HMC is motivated on the basis of the inherent geometric properties of high-dimensional probability distributions. The ideas presented there in a relatively easy to understand manner are mostly based on a more formal foundation of the algorithm in terms of differential geometry which is presented in Betancourt et al. (2017), although that particular discussion is not as accessible without a working knowledge of the field of differential geometry.

HMC is inspired by a theory from physics known as Hamiltonian mechanics. As a consequence much of the terminology used to describe HMC in the statistical literature has carried over from physics. To stay consistent with preceding work, I too use this terminology. HMC augments the parameter space with an additional d momentum variables $\boldsymbol{\gamma}$ - one for each parameter in $\boldsymbol{\theta}$ - that are Gaussian distributed $\mathcal{N}_d(\mathbf{0}, \boldsymbol{M})$ and where the covariance of the momenta $\boldsymbol{M}$ is known as the mass matrix. The momenta are independent from the parameters $\boldsymbol{\theta}$, implying that the negative log of the kernel of their joint distribution, known as the Hamiltonian, is of the following form

$$H(\boldsymbol{\theta}, \boldsymbol{\gamma}) = -\log k(\boldsymbol{\theta}) + \frac{1}{2}\boldsymbol{\gamma}'\boldsymbol{M}^{-1}\boldsymbol{\gamma}.$$

The Hamiltonian decomposes in two parts, the negative log of the posterior kernel $k(\boldsymbol{\theta})$ and the negative log of the kernel of the momenta. The former is labeled the potential energy function $U(\boldsymbol{\theta}) = -\log k(\boldsymbol{\theta})$ and the latter as the kinetic energy function $K(\boldsymbol{\gamma}) = \frac{1}{2}\boldsymbol{\gamma}'\boldsymbol{M}^{-1}\boldsymbol{\gamma}$.

Each draw using HMC starts with a sampling of the momenta variable from $\mathcal{N}_d(\mathbf{0}, \boldsymbol{M})$. Starting with $\boldsymbol{\theta}$ equal to the previous draw $\boldsymbol{\theta}^{m-1}$, a new proposal is generated by following along a vector field defined by a set of differential equations known as Hamilton's equations. In these equations, both $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ are modeled as changing with respect to time. Since this conception of time is continuous, I use τ to denote this continuous time concept, distinguishing it from t, which is reserved for the indexing of time series observations. Hamilton's equations are defined as

$$\begin{aligned}
\frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}\tau} &= \frac{\partial}{\partial \boldsymbol{\gamma}} H(\boldsymbol{\theta}, \boldsymbol{\gamma}) = \boldsymbol{M}^{-1}\boldsymbol{\gamma}, \\
\frac{\mathrm{d}\boldsymbol{\gamma}}{\mathrm{d}\tau} &= -\frac{\partial}{\partial \boldsymbol{\theta}} H(\boldsymbol{\theta}, \boldsymbol{\gamma}) = -\frac{\partial}{\partial \boldsymbol{\theta}} \log k(\boldsymbol{\theta}).
\end{aligned} \tag{9}$$

By integrating Hamilton's equations, starting from an initial state $(\boldsymbol{\theta}^{(m-1)}, \boldsymbol{\gamma})$, for some fixed amount of integration time $\mathcal{T}$ we end up at a new state $(\boldsymbol{\theta}^*, \boldsymbol{\gamma}^*)$. Since $\boldsymbol{\theta}$ is independent of the momenta $\boldsymbol{\gamma}$, we can discard the momenta $\boldsymbol{\gamma}^*$ and treat $\boldsymbol{\theta}^*$ as a draw from the posterior. See e.g. Pakman & Paninski (2014) for a proof of how this procedure serves as a valid Markov transition kernel that

[3]More specifically, I describe - and use throughout this thesis - Euclidean HMC, where the specification refers to the fact that the algorithm uses a Gaussian kinetic energy function that is independent of the parameters $\boldsymbol{\theta}$. The distinction is due to Betancourt (2013), but since most applications of HMC in applied statistics use this Euclidean adaption of the HMC algorithm I will forgo the specification throughout the main text.

preserves the target distribution.

By construction the trajectory described by Hamilton's equation initially guide the parameters $\boldsymbol{\theta}$ away from their starting point ($\boldsymbol{\theta}^{(m-1)}$) allowing for coherent exploration of the parameter space (Betancourt, 2017). Empirically, the resulting transitions often prove to result in nearly independent draws if the integration time $\mathcal{T}$ is properly tuned and the kinetic energy is not too poorly chosen.

In practice the application of HMC is however hindered by the fact that we are rarely able to find exact solutions to Hamilton's differential equations and we need to rely on numerical integration schemes to approximate the trajectories. Proofs, of the validity of the HMC transition kernel all depend critically on a property of Hamilton's equations known as the conservation of the Hamiltonian, which is violated as a result of the numerical approximations. The conservation of the Hamiltonian means $H(\boldsymbol{\theta}^{(m-1)}, \boldsymbol{\gamma}) = H(\boldsymbol{\theta}^*, \boldsymbol{\gamma}^*)$ for any $\mathcal{T}$, which follows from observing the change of the Hamiltonian with respect to time

$$\frac{\mathrm{d}H}{\mathrm{d}\tau} = \frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}\tau}\frac{\partial H}{\partial \boldsymbol{\theta}} + \frac{\mathrm{d}\boldsymbol{\gamma}}{\mathrm{d}\tau}\frac{\partial H}{\partial \boldsymbol{\gamma}} = \frac{\partial H}{\partial \boldsymbol{\gamma}}\frac{\partial H}{\partial \boldsymbol{\theta}} + \frac{\partial H}{\partial \boldsymbol{\theta}}\frac{\partial H}{\partial \boldsymbol{\gamma}} = 0,$$

where the second equality follows from (9). The joint probability of $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ hence remains unchanged if Hamilton's equations could be integrated exactly.

To correct for the potential of lower joint probabilities resulting from the numerical integration, a MH acceptance-rejection step is used. The probability of accepting $\boldsymbol{\theta}^*$, generated by means of a numerical integration of Hamilton's equations is determined as

$$\alpha(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(m-1)}) = \min(1, \exp(H(\boldsymbol{\gamma}, \boldsymbol{\theta}^{(m-1)}) - H(\boldsymbol{\gamma}^*, \boldsymbol{\theta}^*))). \tag{10}$$

The common choice for the numerical integrator is the Leapfrog integration scheme, which, considering we wish to integrate for one discretized time interval (i.e. time τ to $\tau + \epsilon$), looks as follows:

$$\begin{aligned}
\boldsymbol{\gamma}(\tau + \epsilon/2) &= \boldsymbol{\gamma}(\tau) - (\epsilon/2)\frac{\partial}{\partial \boldsymbol{\theta}}\log k(\boldsymbol{\theta}(\tau)), \\
\boldsymbol{\theta}(\tau + \epsilon) &= \boldsymbol{\theta}(\tau) + \epsilon \boldsymbol{M}^{-1}\boldsymbol{\gamma}(\tau + \epsilon/2), \\
\boldsymbol{\gamma}(\tau + \epsilon) &= \boldsymbol{\gamma}(\tau + \epsilon/2) - (\epsilon/2)\frac{\partial}{\partial \boldsymbol{\theta}}\log k(\boldsymbol{\theta}(\tau + \epsilon)).
\end{aligned}$$

This scheme is applied for L steps so that $\epsilon L = \mathcal{T}$.

The tuning of the step size ϵ and the number of Leapfrog steps L is guided by the fact that their product, the integration time $\mathcal{T}$, should be such that the autocorrelation in the chain is as low as it can be, while at the same time targeting a theoretically optimal acceptance rate (10) between 0.6 and 0.8 (Betancourt et al., 2014). A favorable property of the Leapfrog integration scheme is that for relatively well behaved posteriors and a step size sufficiently small to produce stable trajectories, the approximation error does not increase with the number of Leapfrog steps L, and depends only on the step size ϵ (Leimkuhler & Reich, 2004). A straightforward strategy of finding the right ϵ and L is therefore to first fix ϵ to some safe value that produces high acceptance rates and then increase L to the point that the resulting chain no longer improves in terms of autocorrelation. Next, increase ϵ while simultaneously lowering L keeping the integration time $\mathcal{T}$ constant, until the acceptance rate is in the desired range.

The target acceptance rates optimally balance the number of costly gradient evaluations required for the Leapfrog integration, with the efficiency of the draws. The gradient computations typically constitute the main computational costs for the HMC algorithm. This is particularly true for GAS models where, just as the likelihood is expensive to evaluate due to the recursive formulation of the time-varying parameters, the gradient is expensive to evaluate as well - and typically even more so as the likelihood itself.

The final tuning parameter in HMC is the mass matrix $\boldsymbol{M}$. For a posterior distribution that is relatively well-behaved, meaning it is roughly elliptically shaped like a Gaussian, the mass matrix is optimized by setting it to the inverse of the covariance matrix of $\boldsymbol{\theta}$ under the posterior (6). The reason for this is that the space of the momenta variable can be regarded as dual to the parameter space, so when the parameter space is endowed with some Euclidean structure this suggest that the momenta space should be induced with an inverse Euclidean structure, which is naturally achieved by specifying a Gaussian with covariance equal to the inverse of the posterior covariance for the momenta (Betancourt et al., 2017). The effect of setting the mass matrix in this way is comparable to the use of the posterior covariance for the scale matrix in the proposal density of a RW-MH sampler, since both effectively rotate and rescale the parameter space so that the parameters are roughly posteriorly uncorrelated and identically scaled (see Neal (2011)).

The mass matrix is estimated during the warm up period. As an example, I do this in the application for Section 4.1 as follows: for the first 50 draws $\boldsymbol{M}$ is set to the identity matrix; at that point the last 40 of these are used to estimate the mass matrix; the mass matrix is then re-estimated once using draws 50 to 100 to obtain a sufficiently accurate approximation of the true posterior covariance. For models with a much larger number of parameters longer warm ups are obviously required and more than one re-estimation of the mass matrix can be used. The step size and number of Leapfrog steps need to be tuned after each re-estimation of the mass matrix. This procedure proved sufficient for the models considered in Section 4. In certain cases, for instance if much posterior mass is located close to a constraint or if there is much variation in curvature, the use of a dense mass matrix can be detrimental to HMC's performance. In such cases I restrict $\boldsymbol{M}^{-1}$ to a diagonal matrix with the estimated posterior variances of $\boldsymbol{\theta}$ on the diagonal.

GAS models often require the enforcements of parameter constraints. MH samplers usually impose constraints through priors that are zero on the domain of parameter values that violate the constraints. Meaning that proposals that fall outside the constrained region are simply rejected, essentially resulting in a draw wasted. In the case of simple upper and lower bounds, HMC offers a more efficient alternative approach described in Neal (2011, Sec 5.1) in which the trajectory of a parameter is reversed as soon as a Leapfrog step results in the crossing of a bound. Consider for example a parameter θ_i, which violates a lower bound lb. The reversal is achieved by resetting $\theta_i = lb + (lb - \theta_i)$ and negating the momentum variable $\gamma_i = -\gamma_i$.

HMC is known to have trouble exploring the tails of a distribution if the posterior is heavy-tailed. In order to properly explore the tail regions of such heavy-tailed posteriors, might require unreasonably long integration times (Betancourt, 2017). For the applications considered in Section 4, I did not find the posteriors of GAS-models to be sufficiently heavy-tailed to significantly limit HMC, unless the model parameters were not all well identified. For example in the unrestricted multivariate covariance model analyzed in Section 4.3, identification issues caused extremely high posterior kurtosis for several of the parameters and relatively long travel times proved necessary to properly explore the posterior.

3.2 Bayesian Model Comparison

Next I discuss two popular Bayesian techniques that facilitate inferences that are generally not available in a frequentist setting or would at the least be considerably more difficult. The focus will be on highlighting why these approaches are relevant in particular for GAS models. First the Bayesian model comparison approach is introduced. Bayesian model comparison is of interest for GAS model inference as unlike most frequentist model comparison techniques, it allows for comparison of non-nested models. As discussed in Section 2 the different choices of link functions, scaling matrices and factor structures imply that non-nested models for which comparison is likely desired, are very common in the GAS framework .

Consider the case where we are interested in comparing the two models $\mathcal{M}_1$ and $\mathcal{M}_0$ and we assume them to be a priori equally likely to be correct. The posterior odds ratio of these models is known as the Bayes factor ($BF_{1|0}$) and is defined as

$$BF_{1|0} = \frac{p(\boldsymbol{Y}_t|\mathcal{M}_1)}{p(\boldsymbol{Y}_t|\mathcal{M}_0)}, \tag{11}$$

where $p(\boldsymbol{Y}_t|\mathcal{M}_i)$ is known as the marginal likelihood of model i. In the Bayesian model comparison framework, the Bayes factor is the basis upon which to formulate conclusions regarding the strength of evidence in favor of model 1 relative to model 0 - or in favor of model 0 relative to model 1 depending on the interpretation.

As can be derived from (11), the marginal model likelihoods are critical inputs, but they are typically challenging quantities to obtain. They are defined as

$$p(\boldsymbol{Y}_T|\mathcal{M}_i) = \int p(\boldsymbol{Y}_T|\boldsymbol{\theta}_i, \mathcal{M}_i)p(\boldsymbol{\theta}_i)\mathrm{d}\boldsymbol{\theta}_i. \tag{12}$$

where $p(\boldsymbol{Y}_T|\boldsymbol{\theta}_i, \mathcal{M}_i)$ is the likelihood of the data under model i. The expression under the integral sign is simply the unnormalized posterior for model i, which will again be denoted with $k(\boldsymbol{\theta}_i)$. It is important to note that the prior for $\boldsymbol{\theta}_i$ must be proper for the parameters that are not shared between model 1 and model 0. By integrating over the parameters $\boldsymbol{\theta}$, the log marginal likelihood reflects the full extent of parameter uncertainty. Explicit analytical expressions for the marginal likelihood will generally not be obtainable for GAS models and efficient ways of approximating or estimating them is an area of active research and debate. With the improvements in computing power, estimations of the marginal likelihood using MCMC samples has become increasingly popular relative to crude approximations such as the Bayesian information criterion (BIC, see Kass & Raftery (1995) for a treatment of several of such approximating identities). I choose to use one of such simulation approaches.

3.2.1 Bridge Sampling for Marginal Likelihood Estimation

Throughout this paper I use the bridge sampling method for marginal likelihood estimation introduced by Meng & Wong (1996). Bridge sampling makes use of an importance or candidate density $q(\boldsymbol{\theta}_i)$ that should reasonably approximate the posterior $p(\boldsymbol{\theta}_i|\boldsymbol{Y}_T)$ and a so-called bridge function $h(\boldsymbol{\theta}_i)$. The key identity in the bridge sampling method is

$$p(\boldsymbol{Y}_T|\mathcal{M}_i) = \frac{\mathrm{E}_q(h(\boldsymbol{\theta}_i)k(\boldsymbol{\theta}_i))}{\mathrm{E}_p(h(\boldsymbol{\theta}_i)q(\boldsymbol{\theta}_i))}, \tag{13}$$

where $\mathrm{E}_q(\cdot)$ and $\mathrm{E}_p(\cdot)$ denote the expectations with respect to the candidate density $q(\boldsymbol{\theta}_i)$ and the proper posterior density $p(\boldsymbol{\theta}_i|\boldsymbol{Y}_T)$ respectively. By simulating a sample from both the candidate density and the posterior density we can use the identity 13) to construct a simple Monte Carlo estimator of the marginal likelihood for model i as

$$\hat{p}(\boldsymbol{Y}_T|\mathcal{M}_i) = \frac{\frac{1}{M_q}\sum_{m_q=1}^{M_q} h(\boldsymbol{\theta}_i^{(m_q)})k(\boldsymbol{\theta}_i^{(m_q)})}{\frac{1}{M_p}\sum_{m_p=1}^{M_p} h(\boldsymbol{\theta}_i^{(m_p)})q(\boldsymbol{\theta}_i^{(m_p)})}, \tag{14}$$

where M_q and M_p are the sample sizes from the candidate and posterior respectively. I follow standard procedure and use equal sample sizes from both densities. The sample from the posterior is typically generated using a MCMC algorithm such as one of the algorithms described in Section 3.1.

For the bridge function $h(\cdot)$, Meng & Wong (1996) present an optimal form, which minimizes the relative mean squared error of the marginal likelihood estimator $\hat{p}(\boldsymbol{Y}_T|\mathcal{M}_i)$. It is defined up to a constant as

$$h(\boldsymbol{\theta}_i) \propto \frac{1}{M_p q(\boldsymbol{\theta}_i) + M_q p(\boldsymbol{\theta}_i|\boldsymbol{Y}_T)}.$$

The challenge with applying this optimal bridge function comes from the fact that we need to know the normalized posterior $p(\boldsymbol{\theta}_i|\boldsymbol{Y}_T)$ to compute it and to obtain this we need its integrating constant, which is the marginal likelihood $p(\boldsymbol{Y}_T|\mathcal{M}_i)$ - the exact quantity we are trying to estimate using bridge sampling. The implementation of the bridge sampling estimator with optimal bridge function, therefore requires an iterative estimation procedure for the marginal likelihood $\hat{p}(\boldsymbol{Y}_T|\mathcal{M}_i)$, which is obtained as the limit of $\hat{p}^{(j)}(\boldsymbol{Y}_T|\mathcal{M}_i)$ for $j = 1, 2, ...$, where the j+1-th iteration is computed using the following estimate of the normalized posterior

$$\hat{p}(\boldsymbol{\theta}_i|\boldsymbol{Y}_T) = \frac{k(\boldsymbol{\theta}_i)}{\hat{p}^{(j)}(\boldsymbol{Y}_T|\mathcal{M}_i)}.$$

The iterative scheme can be initialized with $\hat{p}(\boldsymbol{\theta}_i|\boldsymbol{Y}_T)$ set to $k(\boldsymbol{\theta}_i)$. The recursion is then given by

$$\hat{p}^{(j+1)}(\boldsymbol{Y}_T|\mathcal{M}_i) = \frac{\frac{1}{M_q}\sum_{m_q=1}^{M_q} \frac{k(\boldsymbol{\theta}_i^{(m_q)})}{M_p q(\boldsymbol{\theta}_i^{(m_q)}) + M_q \hat{p}(\boldsymbol{\theta}_i^{(m_q)}|\boldsymbol{Y}_T)}}{\frac{1}{M_p}\sum_{m_p=1}^{M_p} \frac{q(\boldsymbol{\theta}_i^{(m_p)})}{M_p q(\boldsymbol{\theta}_i^{(m_p)}) + M_q \hat{p}(\boldsymbol{\theta}_i^{(m_p)}|\boldsymbol{Y}_T)}},$$

The estimator typically converges in under 10 iterations and the method can therefore be implemented relatively fast. For GAS models the main computational costs come from the M_q evaluations of the unnormalized posterior of model i, which only need to be performed once.

The main advantage of using bridge sampling over most alternative marginal likelihood estimators is that the choice of candidate density $q(\cdot)$ is a slightly less precarious one. Since the estimator (14) uses draws from both the candidate and the posterior, the variance of the estimator is not overly sensitive to one of the densities being more heavy tailed as the other. Alternative sampling methods often impose requirements on the tail behavior of the candidate density relative to that of the posterior - such as the importance sampling estimator requiring fatter-tailed candidate densities or reciprocal importance sampling requiring a thinner tailed candidate - in order for the variance of the estimator to be finite (Frühwirth-Schnatter, 2004). The bridge sampling estimator has finite variance regardless of the tail behavior of the candidate. This is an important advantage that I find particularly relevant for its applicability to GAS models. For GAS models the tail behavior

often varies significantly among the marginal posteriors of $\boldsymbol{\theta}$ (i.e. certain sets of parameters in $\boldsymbol{\theta}$ can differ greatly in terms of their posterior kurtosis). Since candidate densities are generally constructed with similar tail behavior in all dimensions, we would have to adjust the tail behavior to the most extreme elements in $\boldsymbol{\theta}$. Experience showed however that for instance taking the approach of constructing a candidate density - even using a relatively advanced method such as the AdMit procedure with $\nu = 1$ to be on the safe side - typically did not result in a good fit to the posterior (acceptance rates in independence chain MH of merely 0.25). Not having a good fitting candidate density also greatly inflates the variance of the marginal likelihood estimator.

Using bridge sampling there are thus no requirements on the tail behavior of the density and we can therefore adopt the common approach of simply using a multivariate normal density. The mean and covariance are set equal to the sample estimates of the posterior means and covariance matrix of $\boldsymbol{\theta}$ based on an additional independent sample of M_p draws from the posterior. The separate independent sample is necessary to avoid possible downward bias in the marginal likelihood estimate Gronau et al. (2017). Given that a good fit is still important for the effectiveness of the estimator and the relatively high potential for non-normalities in GAS model posteriors, I also utilize the so called warp 3 transformation described in Meng & Schilling (2002) to account for skewness. This is done by constructing a mixture of the unnormalized posterior

$$\tilde{k}(\boldsymbol{\theta}_i) = \frac{1}{2}(k(\boldsymbol{\theta}_i) + k(2\boldsymbol{\theta}_{i,0} - \boldsymbol{\theta}_i)),$$

.
which is symmetric around $\boldsymbol{\theta}_{i,0}$ and where $\boldsymbol{\theta}_{i,0}$ is typically set equal to the posterior sample mean. It is straightforward to see that $\tilde{k}(\boldsymbol{\theta}_i)$ has the same integrating constant as $k(\boldsymbol{\theta}_i)$, implying that $\tilde{k}(\boldsymbol{\theta}_i)$ can also be used in (14) to estimate the marginal likelihood. The resulting estimator does require twice as many evaluations of the likelihood and is thereby roughly twice as computationally expensive. In section 4.2 and 4.3 I implement the bridge sampling estimator with warp 3 transformation as described above using the R package "bridgesampling" by Gronau et al. (2017).

3.2.2 Prior Sensitivity

When applying Bayesian model comparison it is important to be aware of one major criticism of the approach, which is its sensitivity to the prior specification for $\boldsymbol{\theta}_i$. Since the prior in (12) must be proper for non shared parameters, it is not possible to use overly diffuse or uninformative priors as the excessive probability mass placed on very unlikely values of $\boldsymbol{\theta}_i$ will bias the Bayes factor to favor the model with less parameters (Kass & Raftery, 1995). In addition to thus requiring reasonably informative priors, the influence of the prior is also known to diminish less rapidly with the number of observations as it does for common posterior inferences such as posterior means or confidence intervals for the parameters (Kass, 1993).

So, although GAS models are typically used for applications with considerably large samples, we should be particularly careful with the prior specification when it comes to model comparison. To assess the impact of prior specifications on the conclusions implied by Bayes factors, I consider the approach suggested in Kass & Raftery (1995) to perform a sensitivity analysis with respect to the degree of informativeness of the priors. This requires that Bayes factors are computed for a selection of different prior specification with varying degrees of informativeness and compared for meaningful differences. I apply such a sensitivity analysis in Section 4.2, where Bayesian model comparison is central to the problem considered.

Some final notes on Bayesian model comparison regard terminology and the BIC as an approximation to the Bayes factor. Bayesian model comparison is often referred to as the Bayesian approach to hypothesis testing. The compared models are therefore also often labelled as the alternative model or hypothesis ($\mathcal{M}_1$) and the null model or hypothesis ($\mathcal{M}_0$) (Kass & Raftery, 1995). However, instead of accepting or rejecting hypothesis at certain levels of significance, it is more natural in a Bayesian framework to make statements regarding the strength of evidence in favor of certain models or hypotheses. A common quantity to base such statements regarding the strength of evidence on is $2 \log \mathrm{BF}_{1|0}$. This quantity is of the same scale as the likelihood-ratio statistic popular in ML settings and also as the ΔBIC statistic discussed in Carlin & Louis (2000), which is simply the difference in the BIC for model 1 and model 0. The BIC for model i is defined as

$$\mathrm{BIC}_i = -2 \log p(\boldsymbol{Y}_T | \boldsymbol{\theta}_{i,ML}, \mathcal{M}_i) + d_i \log(T), \tag{15}$$

where d_i is the number of parameters in model i and $\boldsymbol{\theta}_{i,ML}$ is the ML estimate of $\boldsymbol{\theta}_i$ (i.e. the parameter value for which the likelihood for model i obtains its optimum). The BIC is thus equal to double the negative of the maximal log likelihood plus a penalty term for the number of parameters. The BIC is an approximation to the negative of double the log marginal likelihood and as $T \to \infty$ the two should converge. In most cases, when the likelihood is not strongly peaked and symmetric, the approximation is however rather crude and often displays specific biases relative to the true log marginal likelihood (see e.g. Miazhynskaia & Dorffner (2006)). Since the BIC is a popular tool for informal non-nested model comparison in an ML setting, I use both Bayes factors and the ΔBIC statistic for comparison in Section 4.2.

3.3 Hierarchical Priors

Bayesian model comparison thus allows for a coherent approach for testing hypothesis or making model selection decisions among non-nested models, while accounting for the full extent of parameter uncertainty. Hierarchical modeling is an approach to alleviating parametric uncertainty, typically achieved in a Bayesian framework by imposing hierarchical priors on a subset of the parameter in $\boldsymbol{\theta}$ that share certain characteristics. The hierarchical prior allows the information in the data regarding certain parameters to be shared across the parameters in the group to which they belong. The approach is particularly powerful in more complex models where the parameter space is of such a dimension that the data is insufficient to properly identify all parameters.

Hierarchical prior specifications extend the typical Bayesian setup (6) presented in Section 3, by specifying an additional layer of prior distributions for the static parameters of the prior distribution of $\boldsymbol{\theta}$. Under a hierarchical prior the posterior is characterized as

$$p(\boldsymbol{\theta}, \boldsymbol{\delta}) \propto p(\boldsymbol{Y}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{\delta}) p(\boldsymbol{\delta}),$$

where $\boldsymbol{\delta}$ is a vector of parameters that governs the prior distribution of $\boldsymbol{\theta}$. The parameters in $\boldsymbol{\delta}$ are known as hyperparameters.

Currently hierarchical modeling is mostly applied for generalized linear models in the form of random effects and mixed effects models, and also spike-and-slab regression relies on a hierarchical prior setup. Applications to non-linear models such as time-varying parameter models are notably more scarce (one example is Brownlees (2015)), most likely as a result of the substantial computational costs associated with models of large numbers of time-varying parameters. Regardless, there

seems considerable potential for natural extensions of hierarchical modeling techniques for generalized linear models to enable inference in time-varying parameter models of greater complexity than previously attainable.

Although conjugate hierarchical priors that allow for analytical expression of the posterior will likely not be attainable for GAS models, the basic motives for grouping sets of parameters that we a priori know to have common features are still valid. The benefits of pooling information to reduce parameter uncertainty are also especially relevant for GAS models where significant parameter uncertainty arises relatively quickly. Advancements in computational power of desktop computers, combined with relatively recent MCMC techniques such as HMC that allow for efficient sampling in high dimensional parameter spaces, imply that inference in hierarchical models is possible even when the posterior is not known in closed-form (Betancourt & Girolami, 2015).

A popular application of GAS models is to model time-varying covariance matrices and these models serve as a good example of models where the number of time-varying parameters might quickly cause the number of parameters in $\boldsymbol{\theta}$ to exceed that what the data can support. Common resolutions to this problem are to either simply consider only very small time-varying covariance matrix models or to impose parametric restrictions. In Section 4.3 I show how applying a hierarchical normal prior for the subset of autoregressive parameter that govern the correlation dynamics - similar to the priors used for the regression coefficients in random effects models - to provide a more elegant solution to the parameter uncertainty problem that occurs for time-varying covariance models. Comparison by means of Bayes factor also support the hierarchical model over the restricted (and unrestricted) version, particularly as the number of assets under consideration increases. In the discussion I also consider several possible other applications of hierarchical priors for models in the GAS framework.

4 Empirical Applications

In this section I apply the above described Bayesian methods for inference to three different types of GAS models. Although the first serves mostly as a test case to compare the three MCMC methods described in Section 3.1, all three applications illustrate the benefits that Bayesian methods offer in handling the uncertainty associated with implementing GAS models. Besides the comparison of MCMC methods, the first application also highlights how the considerable model complexity, typical of highly non-linear GAS models, affect parameter uncertainty and the associated slow convergence to a normal distribution for the static parameters. The second application uses Bayesian model comparison to account for parameter uncertainty in a non-nested model selection problem prominent in many applications for GAS models due to the multitude of different model specification options described in Section 2. Finally, the third application explores Bayesian hierarchical modeling as a promising approach to dealing with the substantial parameter uncertainty in more complex models. The hierarchical modeling approach is applied to enable inference on a flexible time-varying covariance model of a large number of assets.

4.1 MCMC Method Comparisons: Beta-Gen-t-EGARCH

To illustrate the relative strengths and weaknesses of the three MCMC methods Griddy Gibbs, AdMit-MH and HMC, I apply all three to the Beta-Gen-t-EGARCH by Harvey & Lange (2017)

estimated on daily returns data from the S&P 500 Index for the period 2012-04-16 to 2017-04-21 (source: <https://finance.yahoo.com/quote/%5EGSPC/history?p=%5EGSPC>, see Figure 1).

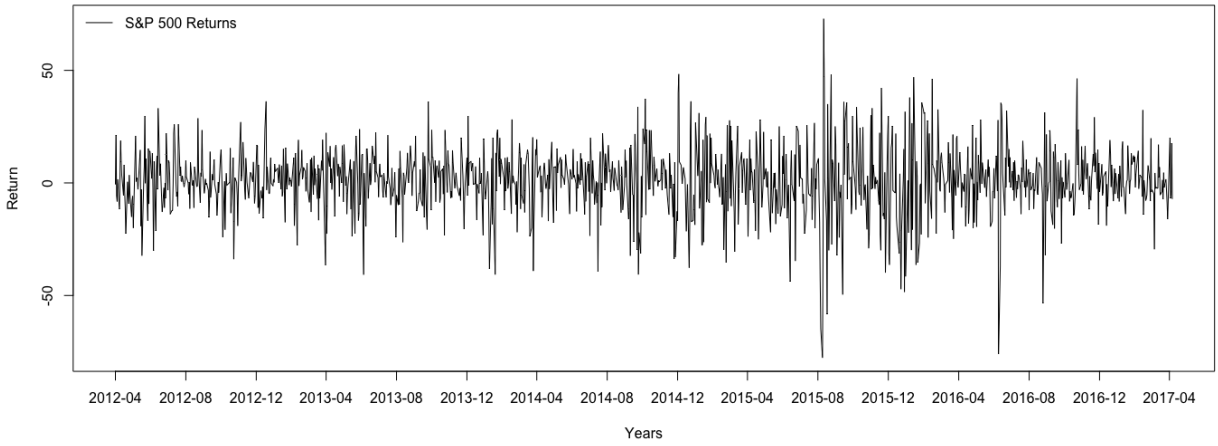


Figure 1: S&P 500 daily returns for the period 2012-04-16 to 2017-04-21.

4.1.1 The Beta-Gen-t-EGARCH Model

The Beta-Gen-t-EGARCH model specifies a generalized Student-t distribution[4] for a univariate dependent variable y_t with time-varying scale parameter φ_t

$$p(y_t|\mu, \varphi_t, \bar{\eta}, \upsilon) = \frac{1}{\varphi_t} \frac{\upsilon \bar{\eta}^{1/\upsilon}}{2\mathrm{B}\left(\frac{1}{\upsilon\bar{\eta}}, \frac{1}{\upsilon}\right)} \left(1 + \bar{\eta} \left| \frac{y_t - \mu}{\varphi_t} \right| \right)^{-\frac{\bar{\eta}+1}{\upsilon\bar{\eta}}}, \tag{16}$$
$$\varphi_t = e^{f_t},$$

where f_t follows the time-varying parameter process (2) and $\mathrm{B}(\cdot, \cdot)$ denotes the beta function. In the Beta-Gen-t-EGARCH model the scaling matrix (4) is set to the identity matrix and thus s_t simply equals the score for f_t

$$\nabla_t = \frac{\bar{\eta} + 1}{\bar{\eta}} b_t - 1, \tag{17}$$

$$b_t = \frac{\bar{\eta}(|y_t - \mu| e^{-f_t})^{\upsilon}}{\bar{\eta}(|y_t - \mu| e^{-f_t})^{\upsilon} + 1}, \tag{18}$$

where b_t is distributed $\mathrm{Beta}(1/\upsilon, 1/\upsilon\bar{\eta})$.

The generalized Student-t distribution has two shape parameters, $\bar{\eta}$ which controls the tails of the distribution, and υ which controls the peak of the distribution. These two shape parameters make the Beta-Gen-t-EGARCH model remarkably versatile and allow it to generalize the Beta-t-EGARCH and Gamma-GED-EGARCH models (Harvey, 2010). In case $\upsilon = 2$ the generalized

[4] In Harvey & Lange (2017) the generalized Student-t distribution is introduced with the parameterization $\eta = 1/\bar{\eta}$, but I consider only the parameterization with $\bar{\eta}$ as it greatly simplifies Bayesian inference and prior specification. See Bauwens & Lubrano (1998) for the need for informative priors on the degrees of freedom parameters to maintain integrability of the posterior (6).

Student-t distribution reduces to a Student-t distribution with $1/\bar{\eta}$ degrees of freedom and in the limiting case where $\bar{\eta} \to 0$ the distribution collapses to the generalized error distribution (GED). Since Harvey & Lange (2017) find that the Beta-Gen-t-EGARCH significantly outperforms the more restrictive Beta-t-EGARCH, the popular Student-t based volatility models are arguably too restrictive meaning that the extra flexibility afforded by the additional shape parameter is needed to capture the complex distributions of financial asset returns. The additional parameter and the high degree of nonlinearities inherent in the log-likelihood of the Beta-Gen-t-EGARCH model, do imply that asymptotic convergence to the normal density for the parameters $\boldsymbol{\theta} = (\omega, A, B, \mu, \bar{\eta}, \upsilon)'$ likely require considerably large samples. The analysis in this section suggest that the 5 years of daily data (1256 observations) used here is insufficient. Although Harvey & Lange (2017) use closer to 10 years of daily data, it would probably still seem quite reasonable to expect the 1256 observations to be adequate, considering that - based on a study of the small sample properties of ML estimates of ARCH type models by Hwang & Valls Pereira (2006) - the suggestion for a standard GARCH model is to use a minimum of 500 observations.

I impose priors on the parameters that are as uninformative as possible while imposing the stationarity constraint that $|B| < 1$ and making sure that the variance of y_t remains bounded to enable the volatility analysis in Section 4.1.3. The bounded variance constraint requires that $\bar{\eta} < 1/2$. This thus implies the following improper flat priors

$$\begin{aligned} p(\omega) &\propto 1, \\ p(A) &\propto 1, \\ p(B) &\propto I_{[-1<B<1]}, \\ p(\bar{\eta}) &\propto I_{[0<\bar{\eta}<1/2]}, \\ p(\upsilon) &\propto I_{[0<\upsilon<\infty]}, \end{aligned}$$

where $\mathrm{I}_{[\cdot]}$ denotes the indicator function, which equals one if the parameter in brackets is in the specified range and is zero otherwise. Note that due to the parameterization with $\bar{\eta}$ - which can be likened to the inverse of the degrees of freedom parameter for a regular Student-t distribution - the issues with flat priors on the full support for degrees of freedom parameters mentioned in Bauwens & Lubrano (1998) are avoided. The uninformative priors imply that the kernel of the posterior (6) simplifies to the likelihood of the Beta-Gen-t-EGARCH on the intervals for which the priors are not zero.

4.1.2 Comparing MCMC Methods

For the Griddy Gibbs sampler, the high correlation between ω and B in the Beta-Gen-t-EGARCH model (see the lower-left plot in Figure 2b), requires that the time-varying parameter process is re-parameterized according to the specification (3). To facilitate comparison, the reported results are for the re-transformed variable $\omega = \tilde{\omega}(1 - B)$. Also, the boundaries of the grid, which need to be limited to regions of high posterior mass, imply that for the Griddy Gibbs the priors are in fact flat on the integration region determined by the upper and lower bounds of the grid.

For convenience I determine the upper and lower bounds of the grids as appropriate multiples of the minimum and maximum values of the parameter draws from a preceding run of the AdMit-MH algorithm. I find that 50 grid points resulted in sufficiently smooth histograms for each variable. The AdMit-MH method is implemented using the R package "AdMit" by Ardia et al. (2009). The fitting procedure resulted in a mixture of 2 Student-t distributions. For HMC the

final tuned step size and number of Leapfrog steps after the first 100 warm up draws is 0.5 and 4 respectively. For the first 100 draws I capped the number of Leapfrog steps at 20 and used a step size of 0.005 for the first 50 and 0.1 for the second 50 draws. Derivatives for the log-likelihood of the Beta-Gen-t-EGARCH model as required for HMC, are reported in Appendix A. In the HMC algorithm, the constraints are enforced using the method of Neal (2011) described in Section 3.1.3. The HMC and Griddy Gibbs algorithms as well as the kernel function used as input to the "AdMit" package are implemented in the C language with an R interface on a 2.6 GHz Intel Core i5 processor.

I let all three samplers generate $40,000$ draws after discarding a $1,000$ draw warm up sample. The main results are summarized in Table 1. The estimates of the parameters are quite close for all three methods. For comparison purposes results on ML estimates are also reported. The ML estimates clearly do deviate from the posterior expectations - both in terms of the point estimates and the ML estimates of the standard errors. These differences are largely due to the skewness of the marginal posteriors displayed in the histograms of Figure 2a. Especially for the estimate of $\bar{\eta}$, which the histogram shows is clearly constrained at 0, this leads to biased estimates. The applicability of the asymptotic convergence properties of the ML estimates are therefore questionable given this sample size.

In terms of efficiency, HMC is superior to the other methods for a MCMC chain of this size. The Effective Sample Size (ESS) normalized for computation time of HMC is at least 5 times that of AdMit-MH and roughly 75 times that of the Griddy Gibbs sampler. The Effective Sample Size gives an autocorrelation adjusted estimate of the sample size (number of posterior draws)

$$\text{ESS} = M(1 + 2\Sigma_{j=1}^{\infty}\rho_j(\theta)),$$

where ρ_j is j-th autocorrelation of the draws of a parameter θ. The estimates of the ESS and also the Geweke Convergence Diagnostic (CD) reported in Table 1 are computed using the R package "coda".

For AdMit-MH and HMC the reported computation times include the initial mixture fitting and the estimation of the mass matrix. For HMC these up-front computational cost are negligible but for AdMit-MH they make up roughly 90% of total cost. For longer runs the relative computational efficiency of AdMit-MH would thus improve. In this scenario however, with the numerical standard errors (NSEs) of the posterior means ($\text{NSE}(\theta) = \sqrt{\text{var}(\theta|\boldsymbol{Y}_T)/\text{ESS}(\theta)}$) all well below 0.01, the $40,000$ draws seem more than adequate.

To check whether the Markov chains have converged to the posterior I use the diagnostic of Geweke (1992)

$$\text{Geweke CD} = \frac{\text{E}(\theta_1|\boldsymbol{Y}_T) - \text{E}(\theta_2|\boldsymbol{Y}_T)}{\sqrt{\text{NSE}(\theta_1)^2 + \text{NSE}(\theta_2)^2}},$$

where θ_1 and θ_2 represent draws of the same parameter but of a different fraction of the Markov chain. I use the default setting in the "coda" package which use the first 10% of the full sample as the first fraction and the last 50% as the second fraction. The diagnostic is essentially a z-score and is used to test for a difference in means between the two fractions of the chain. Applying a 5% significance level, all chains apart from the ω and B chains produced by the Admit-MH method seem to have converged. This is likely just due to the high correlation between the two parameters combined with the generally low efficiency per draw of the AdMit-MH method that causes these chains to have the lowest ESS providing insufficient evidence to support the null-hypothesis of convergence. A visual inspection of trace plots suggests that the AdMit-MH algorithm indeed has no

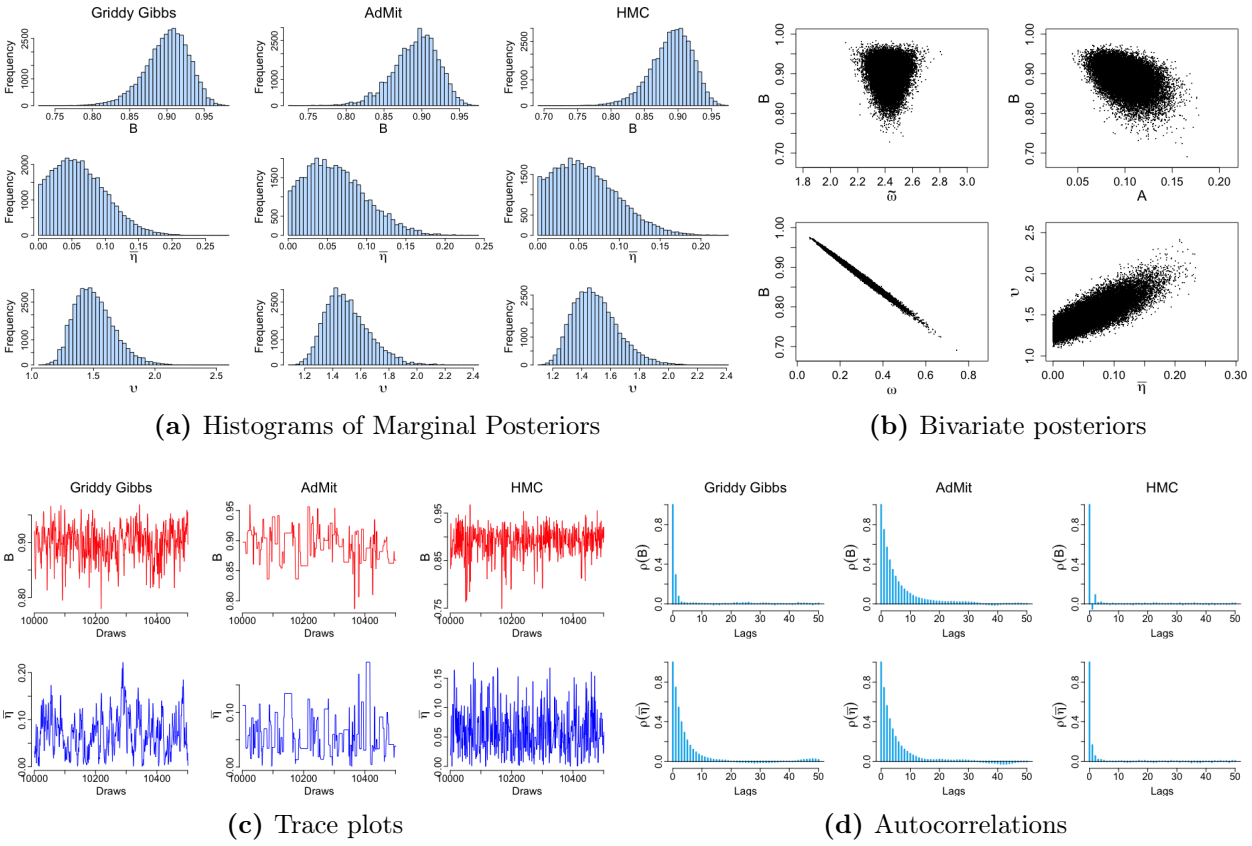
Table 1: Parameter Estimates and MCMC Diagnostics for the Beta-Gen-t-EGARCH Model

		MCMC Methods				
$\boldsymbol{\theta}$		Griddy Gibbs	AdMit-MH	HMC		ML
ω	$\mathrm{E}(\omega\|\boldsymbol{Y}_T)$	0.243	0.264	0.266	$\widehat{\omega}$	0.236
	$\sqrt{\mathrm{var}(\omega\|\boldsymbol{Y}_T)}$	0.073	0.074	0.076	$\sqrt{\widehat{\mathrm{var}}(\omega)}$	0.066
	ESS	21,701	4,967	34,327		
	ESS/time	7.8	84.8	640.5		
	Geweke CD (z-score)	-1.14	-2.10*	0.64		
A	$\mathrm{E}(A\|\boldsymbol{Y}_T)$	0.098	0.100	0.100	$\widehat{A}$	0.093
	$\sqrt{\mathrm{var}(A\|\boldsymbol{Y}_T)}$	0.017	0.016	0.017	$\sqrt{\widehat{\mathrm{var}}(A)}$	0.016
	ESS	19,255	5,183	42,927		
	ESS/time	7.0	88.5	801.0		
	Geweke CD (z-score)	-0.72	0.35	-0.54		
B	$\mathrm{E}(B\|\boldsymbol{Y}_T)$	0.900	0.891	0.890	$\widehat{B}$	0.903
	$\sqrt{\mathrm{var}(B\|\boldsymbol{Y}_T)}$	0.030	0.030	0.031	$\sqrt{\widehat{\mathrm{var}}(B)}$	0.027
	ESS	21,734	4,951	34,044		
	ESS/time	7.9	84.5	635.2		
	Geweke CD (z-score)	1.09	2.00*	-0.54		
μ	$\mathrm{E}(\mu\|\boldsymbol{Y}_T)$	1.189	1.189	1.186	$\widehat{\mu}$	1.146
	$\sqrt{\mathrm{var}(\mu\|\boldsymbol{Y}_T)}$	0.333	0.329	0.329	$\sqrt{\widehat{\mathrm{var}}(\mu)}$	0.345
	ESS	32,064	5,798	49,677		
	ESS/time	11.6	99.0	926.9		
	Geweke CD (z-score)	0.82	1.54	-0.26		
$\bar{\eta}$	$\mathrm{E}(\bar{\eta}\|\boldsymbol{Y}_T)$	0.063	0.062	0.060	$\widehat{\bar{\eta}}$	0.036
	$\sqrt{\mathrm{var}(\bar{\eta}\|\boldsymbol{Y}_T)}$	0.040	0.038	0.039	$\sqrt{\widehat{\mathrm{var}}(\bar{\eta})}$	0.046
	ESS	6,022	5,434	25,515		
	ESS/time	2.2	92.7	476.1		
	Geweke CD (z-score)	0.23	-0.74	0.91		
υ	$\mathrm{E}(\upsilon\|\boldsymbol{Y}_T)$	1.508	1.503	1.498	$\widehat{\upsilon}$	1.423
	$\sqrt{\mathrm{var}(\upsilon\|\boldsymbol{Y}_T)}$	0.163	0.157	0.158	$\sqrt{\widehat{\mathrm{var}}(\upsilon)}$	0.166
	ESS	6,088	5,077	29,427		
	ESS/time	2.2	86.7	462.7		
	Geweke CD (z-score)	0.11	-1.57	1.41		
	Time (s)	2768	59	54		0.4
	Accept rate	_	0.26	0.80		

Notes: Estimation Results for the parameters in the Beta-Gen-t-EGARCH model and Diagnostics for the Markov chains of the three MCMC methods Griddy Gibbs, Adaptive Mixture of Student-t distributions - Metropolis-Hastings (AdMit-MH) and Hamiltonian Monte Carlo (HMC). All three Markov chains are $40,000$ draws long and for all samplers an initial $1,000$ draw warm up sample is discarded. Reported for all three samplers and for all parameters $\omega, A, B, \mu, \bar{\eta}$ and υ are posterior mean ($\mathrm{E}(\cdot|\boldsymbol{Y}_T)$) and standard deviation ($\sqrt{\mathrm{var}(\cdot|\boldsymbol{Y}_T)}$) estimates; the Effective Sample Size (ESS) and the ESS normalized for time in seconds (ESS/time); and the z-score of the Geweke Convergence Diagnostic (CD), where a * is used to denote rejection of the null hypothesis of converged chains at the 5% significance level. Also reported for all parameters are the Maximum Likelihood (ML) point estimates ($\widehat{\cdot}$) and the ML standard deviation estimates ($\sqrt{\widehat{\mathrm{var}}(\cdot)}$). For all MCMC chains the computation time in seconds and the proportion of accepted proposals is reported.

problems exploring the marginal posteriors of ω and B.

Figure 2c presents the trace plots for a subsection of the Markov chains. Clearly, the HMC chains



(a) Histograms of Marginal Posteriors **(b)** Bivariate posteriors

(c) Trace plots **(d)** Autocorrelations

Figure 2: Diagnostic plots of the Markov chains produced by the three MCMC methods Griddy Gibbs, Adaptive Mixture of Student-t distributions - Metropolis-Hastings (AdMit-MH) and Hamiltonian Monte Carlo (HMC). All three Markov chains are 40,000 draws long. The top left corner (a) displays histograms of 50 bins for the marginal posteriors for the parameters $B, \bar{\eta}$ and υ from all three Markov chains. The top right corner (b) contains scatter plots of the draws form the joint posteriors of the parameter pairs B and $\tilde{\omega}$ (from the Griddy Gibbs chain), B and A, B and ω, and υ and $\bar{\eta}$ (all from the HMC chain). In the bottom left (c) are trace plots of 500 draw long subsamples from the chains of all three MCMC methods, for the parameters B and $\bar{\eta}$. Finally, the bottom right (d) shows plots of autocorrelations up to the 50-th lag for all three chains of the parameters B and $\bar{\eta}$.

mix best. The autocorrelation plots in Figure 6d show that for HMC the autocorrelation drops off to zero almost immediately, and for certain parameters the first-order autocorrelation is even negative. This explains the super-efficient sampling of the A and μ parameters - meaning the ESS is greater then the number of draws M. The much higher rejection rate for the AdMit-MH algorithm implies that the chains it produces have much higher autocorrelation. The Griddy-Gibbs suffers when the parameters are highly correlated. This is evidenced by the high autocorrelation for the $\bar{\eta}$ variable, whose correlation with the other shape parameter υ (0.811 in the sample produced by HMC) can not be resolved through re-parameterization. The remarkably high correlation between ω and B of -0.996 does on the other hand seem to be effectively resolved through the re-parameterization (3), as evidenced by the lower and upper plots in Figure 2b. The joint distribution of $\tilde{\omega}$ and B still has an odd shape however.

The histograms and distribution plots of the marginal and bivariate posteriors shown in Figure 2a and 2b, reveal the asymmetries and nonelliptical shapes of the posterior indicating that the 5 years of daily data are inadequate to make the posterior converge to a Gaussian distribution. Excluding the marginal of the μ parameter (not displayed in Figure 2) - which seems to be a partic-

ularly Gaussian and easy parameter to estimate - the average skewness of the marginals in absolute terms is 0.6 and 3.5 for the kurtosis. The posterior is therefore not too heavy-tailed, explaining why HMC is not noticeably limited in its ability to explore the tails of the distribution.

Also note that since the Griddy Gibbs and HMC have higher ESSs relative to AdMit-MH, their chains produce smoother histograms. For the Griddy Gibbs the histogram of υ however has a slightly jagged left side as a result of the finite number of grid points. This is why I choose to use the grid of 50 points, which is high relative to previous work (e.g. Bauwens & Lubrano (1998) and Asai (2006)), but necessary considering the significant correlation between υ and $\bar{\eta}$. Trials with fewer grid points, particularly without the reparameterization (3), produced far more pronounced jagged patterns in the histograms even to the extent that the resulting sample was significantly biased. It is possible that more advanced integration or interpolation rules could improve the performance of Griddy Gibbs and reduce the number of required grid points.

4.1.3 The Posterior Distribution of Volatility

One of the benefits of operating in a Bayesian framework is that it is possible to do inference on essentially arbitrary functions of the parameters. Here I consider the posterior of the volatility of returns on the S&P 500 Index. Although the Beta-Gen-t-EGARCH is designed to model volatility, the variance of the generalized Student-t distribution is a highly nonlinear function of the parameters. Probabilistic statements regarding volatility in a frequentist setting are therefore challenging.

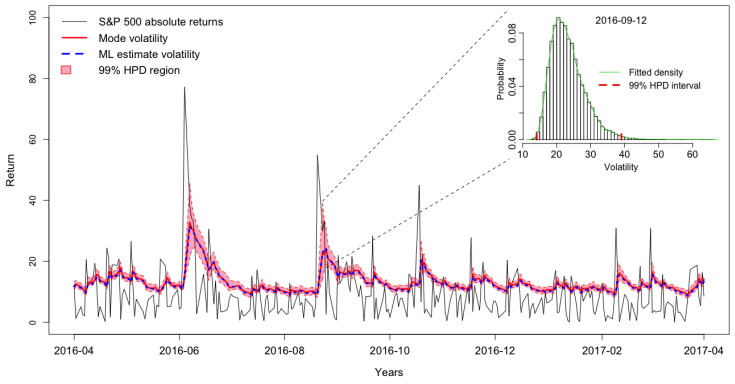
From Harvey & Lange (2017) we have

$$\operatorname{var}(y_t|f_t,\boldsymbol{\theta}) = \frac{\Gamma\left(\frac{3}{\upsilon}\right)\Gamma\left(\frac{1-2\bar{\eta}}{\upsilon\bar{\eta}}\right)}{\Gamma\left(\frac{1}{\upsilon}\right)\Gamma\left(\frac{1}{\upsilon\bar{\eta}}\right)}\frac{1}{\bar{\eta}^{\frac{2}{\upsilon}}}e^{2f_t}, \tag{19}$$

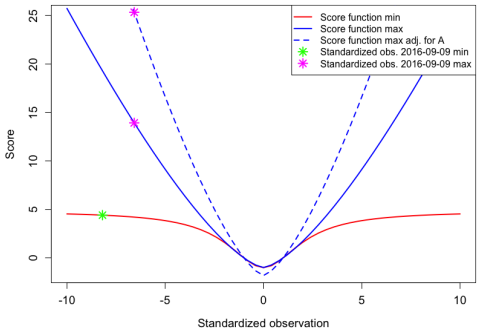
where $\Gamma(\cdot)$ denotes the gamma function[5]. The volatility is then defined as the standard deviation of returns; the square root of (19). Here I consider the Highest Posterior Density (HPD) regions for the volatilities as predicted by the Beta-Gen-t-EGARCH. A HPD region is defined as the set of values that contains $(1-\alpha)100\%$ of the posterior probability mass, for some $\alpha \in (0,1)$. Contrary, to the quantile based confidence intervals that are common in frequentist studies, the HPD region is not necessarily equal-tailed and has the intuitively appealing property that any value outside the region has lower posterior probability as any value inside the region (Gelman et al., 2014, Ch 2.3). The posterior of volatility is easily computed by applying the square root of the formula (19) with draws from the posterior of $\boldsymbol{\theta}$ as input. In this case I use the $40,000$ draws from the chain produced by the HMC algorithm.

Figure 3a depicts the 99% Highest Posterior Density (HPD) region of the volatility as predicted by the Beta-Gen-t-EGARCH, plotted against the absolute value of returns for the last year of the sample period. For comparison a similar plot of volatilities as predicted by the popular Beta-t-EGARCH of Harvey & Chakravarty (2008) - a simpler restricted version of the Beta-Gen-t-EGARCH obtained by fixing υ to 2 - is shown in Figure 3c. The Bayesian analysis is carried out in the same manner and with the same priors as for the corresponding Beta-Gen-t-EGARCH parameters. The HPD regions are particularly noteworthy when the preceding return is abnormally large

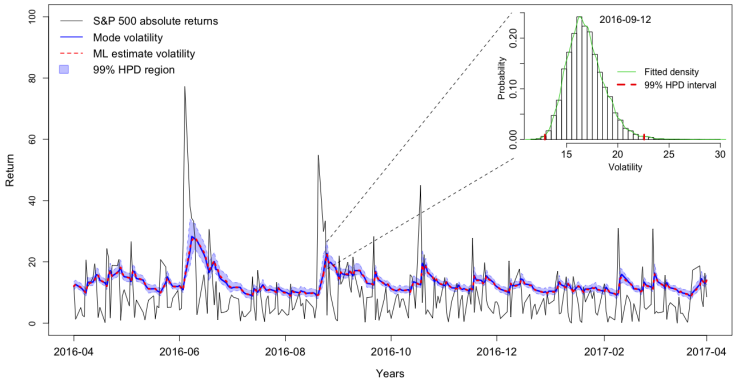
[5]In the limit as $\bar{\eta} \to 0$ numerical evaluation of the gamma function results in overflow. This typically occurred for values of $\bar{\eta} \leq 0.005$ and for these values I compute the variance using the variance formula for the limiting GED, which is $\operatorname{var}(y_t|f_t,\boldsymbol{\theta}) = \Gamma\left(\frac{3}{\upsilon}\right)\upsilon^{\frac{2}{\upsilon}}e^{2f_t}/\Gamma\left(\frac{1}{\upsilon}\right)$



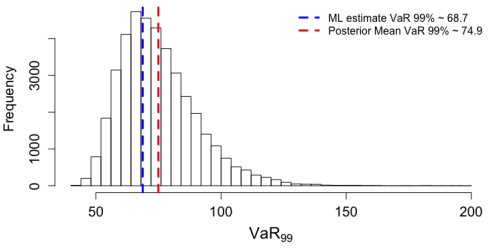
(a) Beta-Gen-t-EGARCH - HPD of predicted volatilities



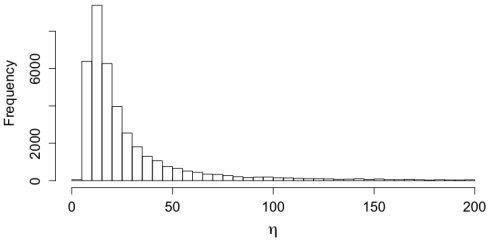
(b) Score functions



(c) Beta-t-EGARCH - HPD of predicted volatilities



(d) One-day-ahead VaR 99% histogram



(e) Degrees of freedom histogram

Figure 3: The 99% highest posterior density (HPD) regions of volatility for the S&P 500 daily returns as predicted by the Beta-Gen-t-EGARCH (top left (a)) and Beta-t-EGACRH (bottom left (c)) plotted in conjunction with absolute returns on the S&P 500, the posterior mode of volatility and the Maximum Likelihood (ML) estimate of volatility for the period 2016-04-21 to 2017-04-21. For illustration, a histogram with fitted density of the volatility on the date 2016-09-12 is also displayed. On the right, other plots for the volatility as predicted by the Beta-Gen-t-EGARCH on the date 2016-09-12 such as the preceding day's score based on the draws corresponding to the minimum and maximum predicted volatilities for the day 2016-09-12 are plotted as a function of the standardized residuals $(y_{t-1} - \mu)e^{-f_{t-1}}$ (b) and a histogram of the posterior of the one-day-ahead 99% value at risk (VaR) (d). The score for the maximum volatility adjusted for A (plotted in b) is defined as a rescaled score obtained by multiplying with the value of A for the maximum draw and dividing by the value of A for the minimum draw to indicate the true relative difference in impact of the score (As_{t-1}) on the next day's f_t. The values for $\bar{\eta}$ and υ are 0.002 and 1.45 for the maximum volatility and 0.21 and 1.97 for the minimum volatility. Also plotted in the bottom right (e) is a histogram of the degrees of freedom η. HPD regions, extrema and histograms are all based on a $40,000$ draw sample from a Hamiltonian Monte Carlo Markov chain.

relative to the predicted level of volatility. The HPD regions for the Beta-Gen-t-EGARCH clearly widen considerably more following such large observations as for the Beta-t-EGARCH. Furthermore, the posteriors of the volatilities as predicted by the Beta-Gen-t-EGARCH are also significantly more positively skewed following large observations, as is evidenced by the position of the mode relative to the 99% HPD region boundaries. As a case in point a histogram of the posterior draws for the predicted volatility on September 12th, 2016 is plotted to illustrate the positive skewness and heavy right tail of predicted volatility (skewness equals 1.1 and kurtosis is 5.0 for the Beta-Gen-t-

EGARCH) following a large negative return of -53 on the S&P500. On the other hand, the average skewness and kurtosis for the volatilities predicted by the the Beta-Gen-t-EGARCH over the entire five year sample (2012-04-16 to 2017-04-21) are only 0.2 and 3.2 respectively, suggesting that the large uncertainty and lack of convergence to a normal distribution is characteristic for the volatilities predicted in response to large absolute returns. For the Beta-t-EGARCH the predicted volatilities following large observations also seem to have larger uncertainty and greater non-normality (see the histogram in Figure 3c), but noticeably less so as for the Beta-Gen-t-EGARCH.

For the Beta-Gen-t-EGARCH, the response to large observations is mostly determined by the shape parameters $\bar{\eta}$ and υ and the large uncertainty in this response can be traced back to the considerable posterior range of the shape parameters. The bottom right plot in Figure 2b shows that the posterior ranges from what is almost a Laplace distribution with $\bar{\eta}$ near zero and υ close to one, to a Student-t distribution with 4 degrees of freedom ($\bar{\eta}$ equal to 0.25 and υ around 2). Both these extremes of the posterior range represent different ways of making the score function of volatility less sensitive to outliers, explaining their high posterior correlation. Typically the effect of a lower degrees of freedom (high $\bar{\eta}$) leads to greater robustness to outliers as the effect of a lower υ. Consequently we find that the lowest values of predicted volatility on dates following extreme observations correspond to draws for which υ is close to two and $\bar{\eta}$ is in the higher end of its posterior range and the highest volatilities correspond to draws for which $\bar{\eta}$ is near zero and υ is closer to one. These effects are evidenced by the plot in Figure 3b where the score functions for the minimum and maximum predicted volatilities (based on the 40,000 HMC draws) and for the date of 2016-09-09 (the Friday preceding 2016-09-12) are displayed. The values of $\bar{\eta}$ and υ corresponding to the plotted score functions are reported in the captions. Clearly the posterior support for the shape parameters encompasses a wide range of score functions or news impact curves, as they are sometimes referred to, which causes the large variation in how strongly the model responds to large absolute returns.

The inverse degrees of freedom parameter therefore seems to be the main determinant of predicted volatility following large observations. To understand the skewness and kurtosis of these predicted volatilities it is more natural to consider the usual - i.e. uninverted - degrees of freedom parameter η, whose posterior range is rather extreme (sampled values range from 4.3 to 8.2×10^5), whilst its mode is around 9.7. Although, as η increases the differences become less and less meaningful, there is still considerable range for which the generalized Student-t distribution transitions to a GED, but where posterior mass is very thinly distributed (see Figure 3e).

The resulting skewness and kurtosis in the posterior of predicted volatilities has significant implications for the application of ML estimates of volatility. As is evident from Figure 3a and 3c the ML estimates for volatility - obtained by plugging in the ML estimates for $\boldsymbol{\theta}$ - are very close to the posterior mode, which is in line with expectation. In practice these ML volatility estimates are often straightforwardly plugged in to compute quantities of interest such as the the VaR or portfolio weights. As demonstrated in Figure 3d, the non-normality of the volatility does carry over into the posterior of the one-day-ahead 99% VaR[6] . The benefits of a Bayesian approach to VaR in which parameter uncertainty can be fully accounted for versus the traditional frequentist plug-in approach is discussed in great detail in e.g. Ardia (2008).

[6] As mentioned in Harvey & Lange (2017), the one-day-ahead VaR can be computed for the Beta-Gen-t-EGARCH by expressing $\varepsilon_t = (y_t - \mu)e^{-f_t}$ as a function of $1 - b_t = \frac{1}{\bar{\eta}\varepsilon_t^{\upsilon}+1}$. Since $1 - b_t$is Beta$(1/\upsilon\bar{\eta}, 1/\upsilon)$ distributed we can use the inverse CDF of a beta distribution to compute quantiles for y_t.

It is important to note that in order to capture the full extent of the uncertainty in the response to extreme observations the uninformative prior on $\bar{\eta}$ is imperative. More than half of the predicted volatilities on 2016-09-12 above the upper bound of the 99% HPD region corresponded to draws for which $\bar{\eta}$ is less than 0.01 (i.e. more than 100 degrees of freedom). Preceding work that applied Bayesian inference methods to volatility models based on Student-t distributions considered the likelihood as parameterized with the regular (non inverted) degrees of freedom parameter and are thereby limited to weakly informative priors that restrict the upper range of the degrees of freedom parameter (see Bauwens & Lubrano (1998)). Even though the intention is usually to be as uninformative as possible, these priors might actually be quite informative as they fully exclude the limiting normal distribution (or the GED in case of the generalized Student-t) and could thereby considerably underestimate the uncertainty in volatility predictions following large observations. The inference presented here on the S&P500 returns suggests that at least for the Beta-Gen-t-EGARCH such a scenario where there is considerable posterior mass for $\bar{\eta}$ close to zero is not necessarily uncommon in practice.

It is however likely that similar uncertainty and high values for the degrees of freedom parameter are less common for simpler Student-t based GAS models such as the GAS-t or Beta-t-EGARCH. Given that for these models there is no second shape parameter that fulfills a similar function in making the model more robust to outliers, the data might be more informative regarding the location of the degrees of freedom parameter. For the Beta-t-EGARCH the range of the degrees of freedom parameter in the 40,000 draw sample is for instance 3.86 to 19.18 and thus much narrower as for the Beta-Gen-t-EGARCH. This also justifies much of the notably smaller and more symmetric HPD regions following large observations shown in Figure 3c relative to Figure 3a.

The analysis presented here shows that using the Beta-Gen-t EGARCH model there is considerably more uncertainty and non-normality in the predicted volatilities following large returns compared to when a simpler model with a single shape parameter such as the Beta-t-EGARCH is used. This is a result of the interactions between the two shape parameters of the generalized Student-t distribution causing significant uncertainty in the marginal posterior of the inverse degrees of freedom parameter. The slower convergence to normality of the volatilities predicted by the Beta-Gen-t-EGARCH suggest that the uncertainty caused by the added shape parameter is not fully accounted for by traditional frequentist methods - at least given the 5 years of return data considered here. The Beta-Gen-t-EGARCH is thus an example of how the model complexity that the GAS framework allows for, might cause larger sample sizes to be required to achieve the degree of convergence to the normal distribution desired in an ML setting.

4.2 Model Comparisons: Dynamic Pooled Marked Point Process Models

Point process models have been gaining traction in financial econometrics in recent decades for their use in modeling time series of events that occur at irregularly spaced time intervals (see e.g. Bauwens & Hautsch (2009) for a review of point process models for high-frequency data). One of point process models applications is to describe the time-varying intensities of credit rating transitions. With this application in mind, Creal et al. (2013) introduced a class of factor GAS models. The models are named dynamic pooled marked point process (which I abbreviate to DPMP) models and represent an easier to estimate observation driven alternative to the parameter driven models with stochastically evolving intensities developed by Koopman et al. (2008) and Bauwens & Hautsch (2006).

In this section the DPMP models are applied to model the intensities of the credit rating transition processes for firms included in Compustats database of Standard & Poor credit ratings. Because Creal et al. (2013) state that no theory for formal comparison of the different DPMP factor models currently exists, there is a very compelling case for considering these models in a Bayesian framework since, as mentioned in Section 3.2, the Bayesian approach to model comparison is not limited to nested models. The focus will therefore be on comparing DPMP models with a variety of factor specifications and I also consider the effects of different scaling matrices. To facilitate comparison with the results obtained in Creal et al. (2013), I stick closely to the factor specifications considered in the work of these authors.

4.2.1 The Model

In order to properly describe DPMP models I need to introduce the counting processes and intensity functions that they describe, both of which are defined in continuous time. Upholding the convention introduced in Section 3.1.3, I use τ to denote continuous time. In addition, we will be interested in the (continuous) time points at which rating class transition events occur, which will be denoted by τ_t for events $t = 1, ..., T$. The index t is thus used to count all events irrespective of event type.

Consider the left-continuous counting processes $N_{ij}(\tau)$, which make unit size jumps at the time an event j occurs for firm i, for event types - or rating class transitions in this applications - $j = 1, ..., k$ and firms $i = 1, ..., l$. The counting processes $N_{ij}(\tau)$ are assumed to be orderly, meaning only one event of type j occurs at time t for firm i (Koopman et al., 2008). In the DPMP models the intensities of these counting processes $\lambda_{ij}(\tau)$ are specified as functions of time. It should be noted however, that the intensities of these processes are naturally only defined at times that firm i is actually at risk of a rating transition of type j occurring. This leads to the following definition of the intensities

$$\lambda_{ij}(\tau) = \lim_{\Delta \downarrow 0} \frac{\Pr(N_{ij}((\tau + \Delta)^-) - N_{ij}(\tau^-) | \mathcal{F}_{\tau^-})}{\Delta},$$

as given in Koopman et al. (2008), where the superscript minus $(\cdot^-)$ is used to denote the time an arbitrary small amount before the point in time to which it is applied (i.e. $\tau^- < \tau$ and $\tau - \tau^-$ is arbitrarily small) and $\mathcal{F}_\tau$ denotes the filtration of the process at time τ, meaning the set of all information up to time τ. The intensity processes $\lambda_{ij}(\tau)$ completely characterize the counting processes $N_{ij}(\tau)$ at time τ (Bauwens & Hautsch, 2006).

Specifying separate dynamics for the l times k intensity processes is generally considered infeasible when modeling credit rating transitions since the number of rating events per firm is normally insufficient. Most datasets typically contain only a few rating events per firm. The DPMP models of Creal et al. (2013) therefore assume the same dynamics for all firms, which is a common approach in the credit risk literature (Koopman et al., 2008). Given this assumption the risk process can be summed across the firm dimension resulting in the k pooled risk processes $N_j(\tau) = \Sigma_{i=1}^{l} N_{ij}(\tau)$, which have intensity processes $\lambda_j(\tau)$.

In the DPMP models the intensity processes are updated only when an event - for any firm and any event type - occurs. Accordingly, it is useful to introduce $\lambda_{j,t}$ to denote the value of the intensity process $\lambda_j(\tau)$ during the time interval $(\tau_{t-1}, \tau_t]$. During this interval, the process behaves as a simple Poisson process with intensity $\lambda_{j,t}$. Letting $\boldsymbol{\lambda}_t = (\lambda_{1,t}, ..., \lambda_{k,t})'$, we can specify the

dynamics for the intensities[7] as

$$\log_\circ(\boldsymbol{\lambda}_t) = \boldsymbol{w} + \boldsymbol{C}\boldsymbol{f}_t, \tag{20}$$

where the operator $\log_\circ$ is used to express the element-wise application of the natural logarithm, $\boldsymbol{w}$ is a vector of baseline log intensities and $\boldsymbol{C}$ a $k \times r$ matrix that maps a set of $r \leq k$ time-varying factors $\boldsymbol{f}_t$ to k dimensions. The matrix $\boldsymbol{C}$ is constructed so that the intensities of the risk processes are driven by particular combinations of the time-varying factors that reflect underlying beliefs regarding the interdependencies between the k competing risk processes. The freedom in the structure of the matrix $\boldsymbol{C}$ and the choice of the number of factors gives this class of models its versatility. However, care has to be taken in that $\boldsymbol{C}$ is constructed in such a way that the model parameters remain identified (see Creal et al. (2011b)) and therefore contains at least a single one per per column, whilst the remaining entries are usually filled with a combination of zeros and parameters that need to be estimated. The time-varying factors follow the process as defined in (2) with $\boldsymbol{A}$ and $\boldsymbol{B}$ restricted to diagonal matrices. In addition, to further ensure model identification, the parameters $\boldsymbol{\omega}$ are set to zero since they interfere with the $\boldsymbol{w}$ parameters.

Karr (1991) showed it is possible to define a likelihood for counting processes based purely on their intensity process. Based on that result and following Koopman et al. (2008), Creal et al. (2013) define the log-likelihood function at event time τ_t for the DPMP models as

$$\ell_t = \boldsymbol{y}_t' \log_\circ(\boldsymbol{\lambda}_t) - (\tau_t - \tau_{t-1})\boldsymbol{K}_t'\boldsymbol{\lambda}_t, \tag{21}$$

where I define $\boldsymbol{y}_t = (y_{1,t}, ..., y_{k,t})$, with $y_{j,t}$ a discrete valued variable equal to the number of firms for which event type j occurred at time τ_t and $\boldsymbol{K}_t = (K_{1,t}, ..., K_{k,t})'$, with $K_{j,t}$ a discrete valued variable equal to the number of firms which are actually subject to the risk of event type j occurring at time t_τ. At the firm level the likelihood is intuitive to interpret. On the one hand, there is the probability of survival (i.e. the probability of the event not occurring) for risk process j if firm i is actually at risk of process j occurring during $(\tau_{t-1}, \tau_t]$ - which is reflected in (21) through the term multiplied by the spell length $(\tau_t - \tau_{t-1})$. On the other hand, in the case that an event of type j does occur for firm i at time τ_t, the probability of survival during the time interval (τ_{t-1}, τ_t) is multiplied by the hazard rate (i.e. the rate or probability of the event occurring) for risk process j (see Koopman et al. (2008)). Note that the latter probability of event j occurring for a certain firm, can alternatively be interpreted as the probability of the time difference between events $(\tau_t - \tau_{t-1})$, which is known to be distributed $\text{Exp}(\lambda_{j,t})$ given that event j occurred at time τ_t.[8] In the log-likelihood (21), the log of all these probabilities are simply pooled (summed) over all firms and all event types.

Given the definition of ℓ_t, the score and the information matrix with respect to the time-varying parameters for the DPMP models are given by

$$\begin{aligned}\boldsymbol{\nabla}_t &= \boldsymbol{C}'\big(\boldsymbol{y}_t - (\tau_t - \tau_{t-1})\boldsymbol{K}_t \odot \boldsymbol{\lambda}_t\big),\\ \boldsymbol{\mathcal{I}}_t &= \boldsymbol{C}'\boldsymbol{P}_t\boldsymbol{C},\end{aligned}$$

where the operator $\odot$ denotes the Hadamard product and $\boldsymbol{P}_t$ is a $k \times k$ diagonal matrix with its j-th element defined as $p_{j,t} = K_{j,t}\lambda_{j,t}/\boldsymbol{K}_t'\boldsymbol{\lambda}_t$, which represents the probability that the event that

[7] The specification used here defers in notation from that used in Creal et al. (2013) in that I specify the dynamics directly for the intensity of the pooled (across firms) process. This is possible since I do not include firm-specific exogenous regressors in the specification for the time-varying intensities, resulting in entirely identical predicted intensities for each firm. The intensity process specifications are thus equivalent to those used by Creal et al. (2013) aside from the fact that no exogenous regressors are used.

[8] This is based on the notion that the risk processes are Poisson in nature in between events and the interarrival times for Poisson processes are exponentially distributed.

occurs at time τ_t is of type j given that the next event occurs at time τ_t. The effect of using an inverse or inverse square root Fisher information matrix as scaling matrix for the time-varying parameter update is thus to amplify the relative impact of the score if the probability of event j being the next to occur is low relative to the other risk processes. The score for the log intensities is negative if at time τ_t the event of type j did not occur for any firm and can be either positive or negative if event type j did occur at time τ_t. Jointly the effect of the score and a scaling matrix based on the information matrix is to amplify the updates of the intensities regardless of whether corresponding event type actually occurs, resulting in a generally more responsive auto-regressive process for the intensities, at times that the probability of event j being the next to occur is low relative to the other risk processes. This effect also shows in the empirical application (see Figure 4). Logically such an update makes sense because low probabilities for a certain event type will generally result from very few events of the type occurring; implying that the events that do occur are more informative relative to the type of events that occur more frequently and should therefore be attributed more weight in the update for the time-varying intensities.

To see that the score $\boldsymbol{\nabla}_t$ is still zero in expectation, we need to recognize that - without conditioning on event j occurring at time τ_t - the interarrival time of the pooled process $(\tau_t - \tau_{t-1})$ is distributed as the minimum of the interarrival times of the k risk processes for all firms at risk at time τ_t and is therefore distributed as $\text{Exp}(\boldsymbol{K}_t'\boldsymbol{\lambda}_t)$. Combining this with the fact that $y_{j,t}$ is in expectation equal to the the probability of event j being the next event to occur summed over all firms, which is just equal to $p_{j,t}$, it follows that the score indeed has expectation zero. These identities related to the minimum of a set of interarrival times can be found in any standard textbook covering point processes (see e.g. Ross (2014)).

Following Creal et al. (2011b, 2013), I group the Standard & Poor credit ratings into two broad groups: investment grade (IG) for corporate bonds rated BBB- or higher and sub-investment grade (SIG) for corporate bonds rated below BBB-. This reduces the number of credit rating transitions to a manageable amount ($k = 4$), as the only possible credit ratings we have to consider are IG $\rightarrow$ SIG, IG $\rightarrow$ DEF, SIG $\rightarrow$ IG and SIG $\rightarrow$ DEF, where DEF refers to a firm defaulting. For the comparison I study three different factor specifications with the following structures for the matrix $\boldsymbol{C}$

$$\begin{bmatrix} C_{1,1} \\ C_{2,1} \\ C_{3,1} \\ 1 \end{bmatrix}, \quad \begin{bmatrix} C_{1,1} & 0 \\ C_{2,1} & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \begin{bmatrix} 1 & 0 & 0 \\ C_{2,1} & 0 & C_{2,3} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

The first corresponds to a model where all rating transitions are driven by one time-varying factor ($r = 1$), the second specifies one factor for downgrades and one factor for upgrades[9] ($r = 2$) and the third specifies one factor for IG downgrades, one factor for upgrades and one factor for downgrades to default ($r = 3$).

For the scaling matrix Creal et al. (2013) use the inverse square root of the Fisher information matrix ($a = 1/2$ in (4)), but in an earlier version of their work (Creal et al., 2011b) they also

[9]The two factor specification is different from the one considered in Creal et al. (2013) where the elements $C_{1,2}$ and $C_{2,2}$ are also treated as parameters to be estimated, making for a total of four free parameters in $\boldsymbol{C}$. This is because the data I consider showed little support for time-variance in the intensity of the upgrade transition SIG $\rightarrow$ IG (see Table 9), causing the second factor to be close to zero throughout the sample period and leading to the $C_{1,2}$ and $C_{2,2}$ parameters to be ill-identified. Hence I apply a more restrictive structure for $\boldsymbol{C}$, whilst still allowing upgrades and downgrades to follow different dynamics.

considered the inverse Fisher information matrix ($a = 1$) as scaling matrix for an equivalent factor DPMP model. Besides testing which of the above factor specification is superior, I also test whether there is a performance difference between the three scaling matrix specifications presented in Section 2 ($a = 0, 1/2, 1$). Although, one would intuitively expect the second-order information to be beneficial, since the different scaling matrices imply non-nested models no preceding work has tested such a hypothesis.

4.2.2 Application to Credit Rating Data and Model Comparisons

The Compustat database of Standard & Poor credit ratings contains the ratings of 5969 companies over the period of 1978-11-01 to 2017-02-01, however the first rating class transition (as defined above) occurred on 1986-01-01. In total 884 transitions from IG $\rightarrow$ SIG, 11 transitions from IG $\rightarrow$ DEF, 745 transitions from SIG $\rightarrow$ IG and 782 transitions from SIG $\rightarrow$ DEF are recorded in the dataset. For the IG to DEF risk process the data is therefore relatively sparse. The rating changes are recorded at the monthly level, but the assumption that the firm specific counting processes are orderly is not violated.

I first discuss the results regarding posterior model probabilities and the conclusions for the model comparisons. Parameter estimation results are part of the analysis in Section 4.2.3. I consider a total of 12 model comparisons: for each of the three factor structures I compare the identity scaling option with the inverse square root scaling and the square root scaling with the inverse scaling, and for each of the scaling matrices I compare the one factor specification with the two factor specification and the one factor specification with the three factor specification. This requires posteriors draws of all nine possible models, which I will refer to using the following naming convention DPMPf-S for f equal to 1, 2, 3 factors and scaling matrix S equal to I for the identity matrix, H for the inverse square root of the Fisher information matrix and Inv for the inverse of the Fisher information matrix. For all nine models $400,000$ posterior draws are obtained using a random walk (RW) MH algorithm with a Student-t proposal density with one degree of freedom. The scale matrix for the proposal density was initialized at the identity and reset several times to the sample posterior covariance based on several warm up runs. The step size was adjusted to achieve acceptance rates between 0.25 and 0.5. The RW-MH method is chosen because of its ease of implementation and because the DPMP model posteriors proved not challenging to the degree that more advanced samplers such as those used in Section 4.1 are necessary. Also the generally easy to implement AdMit-MH method proved challenging to apply across all nine models because the optimization routines did not always converge to the extent that invertible Hessians could be obtained. The ESSs for the parameters ranged between 3,000 and 8,000 and for all models the samples could be collected in under 5 minutes. The RW-MH algorithm is implemented in the C language. Marginal likelihoods are computed using the bridge sampling method with warp 3 transformation as described in Section 3.2 and as implemented in the R package "bridgesampling" (Gronau et al., 2017).

For all models I used a $\mathcal{N}(0.05, 1)$ prior for all diagonal elements of $\boldsymbol{A}$, a truncated normal $\mathcal{N}(0.95, 1)\mathrm{I}_{[-1<b.<1]}$ prior[10] for all diagonal elements of $\boldsymbol{B}$, a $\mathcal{N}(0.5, 5^2)$ prior for the free elements in $\boldsymbol{C}$, apart from for $C_{3,1}$ - the element corresponding to upgrades in the one factor model - for

[10]To obtain a proper truncated prior density function as required for marginal likelihood calculation I use the following expression for the truncated normal: $p(b.|\mu, \sigma, ub, lb) = \frac{\phi\left(\frac{b.-\mu}{\sigma}\right)}{\sigma\left(\Phi\left(\frac{ub-b.}{\sigma}\right)-\Phi\left(\frac{lb-b.}{\sigma}\right)\right)}$, where $\phi(\cdot)$ and $\Phi(\cdot)$ are used here to denote the standard normal density function and CDF respectively and ub and lb are the upper and lower truncation bounds.

which a $\mathcal{N}(-0.5, 5^2)$ prior is used, a $\mathcal{N}(-5, 5^2)$ prior for the baseline intensities w_1, w_3 and w_4, and a $\mathcal{N}(-10, 5^2)$ for the baseline intensity of the IG $\rightarrow$ DEF risk process w_2. The prior means can be considered as inspired by a combination of preceding work (e.g. Creal et al. (2011b, 2013)) and ML estimates, whereas the prior variances are set such that the resulting priors are less informative than the data. The approach is therefore somewhat similar to an empirical Bayes approach and follows the general guidelines on the appropriate informativeness of priors for the purposes of objective model comparison set forth in Kass & Raftery (1995).

Due to the significant influence priors can have on model comparison outcomes, I performed a sensitivity analysis with varying degrees of diffuseness of the prior specifications. The results are reported in Appendix B. The main conclusions are that the prior specifications considered did not significantly affect the relative performance of the models. As expected, the more diffuse specifications more strongly favored the models with fewer parameters but not to the degree that the rankings of the models changed. Ultimately my choice for the priors as stated above is due to the fact that the less diffuse priors considerably reduced the posterior uncertainty for several parameters, hindering objective model comparison.

For all nine models the marginal log-likelihoods along with log-likelihoods and BICs based on ML estimates of the parameters are reported in Table 3a. Double the log Bayes factors and ΔBIC statistics for the hypotheses I test are displayed in Table 3b. Surprisingly - unlike in the study of Creal et al. (2013) - the evidence does not favor the two factor models over the one factor models since the Bayes factors with the one factor model as the null model versus the two factor model as the alternative are negative for all three scaling matrix specifications. The evidence does however strongly favor the three factor models over the one and two factor models. Also there is consistent evidence that the the models with scaling matrices based on the Fisher information matrix outperform the identity scaling matrix models. Especially for the three factor model the data seems to support the inclusion of second order information in the time-varying parameter update. The evidence in favor of the models with inverse vs inverse square root scaling is less conclusive. The Bayes factors comparing these models favor the inverse scaling for two out of three factor specifications, but the evidence is relatively weak (2log BF of 2 suggests little support in favor of the null according to the commonly used guidelines proposed in Kass & Raftery (1995)).

The informal ML estimate based methods for non-nested model comparison seem to perform quite well for these models. The relative rankings based on log-likelihood and BIC differ from those based on marginal log-likelihood only in a few instances. Most notably the relative rank between the inverse and inverse square root scaling matrix, for which the ML based methods prefer the former over the later, contrasts with the Bayesian relative model probabilities. This is the only hypothesis for which a different conclusion is reached based on Bayesian model comparison relative to the informal ML estimation based methods. In general it does seem that the ML based methods favor the alternative models (i.e higher factor models (with more parameters) and more second order information in the scaling matrix) over the simpler null hypothesis models more strongly than the Bayesian method of model comparison. This is evidenced by the ΔBIC statistics reported in Table 3b which are higher than the corresponding 2log BFs for all but one model comparison. Since the ΔBIC statistic serves as a rough approximation to double the log Bayes factor (Carlin & Louis, 2000) the two should be comparable. Especially higher factor models score better on the ΔBIC statistic, suggesting that the penalty term for additional parameters does not fully compensate for the added parameter uncertainty introduced by the higher factor models.

Table 2: Likelihoods and model comparison results

(a) Marginal log-likelihoods, Log-likelihoods and Bayesian information criterion (BIC)

	1 factor			2 factor			3 factor		
Scaling	Marginal log-likelihood	Log-likelihood	BIC	Marginal log-likelihood	Log-likelihood	BIC	Marginal log-likelihood	Log-likelihood	BIC
Identity	−15299.9	−15262.4	30544.6	−15304.6	−15265.5	30553.9	−15290.0	−15246.2	30522.3
Inverse sqrt	−15295.5	−15258.1	30536.1	−15300.6	−15260.7	30544.4	−15285.2	−15240.5	30510.8
Inverse	−15294.3	−15256.6	30533.0	−15299.5	−15258.4	30539.8	−15286.4	−15239.4	30508.5

(b) Bayes factors and ΔBIC statistics

1|0	$2\log \mathrm{BF}_{1\|0}$	ΔBIC	1|0	$2\log \mathrm{BF}_{1\|0}$	ΔBIC
1-H|1-I	8.6	8.5	2-I|1-I	−9.5	−9.4
1-Inv|1-H	2.5	3.1	2-H|1-H	−10.2	−8.3
2-H|2-I	8.0	9.6	2-Inv|1-Inv	−10.5	−6.8
2-Inv|2-H	2.2	4.6	3-I|1-I	19.7	22.3
3-H|3-I	9.7	11.4	3-H|1-H	20.7	25.2
3-Inv|3-H	−2.5	2.3	3-Inv|1-Inv	15.7	24.4

Notes: Estimates for the marginal log-likelihoods, regular log-likelihoods and Bayesian information criterion (BIC) scores for all nine models based on a 400,000 draw sample from a random walk Metropolis-Hastings chain are reported in (a). The implied Bayes factors and ΔBIC statistics for the 12 hypotheses under consideration are reported in (b).

The plots of the posterior means of the log intensities - the predicted $\lambda_{ij,t}$, which are the same for all firms i due to the i.i.d. assumption for the intensities in the cross-section - displayed in Figure 4 help shed light on the observed differences in model probabilities. The plots confirm the finding reported in Koopman et al. (2008) and Creal et al. (2013) that the intensities of credit rating upgrades seem to be driven by different dynamics as downgrades. This follows from the clear differences between the predicted log intensities between the three factor and one factor models. Similarly, the dynamics for default intensities seem to differ from the dynamics of the intensity for regular downgrades from IG → SIG. The mean intensities for the two factor specification are not reported for the sake of the clarity of the plots. As expected the predicted log intensities for the two factor models follow a similar pattern as the three factor models for the upgrade risk process and a similar pattern to the one factor models for the IG → SIG risk process.

The differences stemming from the different scaling matrices are less prominent, explaining why the differences in marginal likelihoods for these models are smaller as well. Nevertheless, there do seem to be some consistent differences that are particularly evident for the default risk process intensities. The bottom plots in Figure 4 show that - when the intensities are relatively low - the predicted log intensities of the model with inverse scaling appear clearly more responsive than those predicted by the model with identity scaling. This effect of the inverse Fisher information scaling matrix is as expected based on the discussion in Section 4.2.1. Since the model comparisons prove the models with inverse scaling to be preferred over those with identity scaling, this greater responsiveness is likely a reflection of the inverse scaling models enhanced capacity to capture the time-variation in intensity. The log-intensities predicted by the models with inverse square root scaling matrix also appear more responsive relative to the models with identity scaling but not to the same degree as those with the inverse scaling.

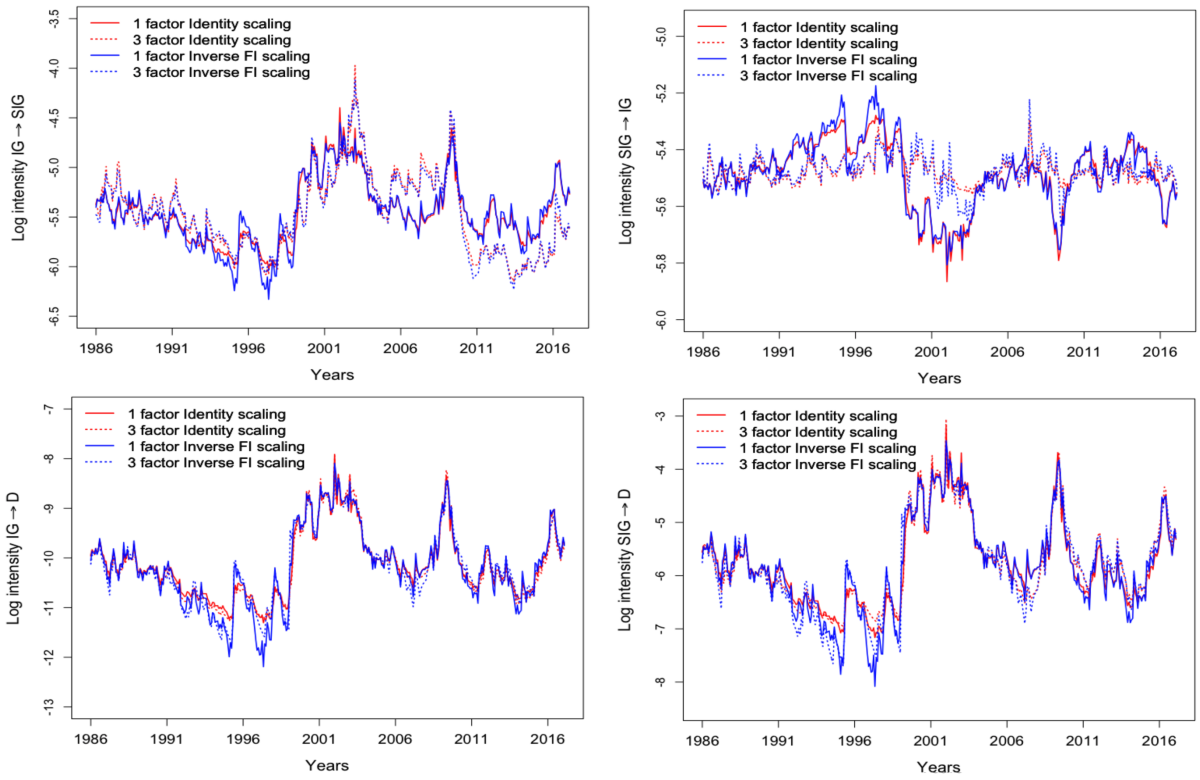


Figure 4: Plots of the mean log intensities for credit rating transitions as predicted by the one and three factor dynamic pooled marked point process models both with identity and Fisher information scaling matrices for the period 1986-01-01 until 2017-02-01. For all four models the means are based on 400,000 posterior draws produced with the random walk Metropolis-Hastings algorithm.

For the upgrade transition the plot of the log intensities predicted by the three factor models suggest that there is actually very little time-variation in intensity for this rating transition. Next, I consider the question of time-variation of the intensities in more detail as it helps understand why the evidence did not support the two factor models over the one factor models. Using the Bayesian approach to hypothesis testing I show that there is indeed little evidence to support that the upgrade intensity is time-varying. In addition I show how the Bayesian framework enables an assessment of the evidence for time-variance of the intensity of the IG $\rightarrow$ DEF process in the three factor models. This hypothesis involves the joint posterior for $C_{2,1}$ and $C_{2,3}$ and hence knows no straightforward alternative in an ML setting.

4.2.3 Time-Variance of the Intensities of Rating Transitions

Evidently the DPMP models inherently assume that the intensities of rating transitions are time-varying. The data might however not necessarily support such time-variance. The suspicion incited by the plot in Figure 4 regarding the lack of time-variance in the log intensity of the upgrade transition, is further raised by the HPD plots in Figure 5a. The 99% HPD bounds for the intensity of the upgrade intensity are wide relative to the IG $\rightarrow$ SIG and SIG $\rightarrow$ DEF intensities and the mode seems relatively constant. The time-variance for the IG $\rightarrow$ DEF intensity also seems questionable mainly due to the large uncertainty proportionate to its level. But for the IG $\rightarrow$ DEF intensity this

is hard to judge from the HPD plots due to its generally very low level.

Traditionally such hypotheses of no time-variation in parameters can be investigated more formally in time-varying parameter models by testing whether some of the autoregressive coefficients (typically the diagonal elements for a diagonal $\boldsymbol{A}$ in the GAS update (2)) are significantly different from zero at some significance level α. This is equivalent to testing whether zero is inside some $1-\alpha$ confidence interval for the autoregressive coefficient of interest. Similarly, using a Bayesian approach we could consider the strength of evidence for time-variance in the log intensity of the SIG $\rightarrow$ IG transition process by considering whether the $(1-\alpha)100\%$ HPD region of the a_2 parameters in the three and two factor models contain zero[11]. We can similarly test for time-variance of the log intensities of the other risk processes for which a row in $\boldsymbol{C}$ contains a one. In certain other instances we must consider the estimated parameters in $\boldsymbol{C}$ to assess time-variation. For example for the log intensity of the SIG $\rightarrow$ IG transition in the one factor model, no time-variation is implied if $C_{3,1}$ is zero. This covers all cases with the exception of the IG $\rightarrow$ DEF transition in the three factor model for which both $C_{2,1}$ and $C_{2,3}$ contribute to the time-variance of the process. I consider this special case separately.

Table 9 contains the estimation results for the parameters of the three models with the inverse Fisher information as scaling matrix. Due to space considerations the estimation results for the other six models are deferred to Appendix C. For the $\boldsymbol{A}$ parameters and for the $\boldsymbol{C}$ parameters of the one and two factor models, asterisks * and ** are used to indicate whether either the 95% or 99% HPD regions of the parameters include zero. This provides an indication of the strength of evidence for the time-variance of the log intensities. The results suggest that as we expected the evidence in support for time-variance is weak for the log intensity of the SIG $\rightarrow$ IG transition. In the three factor model the 95% HPD region of a_2 includes zero and for the two factor model the 99% HPD region includes zero. Oddly, in the one factor model there does seem to be support for time-variance of the the log intensity of the SIG $\rightarrow$ IG transition. The upgrade process does depend negatively on the time-varying factor in the one factor model ($C_{1,3} < 0$). This is consistent with the findings in Creal et al. (2013) and - since the time-varying factor is identified by a downgrade process - suggests that upgrade intensities follow a process that moves oppositely to the downgrade intensities. For the factor models with identity and inverse square root scaling matrices the parameter estimation results tell a similar story. For the two factor model with identity scaling matrix we however find that the 99% HPD region for the $C_{2,1}$ also includes zero, confirming that time-variance for the log intensity of the IG $\rightarrow$ DEF transition is also not supported as strongly by the evidence as are the log intensities of the IG $\rightarrow$ SIG and the SIG $\rightarrow$ DEF transitions.

To assess time-variance for the log intensity of the IG $\rightarrow$ D transition in the three factor models we need to consider the probability that both $C_{2,1}$ and $C_{2,3}$ are zero at the same time. This is not as straightforward as simply evaluating the HPD regions of each of the parameters separately and interpreting the fact that both include zero as limited evidence for time-variance. The bottom plot in Figure 5b shows how such an approach ignores the correlation between the parameters. A more valid approach would be to consider the amount of draws for which both $C_{2,1}$ and $C_{2,3}$ are less than zero. For the inverse scaling model this is only the case for 167 out of 400,000 draws and for the identity and inverse square root scaling models this occurs for only 114 and 60 draws. The respective posterior probabilities for these events are therefore 4.2×10^{-4}, 2.9×10^{-4} and 1.5×10^{-4}

[11] Alternatively, we could formally consider the hypothesis of time-variance in the Bayesian model comparison framework (i.e. estimating a model with a_2 and b_2 restricted to zero and comparing it with an unrestricted model). The HPD interval approach is however significantly less involved as it does not require estimation of a new model.

Table 4: Parameter estimates for the dynamic marked point process one, two and three factor models with inverse information matrix scaling (DPMP1-Inv, DPMP2-Inv and DPMP3-Inv respectively)

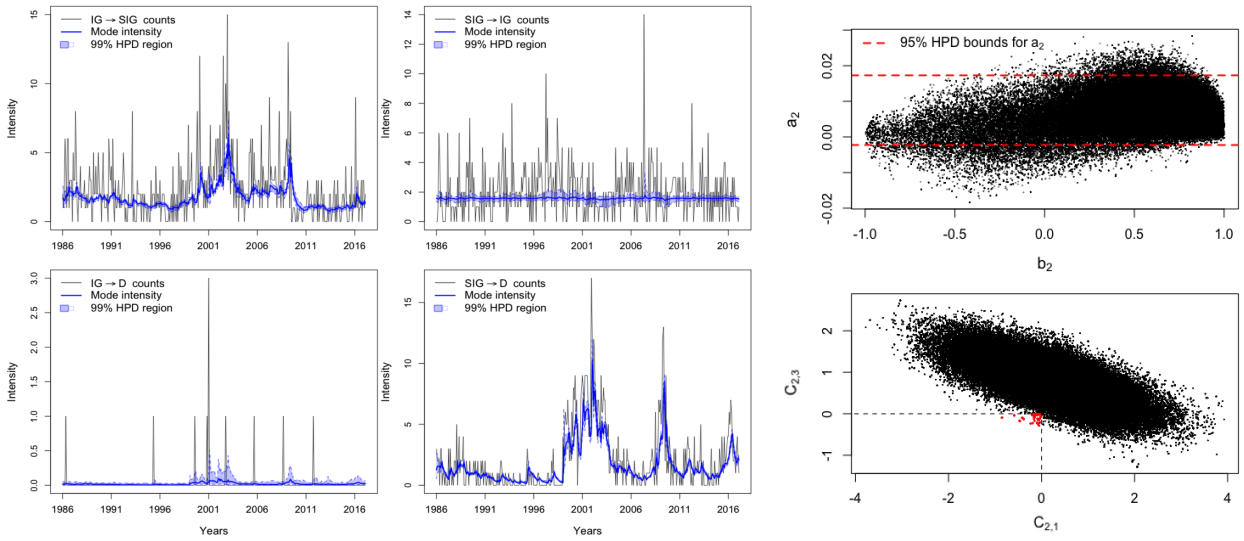
	DPMP1-Inv				DPMP2-Inv				DPMP3-Inv			
$\boldsymbol{\theta}^{[1]}$	$\mathrm{E}(\cdot\|\boldsymbol{Y}_T)$	SE	NSE ($\times 10^{-2}$)	$\boldsymbol{\theta}^{[2]}$	$\mathrm{E}(\cdot\|\boldsymbol{Y}_T)$	SE	NSE ($\times 10^{-2}$)	$\boldsymbol{\theta}^{[1]}$	$\mathrm{E}(\cdot\|\boldsymbol{Y}_T)$	SE	NSE ($\times 10^{-2}$)
a_1	0.047	0.005	0.006	a_1	0.046	0.005	0.007	a_1	0.019	0.003	0.005
				a_2	0.008**	0.004	0.007	a_2	0.007*	0.005	0.008
								a_3	0.038	0.003	0.005
b_1	0.961	0.016	0.020	b_1	0.961	0.016	0.023	b_1	0.964	0.017	0.027
				b_2	0.716	0.313	0.686	b_2	0.571	0.383	0.802
								b_3	0.963	0.016	0.029
$C_{1,1}$	0.387	0.042	0.052	$C_{1,1}$	0.393	0.041	0.064	$C_{2,1}$	0.199	0.980	1.588
$C_{2,1}$	0.890	0.329	0.416	$C_{2,1}$	0.888	0.327	0.510	$C_{2,3}$	0.779	0.471	0.814
$C_{3,1}$	−0.137	0.041	0.049								
d_1	−5.402	0.121	0.163	d_1	−5.418	0.123	0.210	d_1	−5.477	0.160	0.262
d_2	−10.054	0.455	0.576	d_2	−10.073	0.457	0.764	d_2	−10.125	0.472	0.740
d_3	−5.504	0.058	0.076	d_3	−5.447	0.061	0.115	d_3	−5.463	0.047	0.088
d_4	−5.673	0.303	0.403	d_4	−5.706	0.305	0.538	d_4	−5.755	0.299	0.487
Accept	0.34				0.30				0.21		

Notes: Estimation results for the parameters of the DPMP1-Inv, DPMP2-Inv and DPMP3-Inv models based on a $400,000$ draw long Markov chain produced using the random walk Metropolis-Hastings algorithm. Reported for all three models and for all parameters are posterior mean ($\mathrm{E}(\cdot|\boldsymbol{Y}_T)$), standard deviation ($\mathrm{SE} = \sqrt{\mathrm{var}(\cdot|\boldsymbol{Y}_T)}$) and numerical standard error (NSE). For the $a_\cdot$ parameters and the $C_{\cdot,\cdot}$parameters of the one and two factor models, one star (*) is used to denote that the 95% highest posterior density region (HPD) includes zero and two stars (**) are used to denote that the 99% HPD includes zero. Also reported for all three Markov chains are the acceptance rates.

suggesting that the evidence against time-variance for the log intensity of the IG → D transition is quite weak in these models. In contrast, a naive assessment of the individual p-values of the $C_{2,1}$ and $C_{2,3}$ parameters in an ML setting would have led convincingly to the conclusion that there is no time-variance in the log intensity of the IG → D process (i.e. the null hypothesis of no time-variance cannot be rejected).

The top plot in Figure 5b illustrates how low values for a_2 cause considerable uncertainty in the corresponding b_2 parameter. This makes sense because since $\omega_2 = 0$, if a_2 then also equals zero the b_2 parameter is unidentified given that $f_1 = 0$. Since a_2 is close to zero for all models, the resulting uncertainty in b_2 explains why the evidence supports the one factor models over the two factor models. The separate factor for the upgrade process introduces a lot of additional uncertainty in the model without much gain in ability to fit the data since no time-variance for the upgrade factor can also be captured by the one factor model. Consequently the marginal likelihood deteriorates in moving from the one to the two factor model.

The analysis presented here arrives at slightly different conclusions regarding the dynamics of the intensities of credit rating transition as the preceding work of Creal et al. (2013). Although, time-variance for the second factor seems doubtful considering their reported parameter estimates as well, they do find considerable improvement in BIC and likelihood by using the two factor instead of the one factor specification. These differences are most likely due to the different datasets being considered, as they had access to a larger dataset of Moody rating events for all US corporates. As



(a) DPMP3-Inv - HPD of predicted intensities

(b) Bivariate posteriors

Figure 5: The 99% highest posterior density (HPD) regions of the intensities of credit rating transitions for the period 1986-01-01 to 2017-02-01 as predicted by the dynamic pooled marked point process model with three time-varying factors and inverse information matrix as scaling matrix (DPMP3-Inv) are plotted on the left. Posterior draws of the parameters a_2 and b_2 are plotted in the top right and the posterior draws of the parameters $C_{2,1}$ and $C_{2,3}$ are plotted in the bottom right for the DPMP3-Inv model. In the bottom right plot, draws for which both $C_{2,1}$ and $C_{2,3}$ are less than zero are plotted as red dots. The HPD regions and joint distribution plots are based on a $400,000$ draw sample from a random walk Metropolis-Hastings Markov chain.

a consequence of this richer dataset there is probably more support for time-variance in the upgrade rating transitions and, by extension, better relative performance for the two factor specification.

4.3 Multivariate Student-t Random Coefficients Covariance Model

The multivariate counterparts to univariate volatility models come in the form of time-varying covariance and correlation models. Since time-variance in the dependencies between financial asset returns is now commonly accepted (See e.g. Longin & Solnik (1995)), the ability to model the temporal developments of these dependencies is paramount to advancing understanding of financial market dynamics. In addition, covariances and correlations are crucial inputs to many practical applications such as portfolio selection, hedging, option pricing and VaR estimation.

Central to the modeling of time-varying covariance matrices is the trade-off between flexibility in the specification of the dynamics and parsimonious parameterization (Bauwens et al., 2006). For the more flexible models, such as the VEC and BEKK[12] models (Bollerslev et al., 1988, Engle & Kroner, 1995), the dimension of the parameter space quickly turns unmanageable as the number of assets to be modeled increases. Even in the case of a mere three assets the number of parameters for the unrestricted versions of the VEC and BEKK are 78 and 24 respectively. Consequently, most applications of these models in the literature have been to either very small sets of assets - typically

[12]VEC is simply short for vectorized as the model specifies dynamics directly for the vectorized covariance matrix, whereas BEKK is an acronym that stems from synthesized work of Baba, Engle, Kraft and Kroner.

only 2 to 3 - or through imposing restrictions on the parameters. The popular Dynamic Conditional Correlation (DCC) model proposed by Engle & Kroner (1995) is in most applications also estimated with a parametrization that restricts all correlations to follow similar dynamics (scalar DCC, see Caporin & McAleer (2012)). Nevertheless, unrestricted versions do seem to outperform restricted versions as shown in Burda & Maheu (2013) and Hafner & Franses (2009), suggesting that when the data permits the extra flexibility should be utilized.

Burda & Maheu (2013) illustrate that the difficulty in estimating fully parameterized covariance models for even a moderate number of assets stems from the data containing little information for particular parameters, causing the likelihood to be extremely flat in certain dimensions. This greatly challenges the ability of optimization routines to converge, limiting applications of the ML method for fully parameterized covariance models. Several solutions have been proposed, mostly for DCC type models, that address the optimization problem directly. Engle (2007) and Engle et al. (2008) for instance propose to split up the optimization of the likelihood into many smaller optimizations that fit time-varying correlation coefficients for either two assets or small subsets of assets at a time. Using these methods it is however not clear how to ensure positive definiteness of the resulting covariance matrices. Another recent approach that does allow positive definiteness to be enforced, introduced by Bauwens et al. (2016), relies on a highly robust optimization technique based on a quasi-likelihood, which allowed them to fit unrestricted Hadamard DCC models for up to 30 assets. The Bayesian approach is also proving more robust for covariance models than straightforward likelihood optimization approaches, as shown by Burda & Maheu (2013) and Burda (2015) for unrestricted BEKK models for up to 5 assets. For the unrestricted 5 asset model analyzed here I also find the Bayesian approach to prove more robust, since simple optimization routines did not converge. However, beyond this number of assets, even the Bayesian approach breaks down because of excessive numerical instabilities. More importantly, both the Bayesian approach and the robust optimization technique presented by Bauwens et al. (2016), do not directly address the underlying issue identified by Burda & Maheu (2013), regarding the data being highly uninformative for certain parameters in fully parameterized models.

For the unrelated but widely used class of generalized linear models there exists a rich tradition of coping with the data being highly uninformative for certain parameters through employing Bayesian hierarchical modeling techniques (e.g. random effects or mixed effects models and Spike-and-Slab regressions, see Browne & Draper (2006) for a comparison of Bayesian and ML techniques for such models). These techniques allow prior knowledge regarding similarities and relations between certain sets of parameters to be expressed and exploited to enable inference for models of far greater complexity - with the number of parameters frequently exceeding the number of observations - than what would be possible without the hierarchical structure. In this section I consider the approach of imposing hierarchical normal priors on a subset of the autoregressive parameters $\boldsymbol{A}$ and $\boldsymbol{B}$ in the GAS update equation, for the multivariate Student-t GAS model proposed by Creal et al. (2011a), parametrized such that it allows independent dynamics for all elements of the correlation matrix. The hierarchical priors express the belief that the subsets of the parameters for which they are specified, are highly similar and the information in the data regarding these parameters can therefore be shared among the parameters in these subsets. The use of hierarchical modeling techniques in the context of volatility models has been explored previously by Brownlees (2015) to improve estimation in a large panel of GARCH models, but have yet to be applied in a Bayesian framework to time-varying covariance or correlation models.

In this Section I present results for the hierarchical multivariate GAS-t model estimated on

datasets of 5 but also 10 portfolios. The 5 asset case allows for comparison with an equivalently unrestricted GAS-t model without hierarchical priors, whereas the 10 portfolio case serves to illustrate that the hierarchical modeling approach generalizes well to far higher dimensional covariance models than would be possible without the hierarchical priors. For both the 5 and 10 asset applications I also estimate a more restricted covariance model for which the dynamics of all correlations are governed by the same autoregressive coefficients, to show that the hierarchical models additional flexibility does result in observable differences in predicted correlations. In fact, based on the results of the 5 asset case, it appears that the hierarchical covariance model sacrifices very little if any flexibility in its modeling of the dynamics of the correlations relative to the unrestricted model. In terms of model probabilities the evidence favors the hierarchical model over the restricted model for both datasets but is significantly stronger for the 10 asset model. The unrestricted model clearly performed worst in terms of posterior model probabilities, as a result of severe parameter uncertainty.

To place these models in the context of the larger literature on time-varying covariance models, the multivariate Student-t GAS model may be considered as similar in its specification to the DCC model (Creal et al., 2011a). The version of the GAS-t that I refer to as being unrestricted is similar in terms of flexibility of the correlation dynamics as the Hadamard DCC and the restricted version is similar to the scalar DCC. Bar the models estimated using the specialized optimization procedures mentioned above, the 10 asset hierarchical model represents one of the largest time-varying covariance models that allows for separate dynamics for each correlation pairs, considered in the literature thus far. In theory the hierarchical technique should easily extend to covariance models of even far greater number of assets and should arguably only become more powerful as more complex relationships among the assets can start being incorporated into the hierarchical prior structure.

For the GAS-t model however, the constraints for estimating larger models are also computational in nature. Its GAS update specification causes computational costs to increase quadratically with the number of assets, making it ill-suited to extend beyond 10 assets. The hierarchical and computational techniques presented here however easily transfer to simpler DCC models for which the cost scale only linearly in the number of assets meaning even larger sets of assets could be modeled with similar flexibility. I proceed by first introducing the model, the hierarchical prior setup and addressing the computational hurdles that need to be overcome to estimate the models within a reasonable time frame. In Section 4.3.5 and 4.3.6 the empirical applications are discussed.

4.3.1 The Multivariate GAS-t Model

Creal et al. (2011a) specify a multivariate Student-t distribution for a $k \times 1$ vector of asset returns $\boldsymbol{y}_t$, which takes the form[13]

$$p(\boldsymbol{y}_t|\boldsymbol{\Sigma}_t, \bar{\nu}) = \frac{\Gamma\left(\frac{1+k\bar{\nu}}{2\bar{\nu}}\right)}{\Gamma\left(\frac{1}{2\bar{\nu}}\right)\left(\frac{(1-2\bar{\nu})\pi}{\bar{\nu}}\right)^{k/2}|\boldsymbol{\Sigma}_t|^{1/2}}\left(1+\bar{\nu}\frac{\boldsymbol{y}_t'\boldsymbol{\Sigma}_t^{-1}\boldsymbol{y}_t}{1-2\bar{\nu}}\right)^{-\frac{1+k\bar{\nu}}{2\bar{\nu}}}.$$

The scale matrix is defined as a function of a set of time-varying parameters $\boldsymbol{\Sigma}_t = \boldsymbol{\Sigma}_t(\boldsymbol{f}_t)$, and the GAS update equation (2) is defined with the scaling matrix (4) set equal to the inverse of the Fisher information matrix. Creal et al. (2011a) derive the following generic score and information matrix

[13] Just as for the Beta-Gen-t-EGARCH, I reparameterize the distribution as originally formulated by (Creal et al., 2011a) with the inverse degrees of freedom parameter so as to enable uninformative prior specification.

with respect to $\boldsymbol{f}_t$

$$\begin{aligned}\boldsymbol{\nabla}_t =& \frac{1}{2}\boldsymbol{\Psi}_t'\boldsymbol{\Sigma}_{t\otimes}^{-1}\big(w_t\boldsymbol{y}_{t\otimes} - \text{vec}(\boldsymbol{\Sigma}_t)\big),\\ \boldsymbol{\mathcal{I}}_t =& \frac{1}{4}\boldsymbol{\Psi}_t'\boldsymbol{J}_{t\otimes}'\big(g\boldsymbol{G} - \text{vec}(\boldsymbol{I}_k)\,\text{vec}(\boldsymbol{I}_k)'\big)\boldsymbol{J}_{t\otimes}\boldsymbol{\Psi}_t,\end{aligned}$$

where $\boldsymbol{\Psi}_t = \partial\text{vec}(\boldsymbol{\Sigma}_t)/\partial\boldsymbol{f}_t'$. The notation $\text{vec}(\cdot)$ is used to denote the operator that stacks the columns of the matrix to which it is applied on top of each other into one column vector and the application of the subscript $\otimes$ to a matrix is used to denote the Kronecker product of the matrix with itself. Also, the matrix $\boldsymbol{J}_t$ is defined as $\boldsymbol{\Sigma}_t^{-1} = \boldsymbol{J}_t\boldsymbol{J}_t'$, the matrix $\boldsymbol{G} = \text{E}((\boldsymbol{z}_k\boldsymbol{z}_k')_\otimes)$, where $\boldsymbol{z}_k$ is a k-dimensional vector of $\mathcal{N}(0,1)$ distributed random variables, the scalar $w_t = (1+k\bar{\nu})/(1-2\bar{\nu}+\bar{\nu}\boldsymbol{y}_t'\boldsymbol{\Sigma}_t^{-1}\boldsymbol{y}_t)$ and the scalar $g = (1+k\bar{\nu})/(1+2\bar{\nu}+k\bar{\nu})$. Note that since we known the first through to fourth moments of $\boldsymbol{z}_k$, the matrix $\boldsymbol{G}$ is known beforehand and consists of 0's, 1's and 3's.

Creal et al. (2011b) provide several options for the link function that maps $\boldsymbol{f}_t$ to $\boldsymbol{\Sigma}_t$. The scale matrix is first decomposed as

$$\boldsymbol{\Sigma}_t = \boldsymbol{D}_t\boldsymbol{R}_t\boldsymbol{D}_t,$$

where $\boldsymbol{D}_t$ is a diagonal matrix with the time-varying standard deviations on its diagonal and $\boldsymbol{R}_t$ the time-varying correlation matrix. I choose to use the specification with the log link function for the variances and the hyperspherical coordinates transformation for the correlation matrix, such that

$$\boldsymbol{R}_t = \boldsymbol{X}_t'\boldsymbol{X}_t,$$

Where $\boldsymbol{X_t}$ is an upper triangular matrix with its elements defined as

$$x_{ij,t} = \begin{cases}1 & \text{if } i=j=1,\\ \cos(\phi_{ij,t})\Pi_{l=1}^{i-1}\sin(\phi_{lj,t}) & \text{if } i<j,\\ \Pi_{l=1}^{i-1}\sin(\phi_{lj,t}) & \text{if } i=j,\\ 0 & \text{otherwise,}\end{cases} \tag{22}$$

for $i,j = 1,...,k$. The vector of time-varying parameters is then defined as follows

$$\boldsymbol{f}_t = \begin{bmatrix}\log(\text{diag}(\boldsymbol{D}_t^2))\\ \boldsymbol{\phi}_t\end{bmatrix},$$

where $\boldsymbol{\phi}_t$ is the vector containing the $k(k-1)/2$ angles $\phi_{ij,t}$ for $j = 1,...,k-1$ and $i<j$. This specification has several appealing properties. The log link function guarantees that the variances remain positive without requiring further restrictions on the parameters in the GAS update equation (2). The hyperspherical coordinate transformation described in (22) ensure that the generally challenging condition of positive definiteness of the correlation matrix $\boldsymbol{R}_t$ is met and simultaneously restrains the diagonal elements of $\boldsymbol{R}_t$ to one and the off-diagonal elements to be less than one in absolute value (Jaeckel & Rebonato, 1999). Moreover, Pourahmadi & Wang (2015) show that $\boldsymbol{X}_t$ is uniquely identified if the angles are restricted to the range $(0,\pi)$. As mentioned by Creal et al. (2011a), the range restrictions on the angles required for identification do not need to be enforced as it is found in practice that possible minor violations typically do not cause numerical issues. Since, the number of angles is the same as the number of free elements in the correlation matrix, the model has similar flexibility in describing the dynamics of covariances as unrestricted BEKK or

DCC models.

The time-varying factor specification in (23) implies the Jacobian matrix of the function mapping $\boldsymbol{f}_t$ to $\boldsymbol{\Sigma}_t$ takes the following form

$$\boldsymbol{\Psi}_t = \frac{1}{2}\big(\boldsymbol{I}_k \oplus \boldsymbol{\Sigma}_t\big)\boldsymbol{\mathcal{S}_D} + \boldsymbol{D}_{t\otimes}\big(\boldsymbol{I}_{k^2} + \boldsymbol{\mathcal{C}}_k\big)\big(\boldsymbol{I}_k \otimes \boldsymbol{X}_t'\big)\boldsymbol{Z}_t\boldsymbol{\mathcal{S}}_{\boldsymbol{\phi}}, \tag{23}$$

where the operator $\oplus$ denotes the Kronecker sum, the matrix $\boldsymbol{\mathcal{C}_k}$ is known as the commutation matrix and is defined by the relation $\boldsymbol{\mathcal{C}_k}\,\mathrm{vec}(\boldsymbol{P}) = \mathrm{vec}(\boldsymbol{P}')$ for some arbitrary matrix $\boldsymbol{P}$, the matrices $\boldsymbol{\mathcal{S}_D}$ and $\boldsymbol{\mathcal{S}_\phi}$ are defined respectively such that $\mathrm{vec}(\boldsymbol{D}_t) = \boldsymbol{\mathcal{S}_D}\boldsymbol{f}_t$ and $\boldsymbol{\phi}_t = \boldsymbol{\mathcal{S}_\phi}\boldsymbol{f}_t$ and the matrix $\boldsymbol{Z}_t$ represents the matrix of derivatives $\partial\,\mathrm{vec}(\boldsymbol{X}_t)/\partial\boldsymbol{\phi}_t'$, which are defined in Creal et al. (2011a).

Following Creal et al. (2011a) I consider a slightly modified and restricted versions of the time-varying parameter GAS update equation (2). First, the coefficient matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are constrained to be diagonal. Interaction effects between the time-varying factors are still partly incorporated through the scaling matrix. Second, the reparameterization in (3) is used and the $\tilde{\boldsymbol{\omega}}$ vector is fixed at the outset such that $\boldsymbol{\Sigma}_t(\tilde{\boldsymbol{\omega}})$ equals the sample covariance. This approach of fixing the long-run mean of the covariance to an unbiased estimate is commonly used when estimating time-varying covariance models and is known as targeting (Caporin & McAleer, 2012). The usefulness of targeting stems from the fact that sensible parameter restrictions can generally not be applied to the intercept of the time-varying process and therefore the dimension of the parameter space would still be of order $\mathcal{O}(k^2)$. Similarly, hierarchical grouping such as introduced in the next subsection does not make sense for the $\tilde{\boldsymbol{\omega}}$ parameter. The long-run mean of the angles can be obtained by inverting the one-to-one mapping from $\boldsymbol{X}_t$ to $\boldsymbol{R}_t$ and plugging in the sample estimate of the correlation matrix. The exact expressions for the angles given an estimate of the correlation matrix can be found in Pourahmadi & Wang (2015).

4.3.2 Hierarchical Prior Specification

Section 3.3 introduced the notion of a hierarchical prior as the specification of an additional layer of prior distributions for the parameters of the prior on $\boldsymbol{\theta}$. Applications of such hierarchical priors commonly serve the purpose of grouping sets of parameters that we a priori know to have common features. This is mostly useful given a scenario where the data is little informative with respect to certain groups of parameters. In such cases it is likely beneficial to pool the information regarding the parameters of the same group to better inform their joint posterior. Considering the discussion a the start of this Section (4.3), regarding the large parameter uncertainty typical in time-varying parameter models, a hierarchical prior seems appropriate.

In this section I introduce a hierarchical prior setup for the autoregressive parameters in the diagonal $\boldsymbol{A}$ and $\boldsymbol{B}$ matrices that correspond to the time-varying angles $\boldsymbol{\phi}_t$ in the multivariate GAS-t model. In part due to the complex relation that maps the angles to the observable correlation, we can generally consider ourselves ignorant regarding differences in the autoregressive processes of the angles. This contrasts with for instance the autoregressive parameters of the variances or the inverse degrees of freedom parameter, which we assume to be different in nature from the autoregressive angle parameters. This ignorance regarding within group differences implies a priori exchangeability (Gelman et al., 2014, Ch. 5), implying that we may assume a priori that the elements from these two sets of autoregressive parameters are generated by the same prior distributions. Let these sets of parameters be denoted by $\boldsymbol{a}_{\boldsymbol{\phi}}$ and $\boldsymbol{b}_{\boldsymbol{\phi}}$ respectively, and let the priors for their elements be defined

as

$$p(a_{i,\boldsymbol{\phi}}) = \mathcal{N}(\mu_{a,\boldsymbol{\phi}}, \sigma^2_{a,\boldsymbol{\phi}}), \tag{24}$$
$$p(b_{i,\boldsymbol{\phi}}) = \mathcal{N}(\mu_{b,\boldsymbol{\phi}}, \sigma^2_{b,\boldsymbol{\phi}}),$$

for $i = 1, ..., k(k-1)/2$. The data is often inadequate to properly identify all autoregressive parameters for the hyperspherical angles. Even in the moderate dimensional applications presented in Section 4.3.5 ($k = 5$ and $k = 10$), the difficulty in estimation is evidenced by the notably high posterior variance of several of the $\boldsymbol{b}_{\boldsymbol{\phi}}$ parameters (see Table 10).

Generally, the time-varying processes that govern the dynamics of the variances are comparatively well identified and doing inference on the autoregressive parameters of these processes proved nonchallenging. In such cases where the data is highly informative, there is little benefit to imposing hierarchical priors. Therefore, the choice is made to apply the hierarchical priors only to the autoregressive angle parameters. When much larger panels of assets are being modeled however, it is likely that inference for the autoregressive variance parameters would also benefit from the information pooling generated by hierarchical modeling. This effect is demonstrated by Brownlees (2015).

In their empirical application Creal et al. (2011a) take the approach of further restricting the autoregressive parameters for the time-varying angles to single scalars, setting $a_{i,\boldsymbol{\phi}} = a_{1,\boldsymbol{\phi}}$ and $b_{i,\boldsymbol{\phi}} = b_{1,\boldsymbol{\phi}}$ for all i. I consider models both with and without these restrictions and compare them with the model that imposes the hierarchical normal priors (24) on the autoregressive parameters $\boldsymbol{a}_{\boldsymbol{\phi}}$ and $\boldsymbol{b}_{\boldsymbol{\phi}}$. In terms of flexibility and parsimony, the hierarchical model occupies the middle ground between the fully parameterized and restricted versions. The advantage of the Bayesian hierarchical approach is that the data determines the extent to which the model is either flexible, when the data is highly informative regarding the autoregressive angle parameters, or parsimonious, when the data is little informative and the parameters are pulled to a common mean.

Due to the similarity in structure of the prior specification in (24) with the priors on the regression coefficients in random effects models, I refer to the hierarchical covariance model as the multivariate random coefficient GAS-t model.

4.3.3 Efficient Gradient Computation

Doing Bayesian inference for both the hierarchical and nonhierarchical multivariate GAS-t model is challenging for many reasons. Since the dimensionality of the parameter space increases rapidly with k - for $k = 5$ the multivariate GAS-t model without the further restrictions on the diagonal elements of $\boldsymbol{A}$ and $\boldsymbol{B}$ contains 35 parameters and 115 when $k = 10$ - HMC is likely the most viable MCMC method.

The computational costs for applying HMC to the multivariate GAS-t model are however considerable. Likelihoods of time-varying covariance and correlation models are already notoriously expensive to evaluate for even moderate k due to the matrix operations needed at each time series observation t to update the time-varying parameters. Unlike competing multivariate GARCH models that require matrix operations on matrices that are at most of dimension $k \times k$ inside their recursive filters, the GAS-t model requires matrix operations on $k^2 \times k^2$ matrices. Gradient evaluations are unavoidably even more expensive. Particularly, standard symbolic derivatives of matrix

operations can quickly blow up in dimension. For the multivariate GAS-t model, for which the symbolic derivatives are included in Appendix A for reference, the gradient evaluations require matrix multiplications on $k^4 \times k^4$ matrices. Consequently, even for models of a moderate number of assets the computational cost of symbolic derivatives are generally prohibitive for the multivariate GAS-t.

I therefore resort to using techniques from Automatic Differentiation (AD). AD comprises a set of techniques that allow for analytical evaluations of the derivative of a function specified in the form of a computer program. Moreover, the derivative is evaluated at machine accuracy and should in theory require only a small constant multiple of the computational cost of the original computer program (Hascoet & Pascual, 2013). For a general reference on automatic differentiation see Griewank & Walther (2008). In Appendix D I also provide a brief introduction to the two main approaches to AD, forward mode and reverse mode differentiation, and show how the techniques can be adopted to compute the gradient of the multivariate GAS-t log-likelihood far more efficiently compared to evaluating the likelihood using the symbolic expressions in Appendix A.2. In our case we are interested in the gradient of a function with scalar output, making reverse mode differentiation the preferred choice due to what is known as the "cheap gradient principle" Griewank & Walther (2008). For the C language, a free software tool that performs reverse mode automatic differentiation through source code transformation[14] is the Tapenade Automatic Differentiation tool (Hascoet & Pascual, 2013).

However, programs that evaluate the likelihood of GAS models typically contain a long for loop of $T-1$ iterations. This loop can generally not be eliminated due to the recursive definition of the time-varying parameters and challenges the automatic tools in their ability to produce efficient reverse mode derivatives. Reverse mode differentiation namely works by backwards application of the chain rule - computing the derivatives of the inputs with respect to the outputs of all the elementary operation that constitute the complete function. Reverse mode differentiation hence requires most of the intermediate results used in a function evaluation to be stored (see Appendix D for more details). The large for loop needed to evaluate GAS-likelihoods introduces a tremendous number of intermediate products that grows with T, making the default "store-all" approach used by the Tapenade software very memory intensive. This applies especially for the multivariate GAS-t model which requires storage of many intermediate matrix products of $k^2 \times k^2$ dimension (see Appendix A). Knowledge of this special structure can however be exploited through adjustments to the transformed code produced by Tapenade so as to greatly enhance the efficiency of the gradient computation.

The solution I employ is a checkpointing strategy; meaning snapshots of the currents state of a program are stored at certain points in the gradient program. In the backward sweep to compute the gradient, parts of the program are then recomputed starting from these strategically placed checkpoints (see e.g. Griewank (1992)). Checkpointing revolves around the balancing of memory requirements with the additional costs of recomputing parts of the program. For computer programs of GAS model likelihoods a natural strategy would be to place checkpoints at the end of every iteration in the for loop, since a complete snapshot of the current state of the program can be stored in the vector of time-varying parameters $\boldsymbol{f}_t$. The resulting gradient code requires only the storage of a $T \times n$ matrix to store $\boldsymbol{f}_t$ for $t = 1, 2, ..., T$ and the additional derivative arrays to store the intermediate matrix derivative results of one iteration. These derivative arrays are of

[14]This refers to the technique that is used by Tapenade to construct the gradient code, which is more challenging but results in more efficient code as the more common alternative technique of operator overloading (Hascoet & Pascual, 2013).

equal dimension as the original arrays for the intermediate results. Total storage requirements thus equal twice that of the original code plus an amount that diminishes in significance as k increases. Computational cost are fixed at roughly a three times multiple of the original likelihood as a result of one forward pass to compute all $\boldsymbol{f}_t$ and one backward pass that requires a re-computation of intermediate results and a derivative computation at each time t.

In the applications I consider, the additional storage requirements had minimal impact on performance and the tripling of computational time relative to a single likelihood evaluation proved quite close. Computational time for the gradient program tuned out to be roughly a four times multiple of the original program. If however, the model is applied to longer time series, for instance daily data for a 20 year period with moderate $k = 5$, the additional storage of all $\boldsymbol{f}_t$ might have considerable adverse impact on performance. In such instances alternative storage schemes where $\boldsymbol{f}_t$ is stored only at larger intervals or a more advanced binomial checkpointing strategy (see e.g. Griewank (1992)) might be more efficient.

Besides adjusting the code for the gradient produced by Tapenade with the above described checkpointing strategy, I also hand-coded several of the matrix derivatives using the adjoint matrix derivative rules presented in Giles (2008). For certain matrix operations such as matrix inversions the source code transformations are considerably less efficient than analytical matrix expression.

4.3.4 Coping with Large Variation in Posterior Curvature

A further complication to applying HMC to the multivariate GAS-t model comes from the difficulty in estimating the posterior variance needed to set the mass matrix. As mentioned in Section 3.1.3, when the curvature of the posterior varies strongly with position using HMC with the inverse mass matrix set to the posterior covariance tends to result in poor performance. The reason being that the mass matrix attempts to globally decorrelate the parameters. This works well for elliptically shaped Gaussian distributions, but the further removed from Gaussian the posterior is, the less effective a fixed mass matrix.

For the multivariate GAS-t models I encounter two sources of significant variation in curvature. The first is specific to the unrestricted nonhierarchical model where the stationarity constraints on $\boldsymbol{B}$ are highly restrictive for certain angles, causing harsh cut-offs in the posterior range. The second is a well known pathology associated with hierarchical variance parameters, that the specification as in (24) induces strong prior correlations for the parameters $\boldsymbol{a}_{\boldsymbol{\phi}}$ and $\boldsymbol{b}_{\boldsymbol{\phi}}$ with $\sigma_{a,\boldsymbol{\phi}}$ and $\sigma_{b,\boldsymbol{\phi}}$ respectively. This correlation is highly local since for small values of hyper variance parameters the conditional prior variance of the autoregressive parameters is very small and vice versa for large values of the variance parameters, resulting in a funnel shaped prior distribution (see Betancourt & Girolami (2015)). If the data contains little information regarding the autoregressive parameters much of this local variation in prior correlation persists in the posterior.

The first source is hard to address and is simply a consequence of the model not being well identified. Introducing stronger prior information is the logical Bayesian solution. Also due to the fact that the parameters in $\boldsymbol{\theta}$ are relatively independent in the multivariate GAS-t model, the strategy of restricting the mass matrix to be diagonal proves reasonably effective in the application presented below.

Since the second source of nonglobal curvature is common for many hierarchical models there is

actually a body of literature that considers potential solutions. The common solution, discussed in Papaspiliopoulos et al. (2007), involves a different parameterization of the prior setup in (24) known as the noncentered parameterization. In this case the resulting parameterization for the elements in $\boldsymbol{a}_{\boldsymbol{\phi}}$ will for example be of the form

$$\begin{aligned} a_{i,\boldsymbol{\phi}} &= \mu_{a,\boldsymbol{\phi}} + \sigma_{a,\boldsymbol{\phi}}\xi_{a,i} \\ \xi_{a,i} &\sim \mathcal{N}(0,1). \end{aligned} \tag{25}$$

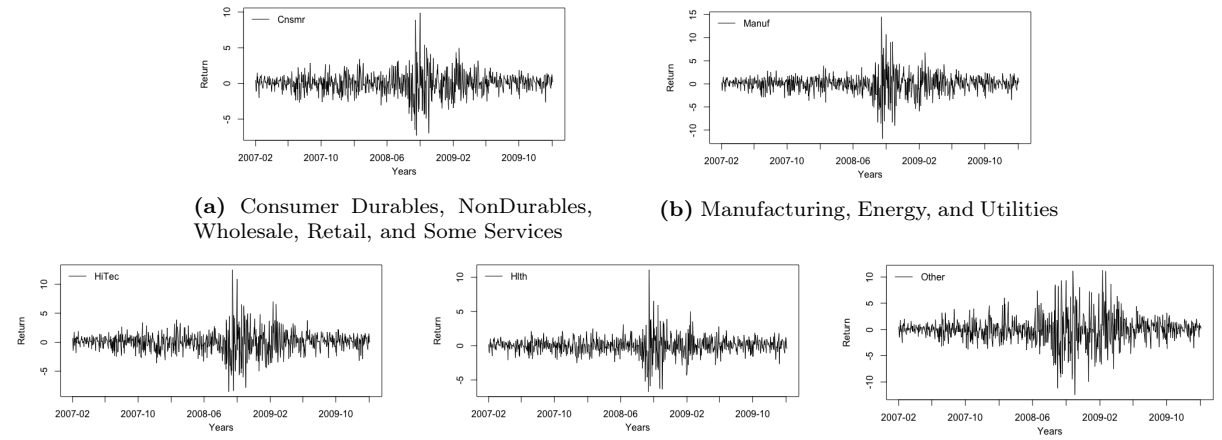
Under this setup we no longer sample $\boldsymbol{a}_{\boldsymbol{\phi}}$ directly. The gain that we achieve under (25) is that $\sigma_{a,\boldsymbol{\phi}}$ and $\xi_{a,i}$ are a priori independent. Thus in cases of uninformative data regarding $\boldsymbol{a}_{\boldsymbol{\phi}}$, there will be no challenging prior dependencies that shape the posterior. The difficulty with applying this setup in the multivariate random coefficient GAS-t model is that if the data actually is informative regarding the autoregressive parameters the noncentered parameterization has the opposite effect and induces strong posterior correlation between $\sigma_{a,\boldsymbol{\phi}}$ and $\xi_{a,i}$ and the centered parameterization (24) should be the preferred choice. In practice it shows that the degree to which the data is informative regarding the autoregressive angle parameters varies greatly among the angles and needs to be deduced by trial and error. In the moderate dimensional $k = 5$ application presented in the next section, the centered parameterization sufficed. For the larger $k = 10$ application, the noncentered parameterization proved to allow for more efficient sampling.

An alternative solution that can be explored in future research is to consider more advanced extension of HMC that can cope with local variation in curvature such as Riemannian HMC (RHMC) (Girolami et al., 2009), which was shown highly effective in hierarchical models by Betancourt & Girolami (2015). RHMC might however by too computationally demanding for high-dimensional GAS models and approximating versions such as proposed by Burda & Maheu (2013) or Zhang & Sutton (2011) might be more appropriate.

4.3.5 Application to 5 Industry Portfolios

I estimate the multivariate random coefficient GAS-t (RC-GAS-t) model, the restricted multivariate GAS-t (s-GAS-t) model, which has all angle factors driven by the same scalars, and the fully parameterized multivariate GAS-t model on the set of 5 industry stock portfolio daily return series for the period 2007-02-27 until 2010-03-02 obtained from the Kenneth French website (source: `http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`, see Figure 6).

To facilitate model comparison I specify weakly informative priors on all autoregressive parameters and the hyperparameters in (24). This implies a $\mathcal{N}(0.05, 0.5^2)$ prior for the nonhierarchical nonzero elements of $\boldsymbol{A}$ for all three models and for the $\mu_{a,\boldsymbol{\phi}}$ hyperparameter, and a truncated $\mathcal{N}(0.95, 0.5^2)I_{[-1<b.<1]}$ prior for the nonhierarchical nonzero elements of $\boldsymbol{B}$ for all three models and for the $\mu_{b,\boldsymbol{\phi}}$ hyperparameter. Following Gelman (2006) and Polson & Scott (2012), I use a weakly informative half-Cauchy prior with the scale parameter set to 0.25 for the hyper variance parameters $\sigma_{a,\boldsymbol{\phi}}$ and $\sigma_{b,\boldsymbol{\phi}}$. Since the inverse degrees of freedom parameter is shared among all three models a diffuse prior on the interval $(0, 1/2)$ is admissible. The constraints for stationarity and finite variance of the $\boldsymbol{y}_t$ are enforced using the approach by Neal (2011) described in Section 3.1.3. For all three models a 1000 iteration warm up run is used to estimate a diagonal mass matrix and tune the step size and the number of Leapfrog steps, after which 40,000 draws were produced. The run time was approximately 8 hours for the unrestricted GAS-t and RC-GAS-t models and 2 hours for



(a) Consumer Durables, NonDurables, Wholesale, Retail, and Some Services

(b) Manufacturing, Energy, and Utilities

(c) Business Equipment, Telephone and Television Transmission

(d) Healthcare, Medical Equipment, and Drugs

(e) Other

Figure 6: Daily returns on 5 Industry Portfolios for the period 2007-02-28 to 2010-03-02.

the s-GAS-t model. Resulting estimates of posterior means and variances are reported in Table 10.

The standard error (SE) estimates reveal that particularly for several of the $\boldsymbol{b}_{\boldsymbol{\phi}}$ parameters in the unrestricted model there is considerable parameter uncertainty. Similar as for the DPMP models the large uncertainty in b parameters can be traced back to limited evidence in support for time-variance of several of the angle parameters causing identification issues for the b parameters. The SEs are generally the largest for the $\boldsymbol{b}_{\boldsymbol{\phi}}$ parameters for which the 95% HPD regions of the corresponding $\boldsymbol{a}_{\boldsymbol{\phi}}$ parameters contain zero. This carries over in odd estimates for the means, such as 0.484 for $b_{5,\boldsymbol{\phi}}$, which are far removed from the average estimates for $\boldsymbol{b}_{\boldsymbol{\phi}}$ produced by the other two models. In the RC-GAS-t model, the $\boldsymbol{b}_{\boldsymbol{\phi}}$ parameters clearly benefit from information pooling as the estimates are all close to the hyper mean parameter $\mu_{b,\boldsymbol{\phi}}$, leading to substantial reduction in parameter uncertainty relative to the unrestricted model. For the $\boldsymbol{a}_{\boldsymbol{\phi}}$ parameters in the unrestricted model the SEs are smaller. As expected this is also reflected in more variation in the estimates of $\boldsymbol{a}_{\boldsymbol{\phi}}$ in the RC-GAS-t model, as the data is more informative regarding these parameters.

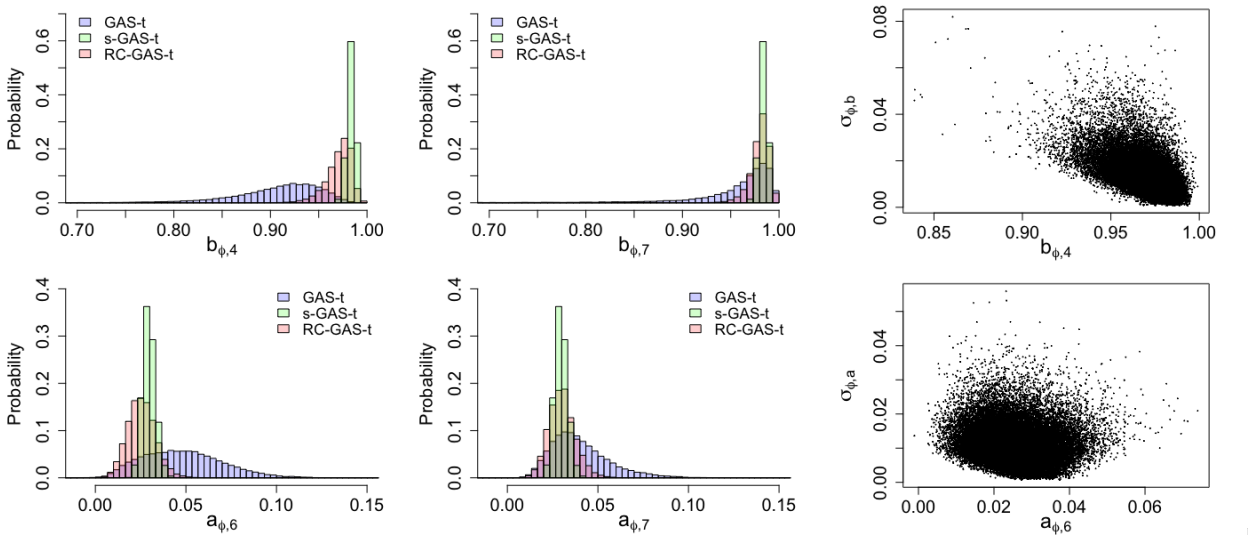
The parameters in the restricted s-GAS-t model are the most precise in terms of both posterior standard deviation and in accuracy of the parameter estimates as reflected in the lower NSEs. The lower NSEs are in part also due to greater efficiency of the HMC algorithm for the s-GAS-t model. Because of the challenging variations in curvature for the posteriors of the unrestricted GAS-t and the RC-GAS-t models HMC is challenging to tune correctly and the resulting Markov chains for these two models still have non-negligible autocorrelation. The variation in curvature poses an additional challenge for HMC because of the fixed step size. The step size might need to be significantly lower in order to allow the algorithm to adequately explore locations of high posterior curvature, relative to the step size required to obtain a targeted acceptance rate between 0.6 and 0.8 (Betancourt & Girolami, 2015). To account for possible regions of high posterior curvature, I set the step size conservatively in the unrestricted GAS-t and RC-GAS-t model. The result is notably high acceptance rates for these chains of around 0.9.

The differences in parameter uncertainty and flexibility in location between the three models are further illustrated in Figure 7a. The RC-GAS-t posterior marginal probability mass is consid-

Table 5: Parameter estimates for the restricted multivariate scalar GAS-t model (s-GAS-t), the fully parameterized multivariate GAS-t model (GAS-t) and the multivariate random coefficient GAS-t model (RC-GAS-t)

	s-GAS-t			GAS-t			RC-GAS-t		
$\boldsymbol{\theta}$	$\mathrm{E}(\cdot\|\boldsymbol{Y}_T)$	$\sqrt{\mathrm{var}(\cdot\|\boldsymbol{Y}_T)}$	NSE ($\times 10^{-2}$)	$\mathrm{E}(\cdot\|\boldsymbol{Y}_T)$	$\sqrt{\mathrm{var}(\cdot\|\boldsymbol{Y}_T)}$	NSE ($\times 10^{-2}$)	$\mathrm{E}(\cdot\|\boldsymbol{Y}_T)$	$\sqrt{\mathrm{var}(\cdot\|\boldsymbol{Y}_T)}$	NSE ($\times 10^{-2}$)
a_1	0.073	0.009	0.015	0.077	0.009	0.015	0.074	0.009	0.018
a_2	0.072	0.011	0.009	0.079	0.011	0.017	0.075	0.011	0.021
a_3	0.072	0.010	0.010	0.078	0.010	0.016	0.074	0.010	0.018
a_4	0.067	0.010	0.009	0.069	0.010	0.014	0.067	0.010	0.015
a_5	0.089	0.010	0.008	0.095	0.011	0.015	0.092	0.011	0.018
$a_{1,\phi}$	0.030	0.004	0.009	0.049	0.009	0.013	0.041	0.007	0.012
$a_{2,\phi}$				0.050	0.010	0.012	0.037	0.008	0.014
$a_{3,\phi}$				0.028	0.013	0.017	0.028	0.008	0.015
$a_{4,\phi}$				0.065	0.017	0.022	0.034	0.009	0.017
$a_{5,\phi}$				0.039*	0.025	0.037	0.023**	0.009	0.013
$a_{6,\phi}$				0.048	0.022	0.029	0.026	0.008	0.013
$a_{7,\phi}$				0.040	0.016	0.033	0.030	0.007	0.011
$a_{8,\phi}$				0.021	0.011	0.019	0.025	0.008	0.012
$a_{9,\phi}$				0.031*	0.025	0.036	0.027	0.009	0.012
$a_{10,\phi}$				-0.006*	0.027	0.038	0.020**	0.009	0.015
b_1	0.981	0.005	0.006	0.978	0.006	0.010	0.980	0.006	0.011
b_2	0.977	0.006	0.007	0.976	0.006	0.010	0.977	0.006	0.013
b_3	0.981	0.005	0.006	0.980	0.006	0.009	0.981	0.005	0.010
b_4	0.979	0.007	0.006	0.977	0.007	0.011	0.978	0.007	0.012
b_5	0.985	0.004	0.004	0.984	0.005	0.007	0.985	0.005	0.009
$b_{1,\phi}$	0.984	0.004	0.003	0.989	0.008	0.012	0.989	0.006	0.009
$b_{2,\phi}$				0.956	0.019	0.026	0.975	0.010	0.024
$b_{3,\phi}$				0.986	0.012	0.020	0.984	0.008	0.015
$b_{4,\phi}$				0.906	0.050	0.084	0.971	0.013	0.035
$b_{5,\phi}$				0.484	0.274	0.806	0.971	0.015	0.040
$b_{6,\phi}$				0.767	0.186	0.357	0.975	0.014	0.039
$b_{7,\phi}$				0.946	0.080	0.397	0.980	0.010	0.018
$b_{8,\phi}$				0.989	0.016	0.042	0.986	0.009	0.017
$b_{9,\phi}$				0.708	0.290	1.616	0.975	0.014	0.031
$b_{10,\phi}$				0.515	0.315	0.758	0.976	0.014	0.029
$\bar{\nu}$	0.109	0.013	0.010	0.102	0.013	0.013	0.105	0.013	0.015
$\mu_{a,\phi}$							0.029	0.005	0.010
$\mu_{b,\phi}$							0.978	0.008	0.021
$\sigma_{a,\phi}$							0.011	0.006	0.011
$\sigma_{b,\phi}$							0.012	0.008	0.024
Leapfrog steps		4			20			20	
Step size		0.4			0.05			0.05	
Accept rate		0.88			0.92			0.90	

Notes: Estimation results for the parameters of the multivariate s-GAS-t, GAS-t and RC-GAS-t models based on a $40,000$ draw long Markov chain produced using Hamiltonian Monte Carlo. Reported for all three models and for all parameters are posterior mean ($\mathrm{E}(\cdot|\boldsymbol{Y}_T)$), standard deviation ($\sqrt{\mathrm{var}(\cdot|\boldsymbol{Y}_T)}$) and numerical standard error (NSE). For the s-GAS-t model only the first autoregressive angle parameters are reported, since all others are restricted to be of the the same value as the first in their set. For the $a_\cdot$ parameters one star (*) is used to denote that the 95% highest posterior density region (HPD) includes zero and two stars (**) are used to denote that the 99% HPD includes zero. Also reported for all three Markov chains are the number of Leapfrog integrator steps, the integrator step size and the acceptance rates.



(a) Histograms of Marginal Posteriors **(b)** Bivariate posteriors

Figure 7: Histograms of the marginal posteriors of a selection of autoregressive angle parameters for all three models - the unrestricted (GAS-t), the scalar (s-GAS-t) and the random coefficient (RC-GAS-t) multivariate GAS-t models - are shown on the left. The joint distributions of the hyper variance parameter and one of their corresponding autoregressive angle parameters from the RC-GAS-t model are displayed on the right. Both plots are based on 40,000 draw samples generated using Hamiltonian Monte Carlo.

erably more concentrated than the unrestricted GAS-t model while still allowing the distribution of the autoregressive angle parameters to vary in location and dispersion. Particularly if the data is informative, such as seen in the top left histogram where the unrestricted model reveals that there is weak information in the data that the mode of $b_{\phi,4}$ is below the average of the $\boldsymbol{b}_\phi$, which is then reflected in the posterior of $b_{\phi,4}$ for the RC-GAS-t model by the probability mass being shifted to values slightly below the average. It is worth mentioning that the histograms in Figure 7a are actually cut off to aid comparison, but are therefore not fully reflective of how heavy tailed some of the marginals for the autoregressive angle parameters in the the unrestricted model are. For $b_{\phi,7}$ for instance the sampled range is actually from 0.14 to 1 and its sample kurtosis is 26.5.

Figure 7 presents the joint distributions of the hyper variance parameters and the autoregressive angle parameters $b_{\phi,4}$ and $a_{\phi,6}$. The plots show that the distributions narrow for low values of the variance parameter and widen significantly for higher values of the variance, which is the known "funnel" characteristic of hyper variance parameters (Betancourt & Girolami, 2015), which is indicative of large variations in posterior curvature. This effect is most prominent for the $b_{\phi,5}$, $b_{\phi,6}$, $b_{\phi,9}$ and $b_{\phi,10}$ parameters, suggesting that the sampling efficiency for these parameters might improve under the noncentered parameterization (25). Moreover, Figure 7b highlights that there there is considerable posterior mass close to zero for the variance parameter. This is to be expected if the data is not very informative for certain parameters in the group and it justifies the choice for a half-Cauchy prior, in favor of the popular inverse-Gamma prior, because of its favorable behavior for values of the variance parameter close to zero Gelman (2006).

The log BFs comparing the three models are reported in Table 6. The bridge sampling method with the warp 3 transformation to account for skewness is again used to compute the marginal

Table 6: The log Bayes Factors ($\text{BF}_{\cdot|\cdot}$) comparing the multivariate random coefficient GAS-t model (model 1), the restricted multivariate scalar GAS-t model (model 2) and the fully parameterized multivariate GAS-t model (model 3)

$2\log \text{BF}_{2\|3}$	$2\log \text{BF}_{1\|3}$	$2\log \text{BF}_{1\|2}$
71.2	74.8	3.6

likelihoods. The log BFs suggest the evidence favors the hierarchical RC-GAS-t model over the other two. The large parameter uncertainty present in the unrestricted model is detrimental to its performance in this comparison as it is clearly dominated by the other two models. The evidence in favor of the RC-GAS-t over the restricted s-GAS-t model is less compelling. The gain from the smaller parameter uncertainty in the s-GAS-t model is thus only barely outweighed by the increased flexibility of the RC-GAS-t. To get a sense of the accuracy of the marginal likelihood estimates, the warped bridge sampling procedure was run another 5 times, but the results were consistent up to one decimal place.

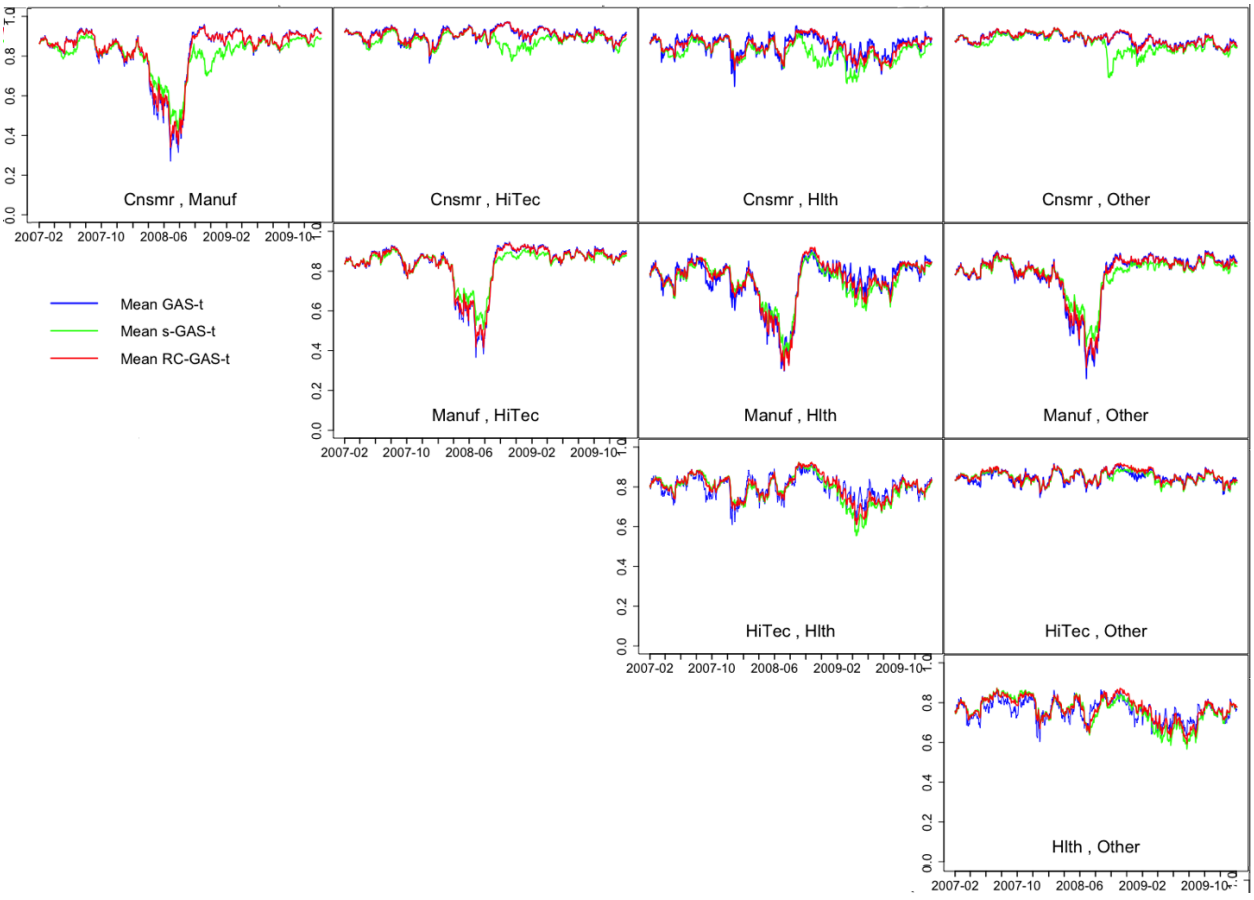
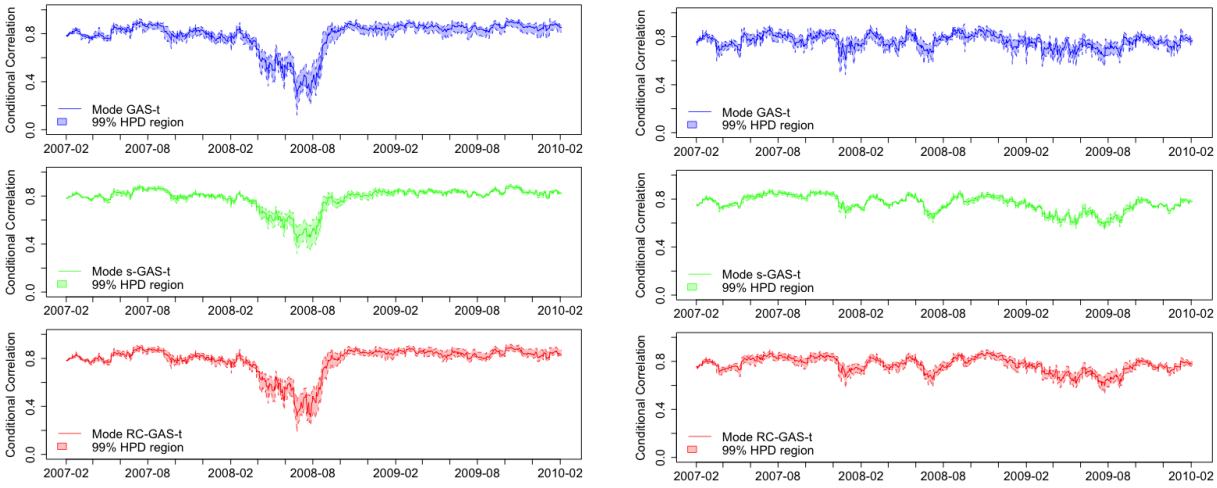


Figure 8: Plots of the mean conditional correlations for a set of 5 industry portfolios for the period 2007-02-27 until 2010-03-02, as predicted by the unrestricted (GAS-t), the scalar (s-GAS-t) and the random coefficient (RC-GAS-t) multivariate GAS-t models. The mean estimates are based on a $40,000$ draw sample from a Hamiltonian Monte Carlo Markov chain.

Another indication of how well the models fit the data is to consider the conditional correla-



(a) Manuf & Other portfolios

(b) Hlth & Other portfolios

Figure 9: The 99% HPD regions of the conditional correlations between the Manufacturing and Other industry portfolio returns (left) and Health and Other industry portfolio returns as predicted by the unrestricted (GAS-t), the scalar (s-GAS-t) and the random coefficient (RC-GAS-t) multivariate GAS-t models, for the period 2007-02-27 until 2010-03-02. The HPD regions are based on a 40,000 draw sample from a Hamiltonian Monte Carlo Markov chain.

tions that the three models predict. Figure 8 displays the mean predicted correlations of for all three models and Figure 9 shows the 99% HPD regions for the conditional covariances between the Manufacturing and Other industry portfolio returns and between the Health and Other portfolio returns. Figure 8 highlights that the parameter restrictions for the s-GAS-t model results do appear to restrict the models ability to capture the time-variation in the correlations, which is particularly evident from the predicted volatilities for the Consumer industry portfolios. The mean predicted correlations for the RC-GAS-t and unrestricted GAS-t are very similar and hard to distinguish from one another apart from the fact that the correlation estimates for the latter appear slightly more noisy for the latter.

The mean estimates however hide the differences in uncertainty associated with the predicted correlations. The HPD regions confirm expectations that the predicted correlations are the most uncertain for the unrestricted model and that there is the least uncertainty for the restricted GAS-t. The RC-GAS-t falls in between. The very large parameter uncertainty for the unrestricted GAS-t does however not seem to carry over fully into uncertainty for the conditional correlations. Most plausibly this due to the the majority of the extreme parameter uncertainty in the unrestricted GAS-t model being related to the identification issue for several of the $\boldsymbol{b}_{\boldsymbol{\phi}}$ parameters resulting from the corresponding $\boldsymbol{a}_{\boldsymbol{\phi}}$ parameters having considerable posterior mass near 0. This type of uncertainty in the autoregressive angle parameters has little implications for the uncertainty in the angles however, similarly as for the application in Section 4.2 the large uncertainty for the b parameter corresponding to the upgrade factor did not translate into uncertainty for the intensity of the upgrade process.

4.3.6 Application to 10 Industry Portfolios

The 5 asset case is useful in that it allows all three models to be compared while still being large enough to illustrate the shortcomings of both the unrestricted and the restricted GAS-t covariance models. In this case, model probabilities already support the notion that the hierarchical model

strikes a favorable balance between these two models. However the true value of hierarchical modeling increases with the complexity of the models and to illustrate this I consider the 10 asset case. A set of 10 industry portfolios is again obtained from Kenneth French's website and the models are estimated on daily returns covering the same period as for the 5 industry portfolios.

The unrestricted model turns unmanageable for this number of assets, so I focus on the s-GAS-t and the RC-GAS-t models. Prior specifications are the same as in the 5 asset case apart from the hierarchical prior setup which is non-centered, as in (25). HMC is used to obtain $20,000$ draws from the posterior, requiring roughly 15 and 30 hours of computational time for the s-GAS-t and RC-GAS-t respectively. Although the MCMC sample size might be smaller, the per draw efficiency of HMC is much improved relative to the 5 asset case by the fact that it was possible to use dense mass matrices for these 10 asset models. The resulting ESSs all exceed 6000.

Rather than present parameter estimation results, which becomes cumbersome for models with this many parameters (a summary of parameter estimation results is included in Appendix E), I focus on model comparison and the estimates of conditional correlations. Double the log Bayes factor of 13.6 suggests that the strength of evidence in favor of the hierarchical model increased considerably with the number of assets. Intuitively this also makes sense, since as the number of assets increases it is more likely that there are more combinations of assets for which the correlations follow different dynamics. Upon close inspection this can also be deduced from the mean correlation plots for the 10 assets in Figure 10. Significant differences in predicted correlations between the models are noticeably more numerous compared to those for the 5 assets in Figure 8.

Both in the 5 and 10 asset applications the hierarchical model proves superior. However, considering the notably greater ease in estimating the s-GAS-t, the restricted model continues to be a valid option when the number of assets is not too large, but its performance deteriorates relative to the hierarchical model as larger number of assets are considered. Since the advantages of the hierarchical specification prove to increase with dimension, it would be interesting for future research to consider simpler multivariate covariance models, such as the Hadamard DCC, with hierarchical prior specifications since the computational costs will not be as extreme as for the GAS-t model. The reverse mode gradient computation approach presented in Section 4.3.3 transfers straightforwardly to multivariate GARCH models and should allow for similar computational speed ups. Given that the matrix operations in these models are with matrices of at most dimension $k \times k$, suggests that much larger models of potentially up to 50 assets can be estimated in a similar time frame as the 10 asset GAS-t models. For datasets of such size, additional levels in the hierarchical setup might also be worth exploring. Grouping autoregressive parameters for correlations based on firm characteristics of the asset pairs such as size, value and industry would be obvious extensions if the model is applied to individual stocks.

Apart from the differences in predictions by the different models, it is also interesting to observe the distinct dip in the mean predicted correlations for the Energy industry portfolio, which is perhaps the most visually obvious pattern in Figure 10. In the 5 portfolio case we see a similar pattern for the Manufacturing industry portfolio which subsumes the Energy portfolio under this broader industry specification. The dip in the correlation initializes just prior to the 2008 financial crisis impacting the stock markets and is due to a strong rise in oil prices during late 2007 and the first half of 2008 (Hamilton, 2009). Oil industry related stocks are naturally strongly linked to oil prices. Energy industry stocks hence saw strong growth during this period, whereas stocks in the other industries were not similarly affected. Once the financial crisis hit stocks across all industries around

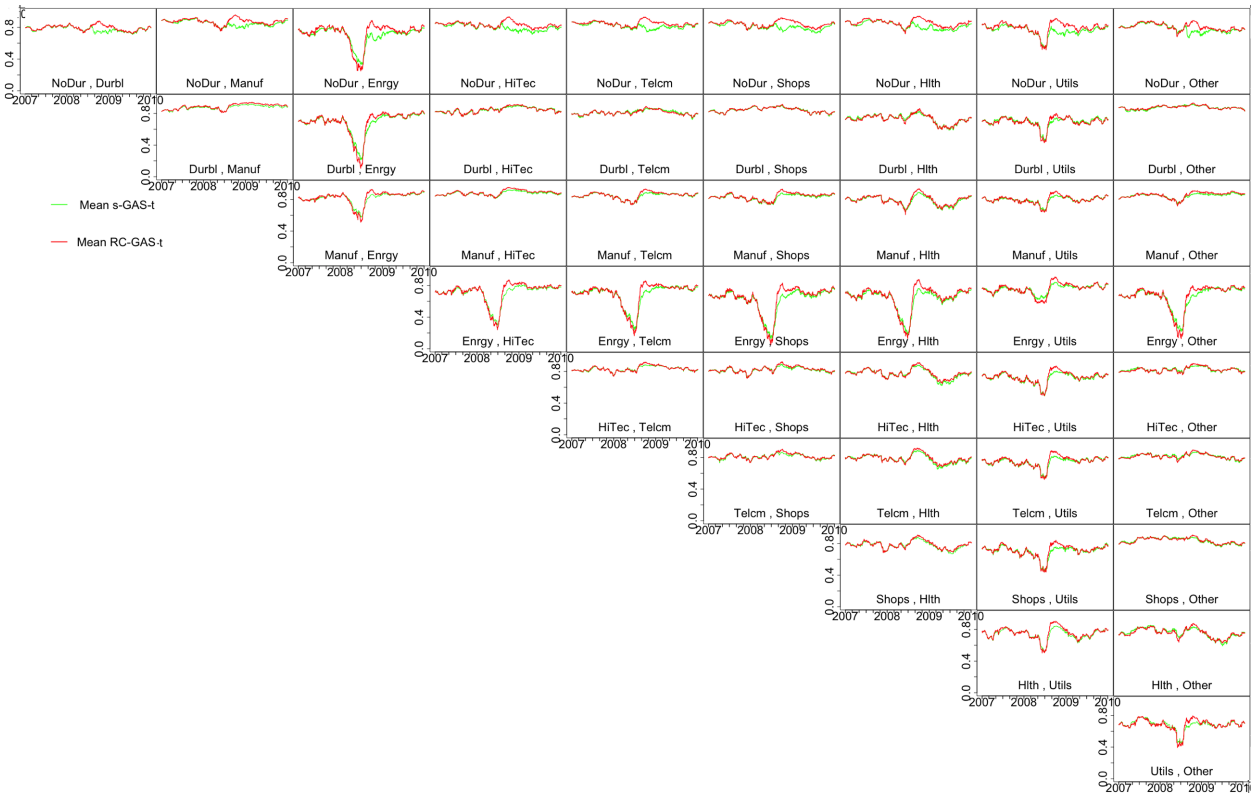


Figure 10: Plots of the mean conditional correlations for a set of 10 industry portfolios for the period 2007-02-27 until 2010-03-02, as predicted by the scalar (s-GAS-t) and the random coefficient (RC-GAS-t) multivariate GAS-t models. The mean estimates are based on a 20,000 draw sample from a Hamiltonian Monte Carlo Markov chain.

July of 2008, correlations between the Energy portfolio and other portfolios rapidly increased. Increased correlations between markets during economic downturns is a common empirical finding and that such was in particular true for the oil sector during the 2008 financial crisis is for instance also found by Filis et al. (2011).

5 Discussion

In the introduction I made the case for Bayesian inference for GAS models on the grounds of four arguments. First, the high degree of nonlinearity in how the parameters enter the likelihood might cause larger samples sizes than typically expected to be required to achieve satisfactory convergence to the normal distribution for the parameters. This is perhaps best illustrated in the Bayesian analysis of the Beta-Gen-t-EGARCH model of Harvey & Lange (2017), where the flexibility afforded by an additional shape parameter is shown to cause considerable uncertainty in the form of skewness and kurtosis for volatility predictions following large absolute returns in spite of a relatively large sample size of 5 years of daily data.

Second, in a Bayesian framework it is much easier to make probabilistic statements regarding quantities of interest that are complex functions of the estimated parameters. Throughout all three empirical applications, that benefit is thoroughly utilized and displayed through the HPD region plots of the time-varying quantities of interest, namely the predicted volatilities, intensities and

correlations. Since the primary quantities of interest in GAS models in nearly all cases involve functions of the time-varying parameters, the HPD plots are useful tools for visualizing the uncertainty associated with these quantities. They also greatly aid in comparing the uncertainty in the predictions made by different GAS model specifications, such as illustrated for the Beta-Gen-t-EGARCH relative to the Beta-t-EGARCH and for the different parameterization of the multivariate GAS-t model. Similar confidence bounds on the predictions would not only be considerably more challenging to obtain in a frequentist setting, they would also completely ignore any residual non-normality. The notion of inferences on non-linear functions of the parameters is further proved useful in Section 4.1.2 for the purpose of testing for time-variance in instances where time-variance is governed by more than one factor.

Third, Bayesian model comparison allows for comparison of non-nested models. In section 2 it is emphasized why this is particularly useful in the case of GAS models as non-nested models that researcher are likely to want to compare, arise naturally as a result of the options for different scaling matrices, link functions and factorizations. In Section 4.2 this technique is extensively applied to test hypothesis that have not been previously considered in the literature regarding the favorability of several scaling matrices and factorization in the existing class of dynamic pooled marked point process factor models of Creal et al. (2013). In line with expectations, the evidence favored scaling matrices that incorporated second order information and a three factor model over one and two factor models for the modeling of intensities in credit rating transitions. In Section 4.3 the technique also proved useful in finding evidence in favor of the hierarchical covariance model.

Fourth, I argued that hierarchical prior specifications can serve to restrain parameter uncertainty as larger GAS models with greater number of time-varying parameters start being explored. Hierarchical priors will generally provide a more elegant and flexible approach to coping with parameter uncertainty in large GAS models relative to the common approach of imposing restriction on the parameters. In Section 4.3 I applied a hierarchical prior setup to a subset of the autoregressive parameters in the multivariate GAS-t model of Creal et al. (2011b). I showed how the technique enabled inference for a multivariate GAS-t model for a set of 10 industry portfolios, which is considerably more as most preceding published work has considered for covariance models of similar flexibility. Bayes factors also provided evidence of superior performance of the hierarchical model relative to restricted and unrestricted versions of the multivariate GAS-t model.

The third and fourth argument warrant some further discussion as they make for potentially exciting areas (particularly hierarchical GAS models) for future work. Section 4.2 showed that for relatively simple GAS models even one of the simplest MCMC methods (RW-MH sampler) was able to produce a sufficiently large sample from the posterior within minutes. Although this is obviously far form as fast as ML, which delivers results in under a second in most cases, the time frame is nevertheless very reasonable. Since, the Bayesian approach has the obvious advantage of enabling the making of very common decisions in the GAS framework, such as which link function or scaling matrix to use - which can not be formally considered when using ML estimation - the extra computational cost seem like a small price to pay. Even for practitioners the need for inferences in much under a minute seems uncommon and the Bayesian approach can thus also be of benefit for this group of GAS model users in improving modeling choices.

Further research might however be needed to establish what are suitable priors for the autoregressive parameters in GAS models to facilitate objective model comparison. Although, a sensitivity analysis is used to asses the impact of the prior variance, the number of prior specification options

extends far beyond the degree of diffuseness. I have for instance taken the approach of using normal or truncated normal priors, but log-normal and uniform priors have alternatively been applied for the autoregressive parameters in GARCH models (see e.g. Miazhynskaia & Dorffner (2006) and Asai (2006)).

Another avenue for further exploration would be to use Bayesian posterior model probabilities to provide a natural way to cope with model uncertainty in the GAS framework. In this thesis I have focused on the impact of parameter uncertainty in comparing two models, but the Bayesian model comparison approach is easily extended to apply to multiple models simultaneously. The resulting Bayes factors are straightforwardly transformed to posterior model probabilities and can be applied for instance in Bayesian model averaging to combine models to produce a single forecast. This might be a more natural way of dealing with the model uncertainty induced by the many different specification options in the GAS framework than simply choosing one model as the preferred choice.

When it comes to hierarchical priors, there are a great many different ways in which they can be applied to start enabling inference in more complex GAS models as previously accessible. One example would be to alleviate the i.i.d assumption across the rating transition risk processes of firms in the DPMP models. This assumption is in place because there are far too little rating transition events per company to model separate dynamics for each company. Using similar hierarchical normal priors on the autoregressive parameter as in the covariance GAS-t model and possibly grouping firms based on firm characteristics it should be possible to start separating the dynamics per firm to a certain degree. Furthermore, different types of hierarchical priors could be considered such as a spike-and-slab prior. Typical extension of the time-varying parameter process as presented in this thesis include adding exogenous variables that enter the update equation (2) with regression coefficients, or allowing for additional lags of either $\boldsymbol{s}_t$ and $\boldsymbol{f}_t$. In both of these cases the coefficients of these additional variables could be specified with a spike-and-slab prior. This has similarly been explored in state space (latent factor) models where a set of exogenous variables enters a time-varying parameter equation by means of a spike-and-slab regression (Scott & Varian, 2014). This would be an interesting extension of the DPMP models as well, since large sets of macroeconomic variables have already proved valuable predictors of the intensities of credit risk processes (Duffie et al., 2009).

What will be more challenging than coming up with interesting new applications for hierarchical priors in GAS models, will be to figure out efficient ways of implementing them. As seen in the multivariate GAS-t applications, the larger size of the models that become accessible through using Bayesian hierarchical modeling, causes us to run into the limits of currently available computational resources. The model presented in Section 4.3 already required innovation in the way that gradients for GAS models are computed, but computation time is still substantial. For other hierarchical models similar efficient approaches are likely needed since, in order to exploit the full potential of hierarchical models, a certain degree of model complexity is simply necessary. To facilitate inference in such models it might also be worthwhile to explore more advanced or specialized MCMC algorithms.

References

Ardia, D. (2008). Financial risk management with Bayesian estimation of GARCH models. *Lecture Notes in Economics and Mathematical Systems*, *612*.

Ardia, D., Boudt, K., & Catania, L. (2016). Generalized autoregressive score models in R: The GAS package [Working paper].

Ardia, D., & Hoogerheide, L. F. (2010). Efficient bayesian estimation and combination of GARCH-type models [MPRA Paper]. (22919).

Ardia, D., Hoogerheide, L. F., & van Dijk, H. K. (2009). Adaptive mixture of Student-t distributions as a flexible candidate distribution for efficient simulation: The R package AdMit. *Journal of Statistical Software*, *29*(3), 1-32.

Asai, M. (2006). Comparison of mcmc methods for estimating garch models. *Journal of the Japan Statistical Society*, *36*(2), 199–212.

Bartels, M., & Ziegelmann, F. A. (2016). Market risk forecasting for high dimensional portfolios via factor copulas with GAS dynamics. *Insurance: Mathematics and Economics*, *70*, 66–79.

Bauwens, L., Grigoryeva, L., & Ortega, J. (2016). Estimation and empirical performance of non-scalar dynamic conditional correlation models. *Computational Statistics & Data Analysis*, *100*, 17–36.

Bauwens, L., & Hautsch, N. (2006). Stochastic conditional intensity processes. *Journal of Financial Econometrics*, *4*(3), 450–493.

Bauwens, L., & Hautsch, N. (2009). Modelling financial high frequency data using point processes. In T. Mikosch, J. Kreiß, R. Davis, & T. Andersen (Eds.), *Handbook of financial time series* (pp. 953–979). Springer.

Bauwens, L., Laurent, S., & Rombouts, J. V. (2006). Multivariate GARCH models: a survey. *Journal of Applied Econometrics*, *21*(1), 79–109.

Bauwens, L., & Lubrano, M. (1998). Bayesian inference on GARCH models using the Gibbs sampler. *The Econometrics Journal*, *1*(1), 23–46.

Betancourt, M. (2013). A general metric for Riemannian manifold Hamiltonian Monte Carlo. In F. Nielsen & F. Barbaresco (Eds.), *Geometric science of information* (pp. 327–334). Springer.

Betancourt, M. (2017). A conceptual introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434*.

Betancourt, M., Byrne, S., & Girolami, M. (2014). Optimizing the integrator step size for Hamiltonian Monte Carlo. *arXiv preprint arXiv:1411.6669*.

Betancourt, M., Byrne, S., Livingstone, S., & Girolami, M. (2017). The geometric foundations of Hamiltonian Monte Carlo. *Bernoulli*, *23*(4A), 2257–2298.

Betancourt, M., & Girolami, M. (2015). Hamiltonian Monte Carlo for hierarchical models. In S. K. Upadhyay, U. Singh, D. K. Dey, & A. Loganathan (Eds.), *Current trends in Bayesian methodology with applications* (Vol. 79, p. 80-100). Chapman and Hall/CRC.

Blasques, F., Koopman, S. J., & Lucas, A. (2014). Stationarity and ergodicity of univariate generalized autoregressive score processes. *Electronic Journal of Statistics*, *8*(1), 1088–1112.

Blasques, F., Lucas, A., & Silde, E. (2016). A stochastic recurrence equations approach for score driven correlation models. *Econometric Reviews*, 1–16.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics*, *31*(3), 307–327.

Bollerslev, T., Engle, R. F., & Wooldridge, J. M. (1988). A capital asset pricing model with time-varying covariances. *Journal of Political Economy*, *96*(1), 116–131.

Browne, W. J., & Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis*, *1*(3), 473–514.

Brownlees, C. T. (2015). Hierarchical GARCH. Retrieved from `https://ssrn.com/abstract=1695649`

Burda, M. (2015). Constrained Hamiltonian Monte Carlo in BEKK GARCH with targeting. *Journal of Time Series Econometrics*, *7*(1), 95–113.

Burda, M., & Maheu, J. M. (2013). Bayesian adaptively updated Hamiltonian Monte Carlo with an application to high-dimensional BEKK GARCH models. *Studies in Nonlinear Dynamics and Econometrics*, *17*(4), 345–372.

Caporin, M., & McAleer, M. (2012). Do we really need both BEKK and DCC? a tale of two multivariate GARCH models. *Journal of Economic Surveys*, *26*(4), 736–751.

Carlin, B. P., & Louis, T. A. (2000). *Bayes and empirical Bayes methods for data analysis* (Vol. 17). Chapman & Hall/CRC Boca Raton, FL.

Chib, S., & Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician*, *49*(4), 327–335.

Creal, D., Koopman, S. J., & Lucas, A. (2011a). A dynamic multivariate heavy-tailed model for time-varying volatilities and correlations. *Journal of Business & Economic Statistics*, *29*(4), 552–563.

Creal, D., Koopman, S. J., & Lucas, A. (2011b). Generalized autoregressive score models with applications [Working paper].

Creal, D., Koopman, S. J., & Lucas, A. (2013). Generalized autoregressive score models with applications. *Journal of Applied Econometrics*, *28*(5), 777–795.

Creal, D., Schwaab, B., Koopman, S. J., & Lucas, A. (2014). Observation-driven mixed-measurement dynamic factor models with an application to credit risk. *Review of Economics and Statistics*, *96*(5), 898–915.

Duane, S., Kennedy, A. D., Pendleton, B. J., & Roweth, D. (1987). Hybrid Monte Carlo. *Physics Letters B*, *195*(2), 216–222.

Duffie, D., Eckner, A., Horel, G., & Saita, L. (2009). Frailty correlated default. *The Journal of Finance*, *64*(5), 2089–2123.

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

Engle, R. F. (2007). High dimension dynamic correlations [NYU working paper]. Retrieved from `https://ssrn.com/abstract=1293628`

Engle, R. F., & Bollerslev, T. (1986). Modelling the persistence of conditional variances. *Econometric Reviews*, *5*(1), 1–50.

Engle, R. F., & Kroner, K. F. (1995). Multivariate simultaneous generalized ARCH. *Econometric Theory*, *11*(01), 122–150.

Engle, R. F., & Russell, J. R. (1998). Autoregressive conditional duration: a new model for irregularly spaced transaction data. *Econometrica*, *66*(5), 1127-1162.

Engle, R. F., Shephard, N., & Sheppard, K. (2008). Fitting vast dimensional time-varying covariance models [NYU working paper]. Retrieved from `https://ssrn.com/abstract=1354497`

Filis, G., Degiannakis, S., & Floros, C. (2011). Dynamic correlation between stock market and oil prices: The case of oil-importing and oil-exporting countries. *International Review of Financial Analysis*, *20*(3), 152–164.

Frühwirth-Schnatter, S. (2004). Estimating marginal likelihoods for mixture and Markov switching models using bridge sampling techniques. *The Econometrics Journal*, *7*(1), 143–167.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, *1*(3), 515–534.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2014). *Bayesian data analysis* (Vol. 2). Chapman & Hall/CRC Boca Raton, FL, USA.

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica*, *57*(6), 1317-1339.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In J. M. Bernardo, J. Berger, A. P. Dawid, & J. F. M. Smith (Eds.), *Bayesian statistics* (Vol. 196, pp. 169–193). University Press.

Giles, M. (2008). An extended collection of matrix derivative results for forward and reverse mode automatic differentiation. In C. Bischof, H. Bücker, P. Hovland, U. Naumann, & J. Utke (Eds.), *Advances in automatic differentiation: Lecture notes in computational science and engineering* (Vol. 64, p. 35-44). Springer.

Girolami, M., Calderhead, B., & Chin, S. A. (2009). Riemannian manifold Hamiltonian Monte Carlo. *arXiv preprint arXiv:0907.1100*.

Gower, R. M., & Gower, A. L. (2016). Higher-order reverse automatic differentiation with emphasis on the third-order. *Mathematical Programming*, *155*(1-2), 81–103.

Griewank, A. (1992). Achieving logarithmic growth of temporal and spatial complexity in reverse automatic differentiation. *Optimization Methods and software*, *1*(1), 35–54.

Griewank, A., & Walther, A. (2008). *Evaluating derivatives: principles and techniques of algorithmic differentiation*. SIAM.

Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., ... Steingroever, H. (2017). A tutorial on bridge sampling. *arXiv preprint arXiv:1703.05984*.

Hafner, C. M., & Franses, P. H. (2009). A generalized dynamic conditional correlation model: simulation and application to many assets. *Econometric Reviews*, *28*(6), 612–631.

Hamilton, J. D. (2009). Causes and consequences of the oil shock of 2007–08. *Brookings Papers on Economic Activity*(1), 215–261.

Hammersley, J., Handscomb, D., & Weiss, G. (1965). Monte Carlo methods. *Physics Today*, *18*, 55.

Harvey, A. (2010). Exponential conditional volatility models [Cambridge Working Papers in Economics]. Retrieved from `https://ideas.repec.org/p/cam/camdae/1040.html`

Harvey, A., & Chakravarty, T. (2008). Beta-t-(E)GARCH [Cambridge Working Papers in Economics]. Retrieved from `https://ideas.repec.org/p/cam/camdae/0840.html`

Harvey, A., & Lange, R. J. (2017). Volatility modeling with a Generalized t distribution. *Journal of Time Series Analysis*, *38*(2), 175–190.

Hascoet, L., & Pascual, V. (2013). The Tapenade automatic differentiation tool: principles, model, and specification. *ACM Transactions on Mathematical Software (TOMS)*, *39*(3).

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*(1), 97–109.

Hoogerheide, L. (2006). *Essays on neural network sampling methods and instrumental variables* (Vol. 253). Rozenberg Publishers.

Hwang, S., & Valls Pereira, P. L. (2006). Small sample properties of GARCH estimates and persistence. *The European Journal of Finance*, *12*(6-7), 473–494.

Jaeckel, P., & Rebonato, R. (1999). The most general methodology for creating a valid correlation matrix for risk management and option pricing purposes. *Journal of risk*, *2*(2), 17–28.

Karr, A. (1991). *Point processes and their statistical inference* (Vol. 7). CRC press.

Kass, R. E. (1993). Bayes factors in practice. *The Statistician*, 551–560.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795.

Koopman, S. J., Lucas, A., & Monteiro, A. (2008). The multi-state latent factor intensity model for credit rating transitions. *Journal of Econometrics*, *142*(1), 399–424.

Leimkuhler, B., & Reich, S. (2004). *Simulating Hamiltonian dynamics* (Vol. 14). Cambridge University Press.

Longin, F., & Solnik, B. (1995). Is the correlation in international equity returns constant: 1960–1990? *Journal of International Money and Finance*, *14*(1), 3–26.

Magnus, J. R., & Neudecker, H. (2007). *Matrix differential calculus with applications in statistics and econometrics* (Third ed.). John Wiley & Sons.

Meng, X. L., & Schilling, S. (2002). Warp bridge sampling. *Journal of Computational and Graphical Statistics*, *11*(3), 552–586.

Meng, X. L., & Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica*, 831–860.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, *21*(6), 1087–1092.

Miazhynskaia, T., & Dorffner, G. (2006). A comparison of Bayesian model selection based on mcmc with an application to garch-type models. *Statistical Papers*, *47*(4), 525–549.

Murray, I. (2016). Differentiation of the Cholesky decomposition. *arXiv preprint arXiv:1602.07527*.

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. In S. Brooks, A. Gelman, G. Jones, & X. L. Meng (Eds.), *Handbook of Markov chain Monte Carlo* (Vol. 2, pp. 113–162). CRC press.

Nelson, D. B. (1996). Asymptotic filtering theory for multivariate ARCH models. *Journal of Econometrics*, *71*(1), 1–47.

Pakman, A., & Paninski, L. (2014). Exact Hamiltonian Monte Carlo for truncated multivariate Gaussians. *Journal of Computational and Graphical Statistics*, *23*(2), 518–542.

Papaspiliopoulos, O., Roberts, G. O., & Sköld, M. (2007). A general framework for the parametrization of hierarchical models. *Statistical Science*, 59–73.

Polson, N. G., & Scott, J. G. (2012). On the half-Cauchy prior for a global scale parameter. *Bayesian Analysis*, *7*(4), 887–902.

Pourahmadi, M., & Wang, X. (2015). Distribution of random correlation matrices: Hyperspherical parameterization of the Cholesky factor. *Statistics & Probability Letters*, *106*, 5–12.

Ritter, C., & Tanner, M. A. (1992). Facilitating the Gibbs sampler: the Gibbs stopper and the Griddy-Gibbs sampler. *Journal of the American Statistical Association*, *87*(419), 861–868.

Ross, S. M. (2014). *Introduction to probability models*. Academic press.

Russell, J. R., & Engle, R. (1998). Econometric analysis of discrete-valued irregularly-spaced financial transactions data using a new autoregressive conditional multinomial model [CRSP working papers].

Scott, S. L., & Varian, H. R. (2014). Predicting the present with Bayesian structural time series. *International Journal of Mathematical Modelling and Numerical Optimisation*, *5*(1-2), 4–23.

Takaishi, T. (2007). Bayesian estimation of GARCH model by hybrid Monte Carlo. *arXiv preprint physics/0702240*.

Virbickaite, A., Ausín, M. C., & Galeano, P. (2015). Bayesian inference methods for univariate and multivariate GARCH models: a survey. *Journal of Economic Surveys*, *29*(1), 76–96.

Zhang, Y., & Sutton, C. A. (2011). Quasi-Newton methods for Markov chain Monte Carlo. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, & K. Weinberger (Eds.), *Advances in neural information processing systems* (pp. 2393–2401).

Appendices

A Derivatives

As a starting point for the derivatives of all log-likelihoods of GAS models we can use the derivatives presented in Creal et al. (2013). Let $p_t = p(\boldsymbol{y}_t|\boldsymbol{Y}_{t-1}, \boldsymbol{F}_t, \boldsymbol{\theta})$ for fixed $\boldsymbol{F}_t$ (i.e. not varying w.r.t. to $\boldsymbol{\theta}$). The gradient of ℓ_t w.r.t. $\boldsymbol{\theta}$ is then

$$\begin{aligned}
\frac{\partial \ell_t}{\partial \boldsymbol{\theta}'} &= \frac{\partial \log p_t}{\partial \boldsymbol{\theta}'} + \boldsymbol{\nabla}_t \frac{\partial \boldsymbol{f}_t}{\partial \boldsymbol{\theta}'}, \\
\frac{\partial \boldsymbol{f}_t}{\partial \boldsymbol{\theta}'} &= \frac{\partial \boldsymbol{\omega}}{\partial \boldsymbol{\theta}'} + \boldsymbol{A}\frac{\partial \boldsymbol{s}_{t-1}}{\partial \boldsymbol{\theta}'} + (\boldsymbol{s}_{t-1}' \otimes \boldsymbol{I}_n)\frac{\partial \operatorname{vec}(\boldsymbol{A})}{\partial \boldsymbol{\theta}'} + \boldsymbol{B}\frac{\partial \boldsymbol{f}_{t-1}}{\partial \boldsymbol{\theta}'} + (\boldsymbol{f}_{t-1}' \otimes \boldsymbol{I}_n)\frac{\partial \operatorname{vec}(\boldsymbol{B})}{\partial \boldsymbol{\theta}'}. \\
\frac{\partial \boldsymbol{s}_{t-1}}{\partial \boldsymbol{\theta}'} &= \boldsymbol{S}_{t-1}\frac{\partial \boldsymbol{\nabla}_{t-1}}{\partial \boldsymbol{\theta}'} + (\boldsymbol{\nabla}_{t-1}' \otimes \boldsymbol{I}_n)\frac{\partial \operatorname{vec}(\boldsymbol{S}_{t-1})}{\partial \boldsymbol{\theta}'},
\end{aligned}$$

where I follow the notational convention used in Magnus & Neudecker (2007) to represent derivatives of matrix functions and/or w.r.t. matrices by vectorizing their differentials. The elements in the above derivatives that need to be derived individually for each GAS model specification are $\partial \log p_t/\partial \boldsymbol{\theta}'$ for the distribution specific parameters in $\boldsymbol{\theta}$, $\partial \boldsymbol{\nabla}_{t-1}/\partial \boldsymbol{\theta}'$ and $\partial \operatorname{vec}(\boldsymbol{S}_{t-1})/\partial \boldsymbol{\theta}'$. The last two are partitioned further as

$$\begin{aligned}
\frac{\partial \boldsymbol{\nabla}_{t-1}}{\partial \boldsymbol{\theta}'} &= \frac{\partial \boldsymbol{\nabla}_{t-1}^*}{\partial \boldsymbol{\theta}'} + \frac{\partial \boldsymbol{\nabla}_{t-1}}{\partial \boldsymbol{f}_{t-1}'}\frac{\partial \boldsymbol{f}_{t-1}}{\partial \boldsymbol{\theta}'}. \\
\frac{\partial \operatorname{vec}(\boldsymbol{S}_{t-1})}{\partial \boldsymbol{\theta}'} &= \frac{\partial \operatorname{vec}(\boldsymbol{S}_{t-1}^*)}{\partial \boldsymbol{\theta}'} + \frac{\partial \operatorname{vec}(\boldsymbol{S}_{t-1})}{\partial \boldsymbol{f}_{t-1}'}\frac{\partial \boldsymbol{f}_{t-1}}{\partial \boldsymbol{\theta}'},
\end{aligned}$$

where the $\cdot^*$ is used to indicate that $\boldsymbol{f}_{t-1}$ is considered fixed.

A.1 Beta-Gen-t-EGARCH

From Harvey & Lange (2017) we find

$$\begin{aligned}
\frac{\partial \log p_t}{\partial \mu} &= \frac{\bar{\eta}+1}{\bar{\eta}}\frac{b_t}{y_t - \mu} \\
\frac{\partial \log p_t}{\partial \bar{\eta}} &= \frac{\psi\left(\frac{1}{\bar{\eta}\upsilon}\right) - \psi\left(\frac{1}{\bar{\eta}\upsilon} + \frac{1}{\upsilon}\right) + \bar{\eta} - \log(1-b_t) - (\bar{\eta}+1)b_t}{\upsilon\bar{\eta}^2}, \\
\frac{\partial \log p_t}{\partial \bar{\eta}} &= \frac{\bar{\eta}\upsilon - \log(1-b_t) - (\bar{\eta}+1)b_t \log b_t + (b_t - \bar{\eta}(1-b_t))\log(\bar{\eta}(1-b_t)}{\bar{\eta}\upsilon^2} \\
&\quad + \frac{\bar{\eta}\psi\left(\frac{1}{\upsilon}\right) + \psi\left(\frac{1}{\bar{\eta}\upsilon}\right) - (\bar{\eta}+1)\psi\left(\frac{1}{\bar{\eta}\upsilon} + \frac{1}{\upsilon}\right)}{\bar{\eta}\upsilon^2},
\end{aligned}$$

where $\psi(\cdot)$ is the digamma function. Since $S_{t-1} = 1$ in the Beta-Gen-t-EGARCH its derivatives w.r.t. to $\boldsymbol{\theta}$ are zero. The derivative of ∇_{t-1}^* w.r.t. to $\boldsymbol{\theta}$ is decomposed as

$$\begin{aligned}
\frac{\partial \nabla_{t-1}^*}{\partial \boldsymbol{\theta}'} &= \frac{\partial \nabla_{t-1}^{**}}{\partial \boldsymbol{\theta}'} + \frac{\partial \nabla_{t-1}^*}{\partial b_{t-1}}\frac{\partial b_{t-1}}{\partial \boldsymbol{\theta}'}, \\
\frac{\partial b_{t-1}}{\partial \boldsymbol{\theta}'} &= \frac{\partial b_{t-1}^*}{\partial \boldsymbol{\theta}'} + \frac{\partial b_{t-1}}{\partial f_{t-1}}\frac{\partial f_{t-1}}{\partial \boldsymbol{\theta}},
\end{aligned}$$

where $\cdot^{**}$ is used to denote that b_{t-1} is considered fixed. The derivative of ∇_{t-1}^{**} w.r.t. to $\bar{\eta}$ are

$$\frac{\partial \nabla_{t-1}^{**}}{\partial \bar{\eta}} = -\frac{1}{\bar{\eta}^2} b_{t-1}.$$

and zero for the other parameters. The derivative

$$\frac{\partial \nabla^*_{t-1}}{\partial b_{t-1}} = \frac{\bar{\eta}+1}{\bar{\eta}}.$$

The derivative of b_{t-1} w.r.t. f_{t-1} is

$$\frac{\partial b_{t-1}}{\partial f_{t-1}} = -\upsilon\, b_{t-1}(1-b_{t-1}).$$

Lastly, the derivatives of b^*_{t-1} are

$$\begin{aligned}
\frac{\partial b^*_{t-1}}{\partial \mu} &= -\frac{\upsilon\, b_{t-1}(1-b_{t-1})}{y_t-\mu},\\
\frac{\partial b^*_{t-1}}{\partial \bar{\eta}} &= \frac{1}{\bar{\eta}}\, b_{t-1}(1-b_{t-1}),\\
\frac{\partial b^*_{t-1}}{\partial \upsilon} &= b_{t-1}(1-b_{t-1})\log(y_t-\mu).
\end{aligned}$$

A.2 Multivariate GAS-t

In the multivariate GAS-t model the only distribution specific parameter in $\boldsymbol{\theta}$ is the inverse degrees of freedom parameter $\bar{\nu}$. The score for $\bar{\nu}$ is

$$\begin{aligned}
\frac{\partial \log p_t}{\partial \nu} = &\frac{1}{2\bar{\nu}^2}\Big(\psi\Big(\frac{1}{2\bar{\nu}}\Big) - \psi\Big(\frac{\bar{\nu}+1}{2\bar{\nu}}\Big)\Big) + \frac{k}{(1-2\bar{\nu})\bar{\nu}} + \frac{1}{2\bar{\nu}^2}\log\left(1+\bar{\nu}\,\frac{\boldsymbol{y}_t'\boldsymbol{\Sigma}_t^{-1}\boldsymbol{y}_t}{1-2\bar{\nu}}\right)\\
&- \frac{1+k\bar{\nu}}{2\bar{\nu}}\left(1+\bar{\nu}\,\frac{\boldsymbol{y}_t'\boldsymbol{\Sigma}_t^{-1}\boldsymbol{y}_t}{1-2\bar{\nu}}\right)^{-1}\frac{\boldsymbol{y}_t'\boldsymbol{\Sigma}_t^{-1}\boldsymbol{y}_t}{(1-2\bar{\nu})^2}.
\end{aligned}$$

Further we find for the derivative of ∇^*_{t-1} w.r.t. $\bar{\nu}$

$$\frac{\partial \boldsymbol{\nabla}^*_{t-1}}{\partial \bar{\nu}} = \frac{1}{2}\boldsymbol{\Psi}'_{t-1}\boldsymbol{\Sigma}^{-1}_{t-1\otimes}\boldsymbol{y}_{t-1\otimes}\frac{k+2-\boldsymbol{y}'_{t-1}\boldsymbol{\Sigma}^{-1}_{t-1}\boldsymbol{y}_{t-1}}{(1-\bar{\nu}(2-\boldsymbol{y}'_{t-1}\boldsymbol{\Sigma}^{-1}_{t-1}\boldsymbol{y}_{t-1}))^2}.$$

Since the scaling matrix for the multivariate GAS-t model is the inverse Fisher Information matrix, it is useful to define the derivative

$$\frac{\partial \operatorname{vec}(\boldsymbol{S}_{t-1})}{\partial \operatorname{vec}(\boldsymbol{\mathcal{I}}_{t-1})'} = -\boldsymbol{\mathcal{I}}^{-1}_{t-1\otimes}. \tag{26}$$

These and other matrix differentiation rules are obtained from Magnus & Neudecker (2007). Using the chain rule the derivatives of $\operatorname{vec}(\boldsymbol{S}_{t-1})$ w.r.t. $\boldsymbol{\theta}$ reduce to the derivatives of $\operatorname{vec}(\boldsymbol{\mathcal{I}}_{t-1})$ w.r.t. $\boldsymbol{\theta}$ pre-multiplied by (26). For $\bar{\nu}$ we therefore have

$$\frac{\partial \operatorname{vec}(\boldsymbol{S}^*_{t-1})}{\partial \bar{\nu}} = \frac{1}{2}\boldsymbol{\mathcal{I}}^{-1}_{t-1\otimes}\boldsymbol{\Psi}'_{t-1}\boldsymbol{J}'^{-1}_{t-1\otimes}\boldsymbol{G}\boldsymbol{J}^{-1}_{t-1\otimes}\boldsymbol{\Psi}_{t-1}(1+\bar{\nu}(2+k))^{-2}.$$

The derrivatives of $\boldsymbol{\nabla}_{t-1}$ and $\operatorname{vec}(\boldsymbol{S}_{t-1})$ w.r.t. $\boldsymbol{f}_{t-1}$ are split as follows

$$\begin{aligned}
\frac{\partial \boldsymbol{\nabla}_{t-1}}{\partial \boldsymbol{f}'_{t-1}} &= \frac{\partial \boldsymbol{\nabla}_{t-1}}{\partial \operatorname{vec}(\boldsymbol{\Sigma}_{t-1})'}\boldsymbol{\Psi}_{t-1} + \frac{\partial \boldsymbol{\nabla}_{t-1}}{\partial \operatorname{vec}(\boldsymbol{\Psi}_{t-1})'}\frac{\partial \operatorname{vec}(\boldsymbol{\Psi}_{t-1})}{\partial \boldsymbol{f}'_{t-1}},\\
\frac{\partial \operatorname{vec}(\boldsymbol{S}_{t-1})}{\partial \boldsymbol{f}'_{t-1}} &= -\boldsymbol{\mathcal{I}}^{-1}_{t-1\otimes}\left(\frac{\partial \operatorname{vec}(\boldsymbol{\mathcal{I}}_{t-1})}{\partial \operatorname{vec}(\boldsymbol{\Sigma}_{t-1})'}\boldsymbol{\Psi}_{t-1} + \frac{\partial \operatorname{vec}(\boldsymbol{\mathcal{I}}_{t-1})}{\partial \operatorname{vec}(\boldsymbol{\Psi}_{t-1})'}\frac{\partial \operatorname{vec}(\boldsymbol{\Psi}_{t-1})}{\partial \boldsymbol{f}'_{t-1}}\right).
\end{aligned}$$

The derivatives on the right-hand side of the equations above are

$$
\begin{aligned}
\frac{\partial \boldsymbol{\nabla}_{t-1}}{\partial \operatorname{vec}(\boldsymbol{\Sigma}_{t-1})'} =& \frac{1}{2}\big(\big(w_{t-1}\boldsymbol{y}_{t-1\otimes} - \operatorname{vec}(\boldsymbol{\Sigma}_{t-1})\big) \otimes \boldsymbol{\Psi}'_{t-1}\big) \frac{\partial \operatorname{vec}(\boldsymbol{\Sigma}^{-1}_{t-1\otimes})}{\partial \operatorname{vec}(\boldsymbol{\Sigma}_{t-1})'} \\
&+ \boldsymbol{\Psi}'_{t-1}\boldsymbol{\Sigma}^{-1}_{t-1\otimes}\Big(\frac{w^2_{t-1}}{\bar{\nu}(1+k\bar{\nu})}\boldsymbol{y}_{t-1}\boldsymbol{y}'_{t-1}\boldsymbol{\Sigma}^{-1}_{t-1\otimes}\boldsymbol{y}_{t-1\otimes} - \boldsymbol{I}_{k^2}\Big), \\
\frac{\partial \operatorname{vec}(\boldsymbol{\Sigma}^{-1}_{t-1\otimes})}{\partial \operatorname{vec}(\boldsymbol{\Sigma}_{t-1})'} =& \big(\boldsymbol{I}_k \otimes \boldsymbol{\mathcal{C}}_k \otimes \boldsymbol{I}_k\big)\big(\big(\operatorname{vec}(\boldsymbol{\Sigma}^{-1}_{t-1}) \otimes \boldsymbol{I}_{k^2}\big) + \big(\boldsymbol{I}_{k^2} \otimes \operatorname{vec}(\boldsymbol{\Sigma}^{-1}_{t-1})\big)\big)\boldsymbol{\Sigma}^{-1}_{t-1\otimes}, \\
\frac{\partial \boldsymbol{\nabla}_{t-1}}{\partial \operatorname{vec}(\boldsymbol{\Psi}_{t-1})'} =& \big(w_{t-1}\boldsymbol{y}_{t-1\otimes} - \operatorname{vec}(\boldsymbol{\Sigma}_{t-1})\big)'\boldsymbol{\Sigma}^{-1}_{t-1\otimes} \otimes \boldsymbol{I}_{k^2}, \\
\frac{\partial \operatorname{vec}(\boldsymbol{\mathcal{I}}_{t-1})}{\partial \operatorname{vec}(\boldsymbol{\Sigma}_{t-1})'} =& \big(\boldsymbol{\mathcal{C}}_n + \boldsymbol{I}_{n^2}\big)\big(\boldsymbol{I}_n \otimes \boldsymbol{\Psi}'_{t-1}\boldsymbol{J}_{t-1\otimes}\big(g\boldsymbol{G} - \operatorname{vec}(\boldsymbol{I}_k)\operatorname{vec}(\boldsymbol{I}_k)'\big)\big) \\
&\big(\boldsymbol{\Psi}'_{t-1} \otimes \boldsymbol{I}_{k^2}\big)\frac{\partial \operatorname{vec}(\boldsymbol{J}_{t-1\otimes})}{\partial \operatorname{vec}(\boldsymbol{\Sigma}^{-1}_{t-1\otimes})'}\frac{\partial \operatorname{vec}(\boldsymbol{\Sigma}^{-1}_{t-1\otimes})}{\partial \operatorname{vec}(\boldsymbol{\Sigma}_{t-1})'}, \\
\frac{\operatorname{vec}(\boldsymbol{\mathcal{I}}_{t-1})}{\partial \operatorname{vec}(\boldsymbol{\Psi}_{t-1})'} =& \big(\boldsymbol{\mathcal{C}}_n + \boldsymbol{I}_{n^2}\big)\big(\boldsymbol{I}_n \otimes \boldsymbol{\Psi}'_{t-1}\boldsymbol{J}_{t-1\otimes}\big(g\boldsymbol{G} - \operatorname{vec}(\boldsymbol{I}_k)\operatorname{vec}(\boldsymbol{I}_k)'\big)\big)\big(\boldsymbol{I}_n \otimes \boldsymbol{J}_{t-1\otimes}\big), \\
\frac{\partial \operatorname{vec}(\boldsymbol{\Psi}_{t-1})}{\partial \boldsymbol{f}'_{t-1}} =& \frac{1}{2}\big(\boldsymbol{\mathcal{S}}'_{\boldsymbol{D}} \otimes \boldsymbol{I}_{k^2}\big)\big(\boldsymbol{I}_k \otimes \boldsymbol{\mathcal{C}}_k \otimes \boldsymbol{I}_k\big)\big(\big(\operatorname{vec}(\boldsymbol{I}_k) \otimes \boldsymbol{I}_{k^2}\big) + \big(\boldsymbol{I}_{k^2} \otimes \operatorname{vec}(\boldsymbol{I}_k)\big)\big)\boldsymbol{\Psi}_{t-1} \\
&+ \big(\boldsymbol{\mathcal{S}}'_{\boldsymbol{\phi}}\boldsymbol{Z}'_{t-1}\big(\boldsymbol{I}_k \otimes \boldsymbol{X}_{t-1}\big)\big(\boldsymbol{I}_{k^2} + \boldsymbol{\mathcal{C}}_k\big) \otimes \boldsymbol{I}_{k^2}\big)\big(\boldsymbol{I}_k \otimes \boldsymbol{\mathcal{C}}_k \otimes \boldsymbol{I}_k\big) \\
&\big(\big(\operatorname{vec}(\boldsymbol{D}_{t-1}) \otimes \boldsymbol{I}_{k^2}\big) + \big(\boldsymbol{I}_{k^2} \otimes \operatorname{vec}(\boldsymbol{D}_{t-1})\big)\big)\boldsymbol{W}_{\boldsymbol{D}_{t-1}}\boldsymbol{\mathcal{S}}_{\boldsymbol{D}} \\
&+ \big(\boldsymbol{\mathcal{S}}'_{\boldsymbol{\phi}}\boldsymbol{Z}'_{t-1} \otimes \boldsymbol{D}_{t-1\otimes}\big(\boldsymbol{I}_{k^2} + \boldsymbol{\mathcal{C}}_k\big)\big)\big(\boldsymbol{I}_k \otimes \boldsymbol{\mathcal{C}}_k \otimes \boldsymbol{I}_k\big)\big(\operatorname{vec}(\boldsymbol{I}_k) \otimes \boldsymbol{I}_{k^2}\big)\boldsymbol{\mathcal{C}}_k\boldsymbol{Z}_{t-1}\boldsymbol{\mathcal{S}}_{\boldsymbol{\phi}} \\
&+ \big(\boldsymbol{\mathcal{S}}'_{\boldsymbol{\phi}} \otimes \boldsymbol{\mathcal{B}}_k\boldsymbol{D}_{t-1\otimes}\big(\boldsymbol{I}_{k^2} + \boldsymbol{\mathcal{C}}_k\big)\big(\boldsymbol{I}_k \otimes \boldsymbol{X}'_{t-1}\big)\big)\frac{\partial \operatorname{vec}(\boldsymbol{Z}_{t-1})}{\partial \boldsymbol{\phi}'_{t-1}}\boldsymbol{\mathcal{S}}_{\boldsymbol{\phi}}
\end{aligned}
$$

Where $\boldsymbol{W}_{\boldsymbol{D}_{t-1}}$ is the $k^2 \times k^2$ diagonal matrix with $0.5\operatorname{vec}(\boldsymbol{D}_{t-1})$ on its diagonal. The derivative of $\boldsymbol{J}_{t-1\otimes}$ requires a rule for the derivative of the Cholesky factor of a matrix, in this case $\boldsymbol{\Sigma}^{-1}_{t-1\otimes}$, since $\boldsymbol{\Sigma}^{-1}_{t-1\otimes} = \boldsymbol{J}_{t-1\otimes}\boldsymbol{J}'_{t-1\otimes}$ by the mixed product-property of the Kronecker product (see e.g. Magnus & Neudecker (2007)). Such a rule for the derivative of the Cholesky factor is given by Murray (2016) and looks as follows

$$
\frac{\partial \operatorname{vec}(\boldsymbol{J}_{t-1\otimes})}{\partial \operatorname{vec}(\boldsymbol{\Sigma}^{-1}_{t-1\otimes})'} = \big(\boldsymbol{I}_{k^2} \otimes \boldsymbol{J}_{t-1\otimes}\big)\boldsymbol{\mathcal{S}}_{\boldsymbol{L}}\big(\boldsymbol{J}^{-1}_{t-1\otimes} \otimes \boldsymbol{J}^{-1}_{t-1\otimes}\big),
$$

where $\boldsymbol{\mathcal{S}}_{\boldsymbol{L}}$ is defined such that for an arbitrary matrix $\boldsymbol{A}$, $\boldsymbol{\mathcal{S}}_{\boldsymbol{L}}\operatorname{vec}(\boldsymbol{A}) = \operatorname{vec}(\boldsymbol{\Phi}(\boldsymbol{A}))$ with $\boldsymbol{\Phi}(\cdot)$ the transformation that selects the lower-triangular part of a matrix and halves its diagonal elements. Lastly, we need the derivative of $\operatorname{vec}(\boldsymbol{Z}_{t-1})$ w.r.t. $\boldsymbol{\phi}_{t-1}$. Since this is simply the second derivative of $\operatorname{vec}(\boldsymbol{X}_{t-1})$ w.r.t. $\boldsymbol{\phi}_{t-1}$. I define this second derivative as

$$
\frac{\partial^2 x_{ijt}}{\partial \phi_{rst}\partial \phi_{mnt}} = \begin{cases} 0, & \text{if } i > j,\ r \geq s,\ m \geq n,\ r > i,\ r = i = j,\ m > i, \\ & m = i = j,\ j \neq n, \text{ or } j \neq s, \\ -x_{ijt}\frac{\tan(\phi_{ijt})}{\tan(\phi_{mjt})}, & \text{if } r = i,\ i > m \text{ and } i \neq j, \\ x_{ijt}\frac{1}{\tan(\phi_{rjt})\tan(\phi_{mjt})}, & \text{if } i > r, \text{ and } i > m, \\ -x_{ijt}\frac{\tan(\phi_{ijt})}{\tan(\phi_{rjt})}, & \text{if } m = i,\ i > r \text{ and } i \neq j, \\ -x_{ijt}, & \text{otherwise,} \end{cases} \tag{27}
$$

for $i, j, r, s, m, n = 1, ..., k$. These derivatives together constitute the elements of the $k^3(k-1)/2 \times k(k-1)/2$ matrix $\partial \operatorname{vec}(\boldsymbol{Z}_{t-1})/\partial \boldsymbol{\phi}'_{t-1}$.

B Prior Sensitivity Analysis DPMP Model Comparisons

For the sensitivity analysis I consider four different normal prior setups that differ only in their variance parameters. The prior setups are specified as follows

Prior 1	Prior 2	Prior 3	Prior 4
$a \sim \mathcal{N}(0.05, 2^2)$	$a \sim \mathcal{N}(0.05, 1)$	$a \sim \mathcal{N}(0.05, 0.5^2)$	$a \sim \mathcal{N}(0.05, 0.25^2)$
$b \sim \mathcal{N}(0.95, 2^2)\mathrm{I}_{[-1<b.<1]}$	$b \sim \mathcal{N}(0.95, 1)\mathrm{I}_{[-1<b.<1]}$	$b \sim \mathcal{N}(0.95, 0.5^2)\mathrm{I}_{[-1<b.<1]}$	$b \sim \mathcal{N}(0.95, 0.25^2)\mathrm{I}_{[-1<b.<1]}$
$C \sim \mathcal{N}(0.5, 10^2)$	$C \sim \mathcal{N}(0.5, 5^2)$	$C \sim \mathcal{N}(0.5, 2.5^2)$	$C \sim \mathcal{N}(0.5, 1)$
$C_{1,3} \sim \mathcal{N}(-0.5, 10^2)$	$C_{1,3} \sim \mathcal{N}(-0.5, 5^2)$	$C_{1,3} \sim \mathcal{N}(-0.5, 2.5^2)$	$C_{1,3} \sim \mathcal{N}(-0.5, 1)$
$w \sim \mathcal{N}(-5, 10^2)$	$w \sim \mathcal{N}(-5, 5^2)$	$w \sim \mathcal{N}(-5, 2.5^2)$	$w \sim \mathcal{N}(-5, 1)$
$w_2 \sim \mathcal{N}(-10, 10^2)$	$w_2 \sim \mathcal{N}(-10, 5^2)$	$w_2 \sim \mathcal{N}(-10, 2.5^2)$	$w_2 \sim \mathcal{N}(-10, 1)$,

where a and b are used to refer to the diagonal elements of $\boldsymbol{A}$ and $\boldsymbol{B}$ respectively, C is used to refer to all free elements in $\boldsymbol{C}$ apart from the $C_{1,3}$ parameter in the one factor model for which a different prior is used and w is used to refer to all elements in $\boldsymbol{w}$ apart from the w_2 parameter for which a different prior is specified. Prior setup 2 is the prior used in the analysis reported in Section 4.2. The double the log Bayes factors for all twelve different hypotheses and all four different prior setups are reported in Table 7.

Table 7: Bayes factors for the 12 hypotheses related to the DPMP factor models for the four prior specifications with varying diffuseness

	$2 \log \mathrm{BF}_{1\|0}$			
1|0	Prior 1	Prior 2	Prior 3	Prior 4
1-H|1-I	8.6	8.6	8.6	8.6
1-Inv|1-H	2.5	2.5	2.5	2.4
2-H|2-I	7.8	8.0	8.1	8.4
2-Inv|2-H	2.1	2.2	2.2	2.3
3-H|3-I	9.6	9.7	9.8	9.7
3-Inv|3-H	−2.5	−2.5	−2.6	−2.8
2-I|1-I	−11.8	−9.5	−7.8	−7.0
2-H|1-H	−12.6	−10.2	−8.3	−7.2
2-Inv|1-Inv	−13.0	−10.5	−13.4	−7.3
3-I|1-I	13.3	19.7	24.8	28.4
3-H|1-H	14.3	20.7	26.0	29.6
3-Inv|1-Inv	9.3	15.7	21.0	24.3

Notes: Bayes factors for the 12 hypotheses under consideration and for all four different prior specifications based on a 400,000 draw sample from a random walk Metropolis-Hastings chain.

C Estimation Results DPMP-I and DPMP-H

Table 8: Parameter estimates for the dynamic marked point process one, two and three factor models with identity information matrix scaling (DPMP1-I, DPMP2-I and DPMP3-I respectively)

	DPMP1-I				DPMP2-I				DPMP3-I			
$\boldsymbol{\theta}^{[1]}$	$\mathrm{E}(\cdot\vert\boldsymbol{Y}_T)$	SE	NSE ($\times 10^{-2}$)	$\boldsymbol{\theta}^{[2]}$	$\mathrm{E}(\cdot\vert\boldsymbol{Y}_T)$	SE	NSE ($\times 10^{-2}$)	$\boldsymbol{\theta}^{[1]}$	$\mathrm{E}(\cdot\vert\boldsymbol{Y}_T)$	SE	NSE ($\times 10^{-2}$)
a_1	0.105	0.008	0.010	a_1	0.105	0.008	0.012	a_1	0.054	0.009	0.016
				a_2	0.016*	0.018	0.028	a_2	0.015*	0.018	0.035
								a_3	0.124	0.008	0.013
b_1	0.968	0.013	0.017	b_1	0.970	0.013	0.020	b_1	0.972	0.017	0.031
				b_2	0.511	0.435	0.737	b_2	0.511	0.442	0.883
								b_3	0.967	0.014	0.026
$C_{1,1}$	0.408	0.041	0.050	$C_{1,1}$	0.411	0.042	0.059	$C_{2,1}$	0.494	0.811	1.344
$C_{2,1}$	0.854	0.317	0.382	$C_{2,1}$	0.830**	0.321	0.481	$C_{2,3}$	0.721	0.367	0.603
$C_{3,1}$	−0.147	0.047	0.061								
d_1	−5.368	0.126	0.167	d_1	−5.364	0.130	0.207	d_1	−5.373	0.200	0.357
d_2	−9.944	0.432	0.536	d_2	−9.914	0.423	0.679	d_2	−9.960	0.461	0.836
d_3	−5.515	0.060	0.077	d_3	−5.481	0.048	0.079	d_3	−5.479	0.046	0.077
d_4	−5.565	0.299	0.399	d_4	−5.552	0.308	0.514	d_4	−5.549	0.326	0.593
Accept	0.31				0.27				0.20		

Notes: Estimation results for the parameters of the DPMP1-I, DPMP2-I and DPMP3-I models based on a $400,000$ draw long Markov chain produced using the random walk Metropolis-Hastings algorithm. Reported for all three models and for all parameters are posterior mean ($\mathrm{E}(\cdot|\boldsymbol{Y}_T)$), standard deviation ($\mathrm{SE} = \sqrt{\mathrm{var}(\cdot|\boldsymbol{Y}_T)}$) and numerical standard error (NSE). For the $a_\cdot$ parameters and the $C_{\cdot,\cdot}$parameters of the one and two factor models, one star (*) is used to denote that the 95% highest posterior density region (HPD) includes zero and two stars (**) are used to denote that the 99% HPD includes zero. Also reported for all three Markov chains are the acceptance rates.

Table 9: Parameter estimates for the dynamic marked point process one, two and three factor models with inverse square root information matrix scaling (DPMP1-H, DPMP2-H and DPMP3-H respectively)

	DPMP1-H				DPMP2-H				DPMP3-H			
$\boldsymbol{\theta}^{[1]}$	$\mathrm{E}(\cdot\|\boldsymbol{Y}_T)$	SE	NSE ($\times 10^{-2}$)	$\boldsymbol{\theta}^{[2]}$	$\mathrm{E}(\cdot\|\boldsymbol{Y}_T)$	SE	NSE ($\times 10^{-2}$)	$\boldsymbol{\theta}^{[1]}$	$\mathrm{E}(\cdot\|\boldsymbol{Y}_T)$	SE	NSE ($\times 10^{-2}$)
a_1	0.074	0.007	0.008	a_1	0.073	0.007	0.010	a_1	0.034	0.005	0.009
				a_2	0.013*	0.009	0.015	a_2	0.012*	0.009	0.015
								a_3	0.082	0.008	0.013
b_1	0.964	0.015	0.019	b_1	0.966	0.015	0.023	b_1	0.965	0.018	0.034
				b_2	0.642	0.375	0.682	b_2	0.618	0.384	0.682
								b_3	0.964	0.017	0.028
$C_{1,1}$	0.387	0.043	0.052	$C_{1,1}$	0.392	0.042	0.065	$C_{2,1}$	0.115	0.942	1.735
$C_{2,1}$	0.889	0.323	0.398	$C_{2,1}$	0.860	0.326	0.493	$C_{2,3}$	0.883	0.444	0.795
$C_{3,1}$	−0.141	0.043	0.054								
d_1	−5.378	0.122	0.160	d_1	−5.387	0.124	0.200	d_1	−5.439	0.175	0.331
d_2	−9.997	0.453	0.552	d_2	−10.006	0.448	0.719	d_2	−10.065	0.522	1.000
d_3	−5.512	0.060	0.077	d_3	−5.467	0.053	0.101	d_3	−5.469	0.050	0.093
d_4	−5.605	0.306	0.409	d_4	−5.626	0.309	0.493	d_4	−5.637	0.344	0.685
Accept	0.34				0.31				0.21		

Notes: Estimation results for the parameters of the DPMP1-H, DPMP2-H and DPMP3-H models based on a $400,000$ draw long Markov chain produced using the random walk Metropolis-Hastings algorithm. Reported for all three models and for all parameters are posterior mean ($\mathrm{E}(\cdot|\boldsymbol{Y}_T)$), standard deviation ($\mathrm{SE} = \sqrt{\mathrm{var}(\cdot|\boldsymbol{Y}_T)}$) and numerical standard error (NSE). For the $a_\cdot$ parameters and the $C_{\cdot,\cdot}$parameters of the one and two factor models, one star (*) is used to denote that the 95% highest posterior density region (HPD) includes zero and two stars (**) are used to denote that the 99% HPD includes zero. Also reported for all three Markov chains are the acceptance rates.

D Automatic Differentiation: A Brief Introduction

To illustrate the principles of automatic differentiation (AD) and its benefits over symbolic expressions, I consider an example of a function that takes a $d \times 1$ vector $\boldsymbol{x}$ as input and has scalar output y. Lets also assume that the function is a composition of four functions that are allowed to have multivariate inputs and outputs such that the composition looks as follows

$$y = \mathfrak{f}_4(\mathfrak{f}_3(\mathfrak{f}_2(\mathfrak{f}_1(\boldsymbol{x})))). \tag{28}$$

Through repeated application of the chain rule we obtain its derivative

$$\frac{\partial y}{\partial \boldsymbol{x}} = \frac{\partial \mathfrak{f}_4}{\partial \mathfrak{f}_3}\frac{\partial \mathfrak{f}_3}{\partial \mathfrak{f}_2}\frac{\partial \mathfrak{f}_2}{\partial \mathfrak{f}_1}\frac{\partial \mathfrak{f}_1}{\partial \boldsymbol{x}}. \tag{29}$$

In essence AD works by distilling any function down to such a composition of elementary functions and systematically applying the chain rule. A conceptual difference with how the chain rule is used in symbolic differentiation however, is that AD applies the chain rule to numerical values.

D.1 Forward Mode AD

AD can be performed in two distinct modes known as forward and reverse mode differentiation. The names refer to the order in which the product in (29) is calculated. In forward mode the expression is evaluated in the order in which the functions $\mathfrak{f}_i$, for $i = 1, 2, 3, 4$, are evaluated

$$\frac{\partial y}{\partial \boldsymbol{x}} = \left(\frac{\partial \mathfrak{f}_4}{\partial \mathfrak{f}_3}\left(\frac{\partial \mathfrak{f}_3}{\partial \mathfrak{f}_2}\left(\frac{\partial \mathfrak{f}_2}{\partial \mathfrak{f}_1}\left(\frac{\partial \mathfrak{f}_1}{\partial \boldsymbol{x}}\right)\right)\right)\right).$$

This approach is the most intuitive and usually most closely resembles the way we would program the symbolic derivatives of the log-likelihoods of GAS functions. Note however, that even forward mode gradient evaluation is different from the usual way gradients are evaluated using symbolic differentiation. In forward mode we only evaluate and store (as numerical values) terms of the form $\dot{\boldsymbol{x}}_i = \partial \mathfrak{f}_i / \partial \boldsymbol{x}$, which can be recursively computed as

$$\dot{\boldsymbol{x}}_i = \frac{\partial \mathfrak{f}_i}{\partial \mathfrak{f}_{i-1}}\dot{\boldsymbol{x}}_{i-1}, \tag{30}$$

for $i = 1, 2, 3, 4$ and $\dot{\boldsymbol{x}}_0$ initialized at $\boldsymbol{I}_d$.[15] At the end of the recursion we find the quantity that we desired since $\dot{\boldsymbol{x}}_4 = \partial \mathfrak{f}_4 / \partial \boldsymbol{x} = \partial y / \partial \boldsymbol{x}$. The difference relative to symbolic expressions in the case that $\mathfrak{f}_i$ are functions with matrix outputs - as are highly common in the log-likelihood of the multivariate GAS-t model - is that we can often avoid instantiating the large symbolic expressions for $\partial \mathfrak{f}_i / \partial \mathfrak{f}_{i-1}$ and that we will usually be able to evaluate the expression (30) at the same order of complexity as d times the original function $\mathfrak{f}_i$. For matrix operations this can be achieved for instance by using the approach to deriving forward and reverse mode update rules for matrix operations as outlined in Giles (2008).

For example, let $\mathfrak{f}_3$ denote matrix inversion, let the output of $\mathfrak{f}_2$ be a square invertible matrix and let us assume $d = 1$, since the forward mode update rules by Giles (2008) are defined for single input variables. The results can be applied to multivariate inputs as well by simply applying the update rules d times and arranging the required intermediate input and output sensitivities correctly. To emphasize that we are dealing with matrices, let the input be denoted by $\boldsymbol{A} = \mathfrak{f}_2(\mathfrak{f}_1(x))$ and let the output be denoted by

[15] Technically this approach of evaluating the derivative w.r.t. to all inputs at once using forward mode differentiation, is known as multi-directional forward mode differentiation (Hascoet & Pascual, 2013). I omit the distinction because we are purely interested in evaluating gradients of functions with multiple inputs and a single output. Multi-directional forward mode is of the same order of computational complexity w.r.t to d as simply applying one-directional forward mode d times, but multi-directional forward mode circumvents the need for intermediate products of a function to also be recomputed d times as all derivatives are computed in one sweep. To see how the initialization makes sense, consider the identity map Id: $\mathbb{R}^d \to \mathbb{R}^d$ such that $\text{Id}(\boldsymbol{x}) = \boldsymbol{x}$. Its derivative $\partial \text{Id}(\boldsymbol{x}) / \partial \boldsymbol{x}$ clearly equals $\boldsymbol{I}_d$. So we can imagine an $\mathfrak{f}_0 = \text{Id}$, from which $\dot{\boldsymbol{x}}_0 = \boldsymbol{I}_d$ logically follows.

$\boldsymbol{C} = \boldsymbol{A}^{-1} = \mathfrak{f}_3(\mathfrak{f}_2(\mathfrak{f}_1(x)))$, so that $\dot{\boldsymbol{C}} = \dot{x}_3$ and $\dot{\boldsymbol{A}} = \dot{x}_2$. Using the terminology convention from Magnus & Neudecker (2007) we find the following differential expression for matrix inversion

$$\mathrm{d}\boldsymbol{C} = -\boldsymbol{C}\mathrm{d}\boldsymbol{A}\boldsymbol{C}. \tag{31}$$

The reason why symbolic expressions for the derivative of $\boldsymbol{C}$ w.r.t $\boldsymbol{A}$ blow up in dimension is because that requires that we compute and store the sensitivity of each element in $\boldsymbol{C}$ for each element in $\boldsymbol{A}$, which - if we let the matrices be of dimension $m \times m$ - amounts to a total of m^4 sensitivities (i.e. the infinitesimal perturbation of each $C_{q,r}$ due to a perturbation in each $A_{s,t}$ for $q, r, s, t = 1, ..., m$). Throughout this paper I use the approach argued for by Magnus & Neudecker (2007) of expressing such derivatives by means of vectorizing the differentials, which results in the following large symbolic expression for matrix inversion

$$\frac{\partial \operatorname{vec}(\boldsymbol{C})}{\partial \operatorname{vec}(\boldsymbol{A})} = -\boldsymbol{C}' \otimes \boldsymbol{C}. \tag{32}$$

Applying this expression in (29) implies that we would need to multiply this $m^2 \times m^2$ by either $\partial\mathfrak{f}_4/\partial\mathfrak{f}_3$ or $\partial\mathfrak{f}_2/\partial\mathfrak{f}_1$, which, for reasonable m, is an expensive operation either way. The forward mode update rules as suggested by Giles (2008), exploit the fact that we are only interested in the sensitivity of $\boldsymbol{C}$ w.r.t. x, which are only m^2 sensitivities. From (31) it follows that we can obtain these sensitivities directly in terms of the sensitivities w.r.t $\boldsymbol{A}$ as

$$\dot{\boldsymbol{C}} = -\boldsymbol{C}\dot{\boldsymbol{A}}\boldsymbol{C}. \tag{33}$$

Considering matrix multiplication and inversion are on the same order of computational complexity, it follows that the update from $\dot{x}_2$ to $\dot{x}_3$ can be performed on the same order of computational complexity as the original function $\mathfrak{f}_3$. Obtaining sensitivities w.r.t d inputs is by extension of the same order of complexity as d times the original function.

In theory these types of forward mode update rules that operate on d times the level of computational complexity as the original function should exist for any function. To highlight the full implications of this in terms of computational cost and storage requirements for the full example function introduced in the beginning of this section, consider the case where $\mathfrak{f}_i : \mathbb{R}^{m\times m} \to \mathbb{R}^{m\times m}$ for $i = 2, 3$, $\mathfrak{f}_1 : \mathbb{R}^d \to \mathbb{R}^{m\times m}$ and $\mathfrak{f}_4 : \mathbb{R}^{m\times m} \to \mathbb{R}$, such that the complete function is a mapping from $\mathbb{R}^d \to \mathbb{R}$. Again adopting the matrix differentiation conventions that matrix derivatives are denoted by vectorizing their differentials, we find that straightforward evaluation of (29) - through computing the products of the terms on the right hand side - will involve the construction of one $1 \times m^2$ row vector, two $m^2 \times m^2$ and one $m^2 \times d$ matrix, all of which will be involved in matrix multiplications.[16] Using forward mode differentiation we are thus typically able to get away with storing only the $m^2 \times d$ dimensional $\dot{\boldsymbol{x}}_i$s, which, if $d < m^2$, should imply lower storage cost. But more importantly, if we assume the functions $\mathfrak{f}_i$ to be elementary matrix operations of similar complexity as matrix multiplications, the resulting gradient program will be of lower computational complexity. This follows from the fact that the forward mode differentiation requires d application of forward mode updates that involve matrix operations on at the most $m \times m$ matrices, which should be compared to the multiplication of the $m^2 \times m^2$ matrices resulting from straightforward computation of the symbolic expressions. Considering most matrix operations - including the most commonly used in the multivariate GAS-t model log-likelihood such as matrix multiplication, inversion and Cholesky decomposition - on $m \times m$ matrices require roughly $\mathcal{O}(m^3)$ elementary operations (i.e. additions and multiplications), forward mode differentiation should in theory be of lower computational complexity as the symbolic expressions as long as $d < m^3$.

If we draw parallels to the computation of the gradient of the GAS-t model log-likelihood, we note that the log-likelihood is also a mapping from $\mathbb{R}^d \to \mathbb{R}$, where d still denotes the dimension of $\boldsymbol{\theta}$. The highest

[16] It should be noted that symbolic expressions for matrix derivatives can sometimes be drastically simplified - by for instance deriving $\partial\mathfrak{f}_3/\partial\mathfrak{f}_1$ directly - potentially allowing significant reduction in the computational burden they induce. This can however be very challenging and it might not be possible to express such derivatives using standard matrix operations. For the symbolic derivatives presented in Appendix A.2 for the multivariate GAS-t log-likelihoods I have made use mostly of the common identities for matrix derivatives expressed in Magnus & Neudecker (2007), with the exception of the Cholesky factor derivative reported in Murray (2016). In this Appendix I operate on the assumption that symbolic derivatives of matrix operations expressed by the functions $\mathfrak{f}_i$ can not be further simplified.

dimensional intermediate matrix functions are multiplications with $k^2 \times k^2$ matrices (e.g. $\boldsymbol{\Sigma}_{t-1\otimes}^{-1}$ and $\boldsymbol{J}_{t-1\otimes}$) and the resulting symbolic gradient expressions involve multiplications with $k^4 \times k^4$ matrices. In terms of the indices of the above example we thus find that k^2 can be interpreted as m and $d = k(k+1)+5$ in the hierarchical GAS-t model. Clearly there is indeed a gain from using forward mode differentiation, since d is still significantly less as $m^3 = k^6$, however the computational complexity relative to the original log-likelihood is still proportionate to d which is on the order of k^2. In the 10 asset case the gradient of the likelihood would thus still be at the least 115 times as computationally intensive as the log-likelihood.

D.2 Reverse Mode AD

Since our interest is with the gradient we should be better off using reverse-mode differentiation. The reason for this is result often mentioned in the AD literature that, using reverse mode differentiation, it should be possible in theory to evaluate the gradient of any function with a scalar output at only three to four times the computational cost of the original function (Griewank & Walther, 2008). In practice it can turn out to be more if the program is inefficiently implemented or the memory requirements are very large. The result is known as the "cheap gradient principle" and is considered as one of the most valuable achievements of the AD field (Gower & Gower, 2016).

To see how we can attain this level of performance that seems almost too good to be true considering the discussion in the preceding subsection, lets consider again the example introduced at the beginning of this section. In direct opposition to forward mode, reverse mode differentiation evaluates the expression (29) in reverse order, starting from $\partial \mathfrak{f}_4/\partial \mathfrak{f}_3$ and working its way down to $\partial \mathfrak{f}_1/\partial \boldsymbol{x}$, as follows

$$\frac{\partial y}{\partial \boldsymbol{x}} = \left(\left(\left(\left(\frac{\partial \mathfrak{f}_4}{\partial \mathfrak{f}_3}\right)\frac{\partial \mathfrak{f}_3}{\partial \mathfrak{f}_2}\right)\frac{\partial \mathfrak{f}_2}{\partial \mathfrak{f}_1}\right)\frac{\partial \mathfrak{f}_1}{\partial \boldsymbol{x}}\right).$$

The convenient notation typically used to denote these types of derivatives is $\bar{y}_i = \partial y/\partial \mathfrak{f}_i$ and the desired result can again be obtained using a recursion

$$\bar{y}_{i-1} = \bar{y}_i \frac{\partial \mathfrak{f}_i}{\partial \mathfrak{f}_{i-1}}, \tag{34}$$

for $i = 4, 3, 2, 1$, where we initialize with $\bar{y}_4 = 1$ and specify $\mathfrak{f}_0 = \boldsymbol{x}$ such that $\bar{y}_0 = \partial y/\partial \boldsymbol{x}$.

The key difference here with forward mode is that the quantities that we evaluate and store in this scenario, the $\bar{y}_i$s, are significantly smaller. If we assume the same sorts of functions $\mathfrak{f}_\mathrm{i}$ as in the preceding subsection we see that the $\bar{y}_i$ are never larger than $1 \times m^2$ or $1 \times d$. Also, just as for the forward mode expression (30), the evaluation of the recursive expressions in (34) can typically be performed without instantiating the $\partial \mathfrak{f}_i/\partial \mathfrak{f}_{i-1}$ terms, however in reverse mode the evaluation can be done at the same order of computation complexity as the original function $\mathfrak{f}_i$ rather than d times the complexity.

For illustration let us again consider the example where $\mathfrak{f}_3$ concerns matrix inversion. Giles (2008) shows how similar updates as for forward mode differentiation can be obtained for reverse mode. It is however not immediately clear from the differential expression in (31) how we can express $\bar{y}_2 = \bar{\boldsymbol{A}}$ directly in terms of $\bar{y}_3 = \bar{\boldsymbol{C}}$. Giles (2008) relies on trivial expressions related to the definition of $\bar{\boldsymbol{C}}$ and the trace (denoted with $\mathrm{tr}(\cdot)$), which is the operator that takes the sum of the diagonal elements of a square matrix. By plugging in the expression for $\mathrm{d}\boldsymbol{C}$ (31) and using the fact that $\mathrm{tr}(\boldsymbol{PQ}) = \mathrm{tr}(\boldsymbol{QP})$ and $\mathrm{tr}(\boldsymbol{P}) = \mathrm{tr}(\boldsymbol{P}')$, for arbitrary

square matrices $\boldsymbol{P}$ and $\boldsymbol{Q}$, we find

$$\begin{aligned}
\mathrm{d}y &= \sum_{q=1}^{m}\sum_{r=1}^{m}\left(\bar{\boldsymbol{C}}\right)_{i,j}\left(\mathrm{d}\boldsymbol{C}\right)_{i,j}, \\
&= \mathrm{Tr}(\bar{\boldsymbol{C}}'\mathrm{d}\boldsymbol{C}), \\
&= \mathrm{Tr}(-\bar{\boldsymbol{C}}'\boldsymbol{C}\mathrm{d}\boldsymbol{A}\boldsymbol{C}), \\
&= \mathrm{Tr}(-\boldsymbol{C}\bar{\boldsymbol{C}}'\boldsymbol{C}\mathrm{d}\boldsymbol{A}), \\
&= \mathrm{Tr}(-\left(\mathrm{d}\boldsymbol{A}\right)'\boldsymbol{C}'\bar{\boldsymbol{C}}\boldsymbol{C}'), \\
&= \sum_{q=1}^{m}\sum_{r=1}^{m}\left(\mathrm{d}\boldsymbol{A}\right)_{i,j}\left(\boldsymbol{C}'\bar{\boldsymbol{C}}\boldsymbol{C}'\right)_{i,j},
\end{aligned}$$

where the second and and last equality follow trivially from the definition of the trace operator. The subscripts refer to the elements of the matrices. From the final expression it follows that

$$\bar{\boldsymbol{A}} = \boldsymbol{C}'\bar{\boldsymbol{C}}\boldsymbol{C}'. \tag{35}$$

This expression is again of the same order of computational complexity a the original function $\mathfrak{f}_3$ and circumvents the large symbolic expression related to the derivative of the inverse. This update needs to preformed only once since, unlike in the forward mode setting, we are interested in the sensitivities w.r.t. one quantity - the scalar y - in contrast to the d entries in $\boldsymbol{x}$. The techniques used above can be used to derive similar reverse mode update expression for most other matrix operations.

As mentioned in the Section 4.3.3 of the main text, reverse mode has one main limitation that challenges its implementation. The evaluation of the expressions in (34) will often require intermediate results; e.g. if $\mathfrak{f}_3$ is the matrix inversion function considered above, the evaluation of $\bar{y}_2$ will then require $\boldsymbol{C}$ - the value coming from $\mathfrak{f}_3$. Although, in this toy example it might not seem like to much effort to store the result $\boldsymbol{C}$ in memory, the task of storing such intermediate products is more daunting when we consider the log-likelihood of the multivariate GAS-t model based on three years of daily data. My solution to this issue is covered in detail the main text.

However effective the solution to the memory management problem, the resulting gradient program will always need a forward pass to compute all intermediate products, followed by a backwards pass to compute the $\bar{y}_i$s. Although, this does imply that the resulting gradient program's computational cost is always some constant multiple of the original program's computational cost, for high dimensional d the resulting gradient program is far more efficient as what is possible using forward mode differentiation. A case in point being the reverse mode gradient program for the multivariate GAS-t log-likelihood for which it turns out that evaluation is possible near the theoretical upper bound of roughly three to four times the computational cost of the original function as opposed to the approximate 115 times for the forward mode program.

E Parameter Estimates Summary 10 Asset GAS-t Models

Table 10: Summary of parameter estimates for the restricted multivariate scalar GAS-t model (s-GAS-t) and the multivariate random coefficient GAS-t model (RC-GAS-t) for the 10 portfolio dataset

	s-GAS-t	RC-GAS-t
$\frac{1}{10}\Sigma_{i=1}^{10}\,\mathrm{E}(a_{i,\boldsymbol{D}}\vert\boldsymbol{Y}_T)$	0.058	0.061
$\frac{1}{55}\Sigma_{i=1}^{55}\,\mathrm{E}(a_{i,\boldsymbol{\phi}}\vert\boldsymbol{Y}_T)$	0.015	0.015
$\frac{1}{10}\Sigma_{i=1}^{10}\,\mathrm{E}(b_{i,\boldsymbol{D}}\vert\boldsymbol{Y}_T)$	0.982	0.981
$\frac{1}{55}\Sigma_{i=1}^{55}\,\mathrm{E}(b_{i,\boldsymbol{\phi}}\vert\boldsymbol{Y}_T)$	0.994	0.989
$\sqrt{\mathrm{var}(\mathrm{E}(a_{\boldsymbol{D}}\vert\boldsymbol{Y}_T))}$	0.010	0.010
$\sqrt{\mathrm{var}(\mathrm{E}(a_{\boldsymbol{\phi}}\vert\boldsymbol{Y}_T))}$	-	0.003
$\sqrt{\mathrm{var}(\mathrm{E}(b_{\boldsymbol{D}}\vert\boldsymbol{Y}_T))}$	0.003	0.003
$\sqrt{\mathrm{var}(\mathrm{E}(b_{\boldsymbol{\phi}}\vert\boldsymbol{Y}_T))}$	-	0.002
$\frac{1}{10}\Sigma_{i=1}^{10}\sqrt{\mathrm{var}(a_{\boldsymbol{D}}\vert\boldsymbol{Y}_T)}$	0.008	0.008
$\frac{1}{55}\Sigma_{i=1}^{55}\sqrt{\mathrm{var}(a_{\boldsymbol{\phi}}\vert\boldsymbol{Y}_T)}$	0.001	0.005
$\frac{1}{10}\Sigma_{i=1}^{10}\sqrt{\mathrm{var}(b_{\boldsymbol{D}}\vert\boldsymbol{Y}_T)}$	0.005	0.005
$\frac{1}{55}\Sigma_{i=1}^{55}\sqrt{\mathrm{var}(b_{\boldsymbol{\phi}}\vert\boldsymbol{Y}_T)}$	0.001	0.005
$\mu_{a,\boldsymbol{\phi}}$	-	0.015 (0.002)
$\mu_{b,\boldsymbol{\phi}}$	-	0.989 (0.002)
$\sigma_{a,\boldsymbol{\phi}}$	-	0.005 (0.002)
$\sigma_{b,\boldsymbol{\phi}}$	-	0.005 (0.002)
$\bar{\nu}$	0.084 (0.008)	0.080 (0.008)
Leapfrog steps	4	7
Step size	0.5	0.25
Accept rate	0.83	0.78

Notes: Summary of estimation results for the parameters of the multivariate s-GAS-t and RC-GAS-t models based on a 20,000 draw long Markov chain produced using Hamiltonian Monte Carlo. Reported for the autoregressive coefficients of (i) the log variance time-varying parameters (denoted by $\cdot_{\boldsymbol{D}}$) and (ii) the hyperspherical angle time-varying parameters (denoted by $\cdot_{\boldsymbol{\phi}}$) are the average and standard deviation ($\sqrt{\mathrm{var}(\cdot)}$) of the posterior mean ($\mathrm{E}(\cdot\vert\boldsymbol{Y}_T)$) estimates, and the average of the posterior standard deviation ($\sqrt{\mathrm{var}(\cdot\vert\boldsymbol{Y}_T)}$). For the hyper parameters and inverse degrees of freedom parameters the posterior means are reported with the posterior standard deviations in brackets. Also reported for all three Markov chains are the number of Leapfrog integrator steps, the integrator step size and the acceptance rates.