ETH
Swiss Federal Institute of Technology Zurich

Seminar for Statistics

Department of Mathematics

Seminar Paper Spring 2023

Nicolas N. Koch

Dirichlet Process GARCH Mixtures for Volatility & Value-at-Risk Estimation

Submission Date: 30 June 2023

Advisor: Prof. Dr. P. Cheridito

Abstract

Bayesian nonparametrics provides a rich set of highly flexible models free of strict distributional assumptions. They circumvent the problem of model selection through data-driven complexity determination and model averaging. This paper explores their application to GARCH models through the Dirichlet process, yielding an infinite mixture GARCH$(1,1)$. In theory, it combines the flexibility of mixture models and Bayesian estimation, helping to model the shape of conditional returns and hence improving robustness to misspecification thereof. The hypothesis is tested on simulated data. As reasonable estimation of the model failed, the numerical analysis is limited to a set of its special cases, including a Bayesian, finite mixture, and Bayesian finite mixture GARCH. The merits of both Bayesian estimation and partly those of mixture modeling are supported, but not necessarily in combination.

Contents

1 Introduction

The generalized autoregressive conditional heteroscedasticity (GARCH) model by Bollerslev (1986) belongs to the most widely used models to estimate time-varying volatility of financial time series. While popular extensions for the volatility dynamics, such as Nelson's (1991) exponential GARCH, can often improve the standard model, conditional returns remain highly skewed and fat-tailed. In response, many modern extensions apply flexible methods, particularly mixture distributions, to directly model innovations (Ausín and Galeano, 2007; Haas, Krause, Paolella, and Steude, 2013). An equivalent model class can be obtained through *regime-switching* (Gray, 1996; Haas, Mittnik, and Paolella, 2004a). These models provide an alternative interpretation of mixture conditional return distributions in the form of additional stochasticity in volatility, in particular latent time-varying volatility regimes. Classical versions thereof still require pre-specification of the (finite) number of regimes.

An alternative approach is Bayesian estimation of GARCH coefficients (Geweke, 1993; Ardia, 2008; Li, Clements, and Drovandi, 2021). Depending on the chosen prior, the thereby obtained posterior parameter distribution can be significantly more complex than a categorical and hence conditional returns richer than a finite mixture. In contrast to regime-switching models, however, coefficients are not assumed to be time-varying. Thus, they implicitly assume constant volatility persistence, which is not usually empirically supported.

Nonparametric Bayesian models extend classical Bayesian models by allowing the prior distribution to be a general stochastic process. Many popular such processes, including the Dirichlet process (Ferguson, 1973) considered in this analysis, produce repeated coefficient values and thus recurrent regimes as in regime-switching models. Yet, in contrast to the latter, they flexibly model the number of regimes and thus render prior selection thereof redundant. Consequently, their application to GARCH models naturally combines the ideas of regime-switching and Bayesian estimation, thereby inheriting the richness of both while avoiding their limitations. Such models have been proposed by Lau and So (2009) and extended by Lau and Cripps (2012), where they were shown to exhibit superior fit to empirical data compared to the classical Bayesian GARCH.

Besides volatility, another key quantity in financial applications is Value-at-Risk (VaR), which constitutes a certain quantile of the conditional return distribution. As classical GARCH-type models are based on point estimates of volatility, they lack flexibility to model characteristics of conditional returns other than their variance. However, the shape, including skewness and kurtosis, of a distribution is an additional key determinant of its quantiles. Stochastic models of volatility can adjust for misspecification of innovation shape, since variance-mixing increases skewness, kurtosis, as well as higher moments. Therefore, it is argued that Bayesian and regime-switching approaches constitute a promising class of models for VaR estimation, all the more their combination.

The hypothesis investigated in the present paper is that the flexibility and Bayesian nature of nonparametric Bayesian GARCH models improves their robustness against misspecification of volatility dynamics and shape of innovations. The hypothesis is put to the test in a simulation study assessing volatility and VaR estimation errors. As reasonable estimation of the model failed, the numerical analysis is limited to the marginal benefits of finite regime-switching and Bayesian estimation.

The text is organized as follows. Section 2 introduces the model, reviews its foundations from Bayesian nonparametrics, and shows how it extends classical approaches. Section 3 analyses the model theoretically, while Section 4 reports on the simulation study. Section 5 reviews key findings and concludes.

2 Model

2.1 Setup

Consider a discrete one-dimensional stochastic process $(X_t)_{t\in\mathbb{Z}}$ adapted to the filtration $(\mathcal{F}_t)_{t\in\mathbb{Z}} = \sigma(\{X_s, s \leq t\})$.[1] Bollerslev's (1986) original GARCH$(1,1)$ model specifies X_t to be of the form

$$\begin{aligned} X_t &= \sigma_t Z_t \\ \sigma_t^2 &= \omega + \alpha X_{t-1}^2 + \beta \sigma_{t-1}^2 \end{aligned} \tag{2.1}$$

where σ_t is the *volatility* at time t and (Z_t) is the innovation process, for which realizations are drawn independently and identically distributed (iid) with $\mathbb{E}[Z_t] = 0$ and $\mathbb{V}(Z_t) = 1$ for all $t \in \mathbb{Z}$. The conditional volatility $\sigma_t \,|\, \mathcal{F}_{t-1}$ is considered fixed. A nonparametric Bayesian GARCH$(1,1)$ model loosens this assumptions by modeling it as the the result of another stochastic process, which is assumed to be of the form

$$\begin{aligned} \sigma_t^2 &= \omega_t + \alpha_t X_{t-1}^2 + \beta_t \sigma_{t-1}^2 \\ (\omega_t, \alpha_t, \beta_t) \mid F &\overset{iid}{\sim} F \\ F &\sim \mathscr{L} \end{aligned}$$

As in Bollerslev's (1986) model, the distribution of Z_t is pre-specified. In contrast, GARCH parameters are assumed to be time-varying and stochastic. They are modeled as iid realizations from some distribution F, which is itself randomly generated from the law $\mathscr{L}$. The law considered in the present analysis is the *Dirichlet process* (DP), which is introduced in Section 2.2. Consequently, the model is referred to as the *DP-GARCH*.

Note that despite pre-specification of the innovation distribution, the shape of conditional returns $X_t \,|\, \mathcal{F}_{t-1}$ is not fixed. This stems from the stochasticity of the GARCH coefficients, which induce conditional returns to be a variance-mixture of the innovation distribution. Therefore, the approach can flexibly model the volatility dynamics and return distribution simultaneously.

Henceforth, unless otherwise specified, GARCH refers to the GARCH$(1,1)$ model. The shorthand notation $\boldsymbol{\theta}_t = (\omega_t, \alpha_t, \beta_t)^\top$ is used for the GARCH coefficients at time t and $\boldsymbol{\theta}_{1:n} = (\boldsymbol{\theta}_t)_{t=1}^n$ to refer to the collection of all coefficients within the observed time interval $t = 1, ..., n$. Further, we impose the variance positivity constraints, so that $\boldsymbol{\theta}_t \in \Theta = \{(\omega, \alpha, \beta) \mid \omega > 0, \alpha \geq 0, \beta \geq 0\}$. Also, any GARCH-type model for which the conditional volatility $\sigma_t \,|\, \mathcal{F}_{t-1}$ is non-deterministic will be referred to as *stochastic (volatility) model*. Lastly, to emphasize the fact that conditional volatility depends on the GARCH coefficients, it may be written as $\sigma_t(\boldsymbol{\theta}_{1:t})$.[2]

2.2 Dirichlet Process

A law $\mathscr{L}$ is a stochastic process generating a random measure F. Amongst the simplest yet most popular laws in Bayesian nonparametrics is the Dirichlet process (DP), formalized by Ferguson (1973). The DP is parametrized by a concentration parameter $\alpha > 0$ and a non-atomic base distribution G_0 on a parameter space Θ, giving it the overall notation $\mathcal{DP}(\alpha, G_0)$.

[1] Informally, (X_t) is adapted to $(\mathcal{F}_t)$ if X_t can be fully determined by information available at time t, i.e. from $\{X_s; s \leq t\}$.

[2] Note that in GARCH-type models, volatility is defined recursively, and hence given $\mathcal{F}_{t-1}$, σ_t depends on $\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_t$ and not only $\boldsymbol{\theta}_t$.

Theoretically, any random measure whose marginals follow a Dirichlet distribution qualifies as a DP, as formalized by Definition 2.1.

Definition 2.1 (Dirichlet process, DP)**.** *Let F be a random measure over a parameter space Θ. Then, F follows a Dirichlet process, $\mathcal{DP}(\alpha, G_0)$, with concentration parameter α and base measure G_0 if for any finite measurable partition A of Θ, i.e. $A = \{A_1, ..., A_n\}$ such that $A_1 \cup ... \cup A_n = \Theta$ and $A_i \cap A_j = \emptyset, i \neq j$, it holds that*

$$\big(F(A_1), ..., F(A_n)\big) \sim Dir\big(\alpha G_0(A_1), ..., \alpha G_0(A_n)\big)$$

where Dir is the Dirichlet distribution.

Its mechanism is connected to various concepts from stochastics. In the following, two intuitively distinct but mathematically equivalent constructions of DPs are presented. These will prove useful for illustrating properties of DPs and DP-GARCH mixtures later on.

Stick-breaking process

Consider the Dirichlet distribution of order K, which is a probability distribution on the $(K-1)$-simplex $\Delta^{K-1} = \{\boldsymbol{\pi} \in \mathbb{R}_+^K \mid \boldsymbol{\pi} \geq \mathbf{0}, \|\boldsymbol{\pi}\|_1 = 1\}$. By taking K to infinity, we approach the GEM distribution (Ewens, 1988, p. 217),

$$\text{GEM}(\alpha) = \lim_{K \to \infty} \text{Dir}\big(\alpha/K, ..., \alpha/K\big)$$

A draw from a GEM is an infinite-dimensional vector whose components sum to 1. Such vectors can be generated through a *stick-breaking* procedure. Therein, the weights $\pi_1, \pi_2, ...$ are determined by repeatedly breaking the residual of a "stick" of initial length 1 at random locations drawn from a $\text{Beta}(1, \alpha)$ distribution, i.e.

$$\pi_k = B_k \prod_{j=1}^{k-1} (1 - B_j), \quad B_k \overset{iid}{\sim} \text{Beta}(1, \alpha) \tag{2.2}$$

A DP can then be constructed by coupling the GEM distribution with the base distribution G_0. Specifically, if $F \sim \mathcal{DP}(\alpha, G_0)$, then, equivalently,

$$\begin{aligned} F &= \sum_{k=1}^{\infty} \pi_k \delta_{\boldsymbol{\theta}_k}, \\ \pi_{1:\infty} &\sim \text{GEM}(\alpha), \\ \boldsymbol{\theta}_k &\overset{iid}{\sim} G_0 \end{aligned}$$

where $\delta_{\boldsymbol{\theta}}$ denotes the Dirac measure concentrated at $\boldsymbol{\theta}$.

Although an intuitive characterization, stick-breaking induces a distribution on samples $\boldsymbol{\theta}_{1:n}$ from the random measure F given by

$$\mathbb{P}(\boldsymbol{\theta}) = \iint \sum_{k=1}^{\infty} \pi_k \delta_{\boldsymbol{\theta}_k} \, G_0(d\boldsymbol{\theta}_{1:n}) \, \text{GEM}(\alpha; d\pi_{1:\infty}), \tag{2.3}$$

which is difficult to work with. The second representation of the Dirichlet process helps to make it more practical.

Algorithm 1 Chinese restaurant process (Aldous, 1985)

Draw $\boldsymbol{\theta}_1$ from G_0
for $t = 1$ to n **do**
  Draw $\boldsymbol{\theta}_t$ from...
  - ...G_0 with probability $\propto \alpha$
  - ...$\widehat{F}_{t-1}$ with probability $\propto t-1$, where $\widehat{F}_{t-1}$ is the empirical distribution of $\boldsymbol{\theta}$ at iteration $t-1$

end for

Chinese restaurant process

A draw from $\mathcal{DP}(\alpha, G_0)$ can also be obtained through a procedure referred to as the *Chinese restaurant process* (CRP), as outlined in Algorithm 1.

The CRP allegedly emulates the allocation procedure of customers to tables in Chinese restaurants. In this analogy, indices $t = 1, ..., n$ correspond to customers entering the restaurant. With probability proportional to α, the t-th customer sits at a new table. With probability proportional to $t-1$, they join the table of a random previous customer $j = 1, ..., t-1$. Whenever a new table k is opened, that table receives a random dish $\boldsymbol{\theta}_k$ drawn from G_0. After all n customers have entered the restaurant, they occupy a certain number K of distinct tables associated with distinct dishes according to a particular seating arrangement.

More formally, the outcome of a CRP can be regarded as two-fold. First, we obtain a sequence $\boldsymbol{\theta}_{1:n} = (\boldsymbol{\theta}_t)_{t=1}^n$. Second, since there is a strictly positive probability of $\boldsymbol{\theta}_{1:n}$ containing repeated values, let $\boldsymbol{\theta}^*_{1:K} = (\boldsymbol{\theta}^*_k)_{k=1}^K$ be a vector of the unique components of $\boldsymbol{\theta}_{1:n}$ and $\boldsymbol{r} \in \mathbb{N}^n$ be a corresponding assignment vector assigning indices $t \in \{1, ..., n\}$ to elements of $\boldsymbol{\theta}^*_{1:K}$. Let $\boldsymbol{p} = \{p_k\}_{k=1}^K$ then be the corresponding partition of $\{1, ..., T\}$, so that $p_k = \{j : z_j = k\}, k = 1, ..., K$.[3] Consequently, the CRP provides a distribution on partitions. Additionally, $|p_k|/n, k = 1, ..., K$, correspond to the weights (π_k) from Eq. 2.2 associated with clusters from which at least one member was sampled. In other words, the CRP represents another way to obtain samples from a GEM distribution. Simultaneously, the components of $\boldsymbol{\theta}^*_{1:K}$ are being sampled iid from G_0. Consequently, the CRP and stick-breaking construction are equivalent. Yet, notice that the CRP integrates out the random measure F, and hence it allows to directly sample from Eq. 2.3. It also allows to write the distribution of a sequence $\boldsymbol{\theta}_{1:n}$ as

$$\mathbb{P}(\boldsymbol{\theta}_{1:n}) \;=\; \alpha^K \prod_{k=1}^{K} \frac{(|p_k| - 1)!}{\alpha + k - 1} G_0(\boldsymbol{\theta}^*_k) \tag{2.4}$$

Other important representations of random measures are the *Pólya* or *Blackwell-MacQueen urn* predictive distribution (Blackwell and MacQueen, 1973) and the *exchangeable partition probability function* (Pitman, 1996). These characterizations are of particular relevance in the development of posterior sampling algorithms, which is an advanced issue of Bayesian nonparametrics and outside the scope of the present analysis. See Lau and So (2009); Lau and Cripps (2012) for Markov chain Monte Carlo (MCMC) methods for the DP-GARCH and Blei and Jordan (2006) for methods for general DP mixture models.

[3] As an example, consider the sequence $\boldsymbol{\theta} = (3, 8, 3, 5)$. Then, $\boldsymbol{\theta}^* = (3, 8, 5)$, $\boldsymbol{r} = (1, 2, 1, 3)$, and $\boldsymbol{p} = \{\{1, 3\}, \{2\}, \{4\}\}$

3 Theoretical Analysis

Owed to the fact that volatility and VaR are latent quantities, model parameters cannot be chosen to minize volatility and VaR estimation errors, but are most commonly estimated to optimize fit to data as measured by (marginal) likelihood. This gives rise to posterior coefficient estimates, i.e. conditioned on the observed data. To enable a more general analysis allowing comparisons of different models, denote by $\mathcal{Q}'_t$ a general prior measure for the time t parameters $\boldsymbol{\theta}_t$. Then, let $\mathcal{Q}_t := \mathcal{Q}'_t \mid X_{1:n}$ denote their posterior measure, i.e. $\boldsymbol{\theta}_t \mid X_{1:n} \sim \mathcal{Q}_t$, which is obtained by conditioning $\mathcal{Q}'_t$ on the observed data. In the case of frequentist models, this corresponds to setting the parameters to their maximum likelihood estimate $\hat{\boldsymbol{\theta}}(X_{1:n})$, so that $\mathcal{Q}_t = \delta_{\hat{\boldsymbol{\theta}}_t(X_{1:n})}$, and, in the case of Bayesian models, to updating $\mathcal{Q}'_t$ according to Bayes' rule. If the model assumes no time-varying coefficients, we define $\mathcal{Q}_t = \delta_{\boldsymbol{\theta}_1}$ for $t \geq 2$.

Knowledge of the posterior coefficient measure $\mathcal{Q}_t$, or at least of properties thereof, allows to analyze how the quantities of interest are estimated. In the following, some properties of $\mathcal{Q}_t$ induced by DP priors are analyzed, compared to those of more classical GARCH-type models, and their adequacy for volatility and VaR estimation examined.

3.1 Model Properties

From the stick-breaking and CRP representation (see Section 2.2), many central properties of the prior measure can be deduced. However, it is not trivial that analogous properties apply to the posterior. Lemma 3.1 ensures tractability of properties of the posterior coefficient measure.

Lemma 3.1 (Conjugacy of Dirichlet process)**.** *If* $F \sim \mathcal{DP}(\alpha, G_0)$ *and* $\boldsymbol{\theta}_{1:n} \mid F \overset{iid}{\sim} F$*, then*

$$F \mid \boldsymbol{\theta}_{1:n} \;\sim\; DP(\alpha', G'_0)$$

where $\alpha' = \alpha + n$ *and* $G'_0 = \frac{1}{\alpha+n}\left(\alpha G_0 + \sum_{i=1}^{n} \delta_{\boldsymbol{\theta}_i}\right)$

Much like the Dirichlet distribution is conjugate w.r.t. the categorical distribution, the DP is conjugate w.r.t. infinite discrete probability distributions. Of course, in the context of mixture models, $\boldsymbol{\theta}_{1:n}$ is not observed but rather used as an input for the next hierarchical layer. Nonetheless, Lemma 3.1 ensures that properties analogous to that of the prior apply to the posterior measure.[4]

First, a random measure F generated from a DP is almost surely discrete. Consequently, there is always a strictly positive probability that there are repeated values in $\boldsymbol{\theta}_{1:n}$. In the context of time series data, this property is desirable, since it introduces recurrent regimes and dependence between observations.

Second, the DP induces not only almost sure repetitions in $\boldsymbol{\theta}_{1:n}$, but samples from its random measure inhibit *preferential attachment*. That is, the more frequently a sample has been re-sampled, the higher the probability that it will be re-sampled again. In GARCH models, this encourages the formation of few common and many rare regimes. This allows it to capture the risks of regime *shocks*, i.e. the possibility of brief yet sudden changes of the volatility dynamics, and is arguably in line with assumptions about empirical financial time series.

[4] If $X_t \mid \boldsymbol{\theta}_{1:n} \sim p(\cdot \mid \boldsymbol{\theta}_{1:n})$ and G_0 was conjugate w.r.t. $p(\cdot)$, then there would be conjugacy of the overall hierarchy (Gershman and Blei, 2012). Still, establishing conjugacy within the DP-GARCH would be hard, since $X_t \mid \boldsymbol{\theta}_{1:n}$ depends on an additional transformation $\sigma_t^2(\cdot)$ and finding a conjugate pair $(G_0, p(\cdot))$ respecting GARCH constraints is difficult.

Third, the distinct values $\boldsymbol{\theta}^*_{1:K}$ of $\boldsymbol{\theta}_{1:n}$ are all drawn iid from the base measure G_0. Thus, through the choice of G_0, prior knowledge about the average regime as well as the variety of regimes – essentially the *magnitude* of shocks – can be incorporated.

Fourth, the number K of distinct regimes produced by a DP grows like $\mathcal{O}(\alpha \log n)$ with the concentration parameter α and the sample size n. Particularly, $K \to 1$ as $\alpha \to 0$ and $K \to n$ as $\alpha \to \infty$. That is, the model DP-GARCH may infer anything from a single GARCH regime up to a separate regime at each time step. Prior knowledge about the probability of any number of regimes can be incorporated through the concentration parameter α.

Fifth, the regime assignments $\boldsymbol{r}$ generated by a DP are *exchangeable*. That is, any two indices, from identical or different clusters, can be swapped without affecting the probability of the overall partition. For GARCH models, this entails that there is no regime *persistence*, since for any permutation $s(\cdot)$,

$$\mathbb{P}(\boldsymbol{\theta}_{1:n}) \;=\; \mathbb{P}\big(\boldsymbol{\theta}_{s(1):s(n)}\big).$$

Arguably, this is not in line with what would be considered realistic in the context of time series models. Nonetheless, it is important for preserving tractability of the model.[5]

Lastly, in expectation, a random measure F from a DP assigns the same probability mass to every subset $A \subseteq \Theta$ of the domain as the base measure G_0. Additionally, the larger α, the more closely F will resemble G_0, which justifies its name of a "concentration" parameter. In particular, it holds that

$$\mathbb{E}_{F\sim\mathcal{DP}(\alpha,G_0)}\big[F(A)\big] \;=\; G_0(A), \tag{3.1}$$

$$\mathbb{V}_{F\sim\mathcal{DP}(\alpha,G_0)}\big(F(A)\big) \;=\; \frac{G_0(A)\big(1-G_0(A)\big)}{\alpha+1}. \tag{3.2}$$

3.2 Relation to Other Models

Knowledge of the key properties of the DP-GARCH facilitates illustration of how it relates and, often, directly extends more classical GARCH-type models.

First and most evidently, the DP-GARCH extends the standard GARCH model by Bollerslev (1986), which assumes that the coefficients (ω, α, β) are deterministic and fixed over time as opposed to stochastic and time-varying. In principle, the standard GARCH could be recovered from the DP-GARCH by specifying $G_0 = \text{Unif}(\mathbb{R})$, letting $\alpha \to \infty$, and collapsing the resulting posterior distribution to its mode ("maximum a posteriori").

Second, the DP-GARCH contains the classical Bayesian GARCH model as a special case. This model loosens the assumption that coefficients are deterministic by specifying a parametric prior distribution F_ϕ over them, which is updated after observing data $X_{1:n}$ according to Bayes' rule. Yet, they retain the assumption that coefficients are constant over time. From the DP-GARCH, the classical Bayesian GARCH model can be recovered by specifying $G_0 = F_\phi$ and letting $\alpha \to \infty$, which follows from Eqs. 3.1 and 3.2.

Third, consider the possibility of a hierarchical Bayesian GARCH model with separate distributions for each time step's parameters, i.e. $\boldsymbol{\theta}_t \,|\, \phi_t \sim F_{\phi_t}$ and $\phi_t \sim H$ for some distribution H. Then, as with the DP-GARCH, coefficients would follow a random measure. A continuous first-level prior distribution, e.g. $F_\phi \equiv G_0$, could incorporate the same distribution of distinct coefficients but would fail to produce recurrent regimes. The latter could be achieved by a

[5] Exchangeability allows to construct Gibbs-type samplers to conduct posterior inference, since any value can be assumed to be the last observed and its distribution conditioned on all other quantities.

discrete F_ϕ, yet this would come with a loss of flexibility of the distribution of distinct coefficients. For instance, specifying F_ϕ as a categorical distribution would require pre-specifying the distinct coefficient values, while most other discrete distributions have a pre-defined domain. Larger hierarchies would not be helpful, either. Hence, hierarchies of classical Bayesian models cannot simultaneously achieve the same control and richness as their nonparametric Bayesian counterparts.

Lastly, the DP-GARCH belongs to the class of regime-switching models. These models assume that the volatility is determined by one of K functions, or *regimes*, usually assumed to be of the form $f_k(X_{t-1}, \sigma_{t-1})$. The regime r_t at time t is considered latent. The class' most popular members are mixture and Markov-switching models (Gray, 1996; Bauwens, Preminger, and Rombouts, 2006; Haas and Paolella, 2012). Mixture models assume that regimes are independent. In contrast, Markov-switching models assume that r_t evolves according to a (first-order) Markov chain with transition probabilities $\mathbb{P}(r_t = j \,|\, r_{t-1} = j) = \boldsymbol{P}_{ij}$. Hence, mixture models are a special case of Markov-switching models, with the constraint that $\boldsymbol{P}_{ij} = p_j$ for all i, j. As coefficients $\boldsymbol{\theta}_t \,|\, F \sim F$ and hence regimes are iid, the DP-GARCH belongs to the class of mixture models. In stark contrast, however, stands that classical regime-switching models require pre-specifying the number of regimes K, which is hard in practice. The DP-GARCH avoids this issue by using the data to model a probability distribution over $K \in \{1, ..., n\}$.

3.3 Volatility & VaR Estimation

This section analyzes the adequacy of properties of the posterior coefficient measure of the DP-GARCH examined in Section 3.1 compared to those of related models outlined in 3.2 for estimation of volatility and VaR.

Volatility

Given a posterior coefficient measure $\mathcal{Q}_t$, a point estimator for volatility which is optimal in a least square and a least absolute error sense is the expectation and the median (Med) of its distribution induced by $\mathcal{Q}_t$, respectively:

$$\begin{aligned} \widehat{\sigma}_t^{LS}(X_{1:n}) &= \mathbb{E}_{\boldsymbol{\theta}_{1:n} \sim \mathcal{Q}_t}\left[\sigma_t(\boldsymbol{\theta}_{1:t})\right], \\ \widehat{\sigma}_t^{LA}(X_{1:n}) &= \mathrm{Med}_{\boldsymbol{\theta}_{1:n} \sim \mathcal{Q}_t}\left(\sigma_t(\boldsymbol{\theta}_{1:t})\right), \end{aligned}$$

where $\widehat{\mathrm{Med}}(A)$ denotes the sample median of the set A. If samples $\boldsymbol{\theta}_{i,t}$ can be obtained from $\mathcal{Q}_t$, these can be approximated relatively easily through Monte Carlo simulations.

A core advantage of stochastic volatility models is that they allow integration of estimation uncertainty through model averaging, in which each model is weighted by its posterior probability. The degree of model averaging then depends on the flexibility of the posterior coefficient measure $\mathcal{Q}_t$. In principle, higher flexibility allows to incorporate the possibility of a wider range of models and hence increases robustness.

For regime-switching GARCH models, flexibility is limited by the specified regimes. In particular, if identical models are specified as regimes, model averaging does not help modeling other volatility dynamics. For instance, if all regimes are standard GARCH models, volatility estimates are

$$\mathbb{E}_{\boldsymbol{\theta}_{1:n} \sim \mathcal{Q}_t}\left[\sigma_t(\boldsymbol{\theta}_{1:t})\right] = \left(\frac{1}{K}\sum_{k=1}^{K} p_k \omega_k\right) + \left(\frac{1}{K}\sum_{k=1}^{K} p_k \alpha_k\right) X_{t-1}^2 + \left(\frac{1}{K}\sum_{k=1}^{K} p_k \beta_k\right) \sigma_{t-1}^2$$

which is still linear in X_{t-1}^2 and σ_{t-1}^2. This argument remains true if K and p_k are themselves random and for $K \to \infty$, and hence the DP-GARCH.

A most commonly exploited benefit of Bayesian or general stochastic models lies in uncertainty quantification. In contrast to their non-stochastic counterparts, stochastic volatility models can adapt to model misspecifications by adjusting the uncertainty in the parameter estimates. This is formalized in Lemma 3.2.

Lemma 3.2 (Standardized moments of volatility)**.** *Let* $\sigma_t = \boldsymbol{\psi}_t^\top \boldsymbol{\theta}_t$*, and* $\boldsymbol{\theta}_t \sim \mathcal{Q}_t$*, where* $\boldsymbol{\psi}_t = \left(1, X_{t-1}^2, \sigma_{t-1}^2\right)^\top$*. Then,*

$$\widetilde{\mu}_k(\sigma_t \mid \mathcal{F}_{t-1}) \;=\; \frac{\mathbb{E}\left[\left(\boldsymbol{\psi}_t^\top \left(\boldsymbol{\theta}_t - \mathbb{E}[\boldsymbol{\theta}_t]\right)\right)^k\right]}{\left(\boldsymbol{\psi}_t^\top \mathbb{C}ov(\boldsymbol{\theta}_t)\boldsymbol{\psi}_t\right)^{k/2}}.$$

That is, $\mathcal{Q}_t$ may not only flexibly model overall uncertainty, i.e. variance, of volatility but additionally skewness, kurtosis, and other aspects of the distribution of σ_t.

Uncertainty quantification may also provide direct benefits for volatility estimation. For instance, in case of underspecification of innovation kurtosis, returns of large magnitude may be attributed to large variance instead of fat tails, thereby inducing overestimation of volatility. Hence, in principle, stochastic volatility models may be expected to improve volatility estimates in such cases. However, in the case of Gaussian distributions, maximum likelihood estimation is equivalent to matching empirical first raw and second central moments, and therefore empirical volatility. Nonetheless, the argument persists for the general case.

In summary, it is not expected that stochastic volatility models based on standard GARCH regimes outperform the base model in volatility estimation under misspecification of volatility dynamics. Their merits are theorized to be limited to underspecification of innovation kurtosis, but not in the case where specified innovations are Gaussian.

Value-at-Risk

VaR at the $100\alpha\%$ level, also simply α-VaR or VaR_α, is defined as the left α-quantile of the conditional return distribution. The present analysis is concerned with in-sample estimation thereof, and hence we define it conditional on all observations:

$$\begin{aligned}\text{VaR}_\alpha\left(X_t \mid X_{1:n}\right) \;&:= q^-_{X_t \mid X_{1:n}}(\alpha) \\ &:= \min\left\{x \in \mathbb{R} : \mathbb{P}(X \le x) \ge \alpha\right\}\end{aligned}$$

A distribution's quantiles are primarily determined by its location and scale. Hence, accurate volatility estimation is essential also for VaR estimation. However, a third strong determinant is shape. Lemma 3.3 illuminates how stochastic volatility models can flexibly model different shapes of returns.

Lemma 3.3 (Standardized moments of conditional returns)**.** *Let* $X_t = \sigma_t Z_t$*,* $\sigma_t = \boldsymbol{\psi}_t^\top \boldsymbol{\theta}_t$*, and* $\boldsymbol{\theta}_t \sim \mathcal{Q}_t$*, where* $\boldsymbol{\psi}_t = \left(1, X_{t-1}^2, \sigma_{t-1}^2\right)^\top$*. Then,*

$$\widetilde{\mu}_k(X_t \mid X_{1:n}) \;=\; \frac{\mathbb{E}\left[(\boldsymbol{\psi}_t^\top \boldsymbol{\theta}_t)^k\right]}{\mathbb{E}\left[(\boldsymbol{\psi}_t^\top \boldsymbol{\theta}_t)^2\right]^{k/2}} \widetilde{\mu}_k(Z_t) \;\ge\; \widetilde{\mu}_k(Z_t)$$

for $k \ge 0$ *and where expectations are taken w.r.t.* $\boldsymbol{\theta}_t \sim \mathcal{Q}_t$*.*

If $\boldsymbol{\theta}_t$ is deterministic, then $\widetilde{\mu}_k(X_t \mid \mathcal{F}_{t-1}) = \widetilde{\mu}_k(Z_t)$, i.e. the shape of conditional returns is fixed at that of the innovation distribution. However, if $\boldsymbol{\theta}_t$ is random, conditional returns can exhibit increased standardized moments. Therefore, although they may result in identical point estimates for volatility, stochastic models can flexibly adjust the shape of the conditional returns, also when the innovation distribution is not explicitly modeled.

A standardized moment of particular relevance in financial applications is the fourth one, which describes the probability weight concentrated in a distribution's tails and is referred to as *kurtosis*. In the context of GARCH models, from Lemma 3.3, conditional return kurtosis can be shown to be

$$\widetilde{\mu}_4(X_t \mid X_{1:n}) = \left(1 + \frac{\boldsymbol{\psi}_t^\top \mathbb{C}ov(\boldsymbol{\theta}_t)\boldsymbol{\psi}_t}{\left(\boldsymbol{\psi}_t^\top \mathbb{E}[\boldsymbol{\theta}_t]\right)^2}\right) \widetilde{\mu}_4(Z_t)$$

Since $\boldsymbol{\psi}_t \geq \mathbf{0}$ for all t, it follows that for fixed expectations, return kurtosis increases with the (co)variance of GARCH coefficients. The factor increase in kurtosis then depends on how flexibly the covariance structure of the GARCH parameters can be modeled.

Another important determinant of VaR is the asymmetry, or *skewness*, of the return distribution, corresponding to $k = 3$. Lemma 3.3 shows that, like other standardized moments, skewness increases proportionally. Unlike standardized moments of even order, however, skewness may be zero. Hence, this added flexibility only comes into play when the specified innovation distribution has non-zero skewness. Additionally, stochastic models cannot *reverse* the sign of the skewness of conditional returns specified in the innovation distribution.

Lemma 3.3 shows that the shape-modifying effect depends on $\boldsymbol{\psi}_t$, i.e. the previous time step's squared returns and volatility. This suggests that a *time-specific* coefficients distribution may be required to adequately adjust shape, where X_{t-1}^2 and σ_{t-1}^2 would have to be considered in their posterior distribution, regime transition probabilities, or similar. Otherwise, conditional returns may necessarily exhibit uncontrollably time-varying shape. However, no such models are considered in this analysis.

In summary, stochastic models are able to flexibly model the general shape conditional returns, albeit only by proportionally increasing standardized moments of the innovation distribution. The specific mechanism depends on the posterior coefficient measure, which in the case of infinite Bayesian mixtures is considerably richer than for its special cases. Hence, they are expected to improve VaR estimation performance.

Practical Limitations

Since posterior inference is hard in practice, a computational note is in order. For VaR, closed-form computation requires knowledge of the cumulative distribution function (CDF) of $X_t \mid X_{1:n}$. For non-stochastic volatility models, this corresponds to the innovation distribution with the corresponding standard deviation. Assuming Gaussian or Student-t innovations, VaR is simply a scaled version of volatility:

$$\widehat{\mathrm{VaR}}_\alpha(X_t \mid X_{1:n}) = \sigma_t(\hat{\boldsymbol{\theta}}_{1:t})\, q_{Z_t}(\alpha)$$

For stochastic models, however, the CDF and hence VaR is intractable due to their hierarchical structure involving multiple stochastic layers, in which a random output is used as input to generate another random quantity. Hence, they have to be estimated through simulation. One possibility is the procedure outlined in Algorithm 2. This approach is motivated by a law-of-large-numbers-like argument. Clearly, the empirical quantile of X converges almost surely to the true quantile of the distribution that X is being sampled from. This remains

Algorithm 2 VaR estimation procedure

for $t = 1$ to n **do**
  for $i = 1$ to M **do**
    Draw $\boldsymbol{\theta}_{i,t} \sim \mathcal{Q}_t$
    Compute $\sigma_{i,t} = \sigma_{i,t}(\boldsymbol{\theta}_{1:t})$
    Draw $Z_{i,t}$ from the innovation distribution
    Set $\widehat{X}_{i,t} = \sigma_{i,t}^2 Z_{i,t}$
  end for
  Set $\widehat{\mathrm{VaR}}_\alpha(X_t \mid X_{1:n}) = \widehat{q}_{\{\widehat{X}_{1,t},\ldots,\widehat{X}_{M,t}\}}(\alpha)$, where $\widehat{q}_A(\alpha)$ empirical α-quantile of set A
end for

true for the quantile of $f(X)$ if a transformation $f(\cdot)$ is applied to each sample (Doss, Flegal, Jones, and Neath, 2014, p. 2451). It could be argued that the law of large numbers still applies for hierarchical stochastic structures, and consequently that $\widehat{\mathrm{VaR}}_\alpha(X_t \mid X_{1:n})$ converges to $\mathrm{VaR}_\alpha(X_t \mid X_{1:n})$ for large enough M.

In practice, limitations are three-fold. First, only a finite number of samples can be obtained. For extreme quantiles, the required sample size can be large. In particular, the variance of the empirical α-quantile of an iid sample of size n is

$$\mathbb{V}\left(\widehat{q}_{X_{1:n}}(\alpha)\right) = \frac{\alpha(1-\alpha)}{n\, f\left(q_X(\alpha)\right)^2}$$

where $f(\cdot)$ is the density function of X. This number is only further increased by the hierarchical stochastic structure. Second, for most models, samples from $\mathcal{Q}_t$ can only be obtained through Markov chain algorithms and hence while incurring autocorrelation. This further slows down convergence compared to an iid setting. Third, the finite-sample properties of $\widehat{\mathrm{VaR}}_\alpha$ will heavily rely on the properties of the MCMC sampler. If certain regions of the domain are only inefficiently reached, $\widehat{\mathrm{VaR}}_\alpha$ will be characterized by high variance.

4 Simulation Study

The initial aim of this section was to test the performance of the DP-GARCH model for estimating volatility and VaR based on simulated data and compare it to more classical benchmark models. As it failed to produce reasonable results (see Section 2.2 of the appendix for details), only its special cases and otherwise related models are examined. The objective is to assess how the models deal with misspecifications in the volatility process as well as the innovation distribution.

4.1 Ground Truths

To test model performance under different types of misspecifications, data was generated from three distinct GARCH-type processes, each with two innovation distributions of varying shape.

Processes

To assess performance under misspecification of volatility dynamics, data was generated from the following three models:

$$\begin{aligned}
\sigma^2_{t,\text{GJR}} &= \omega + (\alpha + \gamma \mathbb{1}_{\{X_{t-1}<0\}}) X^2_{t-1} + \beta\, \sigma^2_{t-1} \\
\sigma^2_{t,\text{Exp}} &= \exp\left(\omega + \tilde{\alpha} Z_{t-1} + \gamma\big(|Z_{t-1}| - \mathbb{E}|Z_{t-1}|\big) + \beta \log \sigma^2_{t-1}\right) \\
\sigma^2_{t,\text{MS}} \mid r_t &= \begin{cases} \sigma^2_{t,\text{Bol}} & , r_t = 1 \\ \sigma^2_{t,\text{GJR}} & , r_t = 2 \\ \sigma^2_{t,\text{Exp}} & , r_t = 3 \end{cases} \\
\mathbb{P}(r_t = j \mid r_{t-1} = i) &= \boldsymbol{P}_{ij}
\end{aligned}$$

where $\sigma^2_{t,\text{Bol}}$ is Bollerslev's (1986) standard GARCH process from Eq. 2.1. The parameters are set to $\omega = 1$, $\alpha = 0.2$, $\tilde{\alpha} = 0$, $\beta = 0.4$, and $\gamma = 0.5$, as well as

$$\boldsymbol{P}_{ij} = \begin{cases} 0.8, & i = j \\ 0.1, & i \neq j \end{cases}.$$

The first two processes correspond to the GJR-GARCH model by Glosten, Jagannathan, and Runkle (1993) and the exponential GARCH (E-GARCH) by Nelson (1991), respectively. For the GJR-GARCH process, specifying $\gamma > 0$ incorporates an asymmetry w.r.t. the sign of returns, inducing volatility to increase more strongly in response to negative returns than to positive returns of equal magnitude ("leverage effect"). For the E-GARCH process, this effect is captured by $\tilde{\alpha}$. The third process is a Markov-switching GARCH (MS-GARCH) process, in which the volatility dynamics change over time in a first-order Markovian manner.

GARCH-type processes are characterized by volatility clustering, where high (low) magnitude returns tend to be followed by high (low) magnitude returns. The degree thereof is captured by the specific process' volatility persistence $\mathcal{P}$, which constitutes the rate at which the conditional variance regresses to the long-run level. For the specified processes, these can be shown to be

$$\begin{aligned}
\mathcal{P}_{\text{Bol}} &= \alpha + \beta &&= 0.6 \\
\mathcal{P}_{\text{GJR}} &= \alpha + \beta + \frac{1}{2}\gamma &&= 0.85 \\
\log \mathcal{P}_{\text{Exp}} &= \beta &&= 0.4
\end{aligned}$$

where $\log \mathcal{P}$ is the persistence of the log-volatility.

Volatility persistence is tightly related to long-run, or unconditional, volatility. The unconditional variances of the specified processes are [6]

$$\begin{aligned}
\mathbb{E}\big[\sigma^2_{*,\text{Bol}}\big] &= \frac{\omega}{1 - \alpha - \beta} &&= 2.5 \\
\mathbb{E}\big[\sigma^2_{*,\text{GJR}}\big] &= \frac{\omega}{1 - \alpha - \beta - \frac{1}{2}\gamma} &&= 6.\overline{6} \\
\mathbb{E}\big[\sigma^2_{*,\text{Exp}}\big] &\geq \exp\left(\frac{\omega}{1 - \beta}\right) &&\approx 5.3
\end{aligned}$$

[6] Processes' unconditional variances are obtained by taking expectations on both sides, setting $\sigma_t = \sigma_{t-1} = \sigma_*$, and solving for $\mathbb{E}[\sigma^2_*]$. For the E-GARCH process, $\exp(\omega/(1-\beta))$ constitutes the exponentiated unconditional log-variance. The inequality then follows from Jensen's inequality. Theoretically, however, its unconditional variance is unbounded, and in fact inexistent for overly leptokurtic innovations such as the Student-t.

which roughly correspond to the scale of log-returns observed in practice scaled by 100.

The necessary and sufficient conditions for covariance-stationarity are well-known to be $\alpha + \beta < 1$ for the standard GARCH process, $\alpha + \beta + \frac{1}{2}\gamma < 1$ for the GJR-GARCH process, and $|\beta| < 1$ for the E-GARCH process. These are well satisfied by the specified parameters. For the MS-GARCH process, covariance-stationarity of each individual process is sufficient for covariance-stationarity of the overall process. Hence, all simulated processes are covariance-stationary and, given correct specification, would be easily estimated.

Ultimately, with the specified parameters, the first process poses the challenge of an asymmetry with respect to the sign of returns. The second process is symmetric w.r.t. sign, yet its volatility dynamics are non-linear in nature. Finally, the third process entails time-varying volatility dynamics, switching between different structures, scales, and persistences of volatility. The specified transition matrix implies regime persistence and thereby tests the sensitivity of the assumption of independent regimes of mixture models, including the DP-GARCH.

Innovations

To assess performance under misspecification of the innovation distribution, innovations are generated either from a standard Gaussian or standardized Student-t distribution,

$$\begin{aligned} Z_t &\sim \mathcal{N} = \mathcal{N}(0,1) \\ Z_t &\sim t = \sqrt{\frac{\nu-2}{\nu}}\, t_\nu \end{aligned}$$

The Student-t distribution's degrees of freedom are set to $\nu = 5$. The latter's kurtosis is $\frac{\nu-2}{\nu-4}$ times larger than that of the Gaussian distribution, i.e. by a factor of 3 for $\nu = 5$.

Although in reality return kurtosis may be even more extreme, also models used in practice typically incorporate leptokurtosis. Ultimately, the analysis sheds light on relative underspecification of return kurtosis.

4.2 Benchmark Models

The volatility and VaR estimation performance of several models capturing a varying number of properties of the DP-GARCH, including Bayesian model averaging and regime-mixing, is compared. Their specific implementations are briefly outlined here and more extensively in Section 2 of the appendix.

All models assume Gaussian innovations and hence model conditional returns as

$$X_t \mid \sigma_t \sim \mathcal{N}(0, \sigma_t^2).$$

Thus, in the case of Student-t_5 innovations, underestimation of the innovation kurtosis by a factor of 3 is incorporated into all models. Yet, they vary in the way volatility σ_t and, the parameters $\boldsymbol{\theta}_{1:n}$ (if any) are modeled.

Higher-Order GARCH

Higher-order GARCH models consider past returns and volatilities going back more than one period. A GARCH(p, q) models the conditional variance as

$$\sigma_t^2 = \omega + \sum_{j=1}^{p} \alpha_j X_{t-j}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2$$

The parameters are estimated through MLE. For this analysis, the order is set to $p = q = 3$.

T-GARCH

The T-GARCH is another model incorporating the leverage effect. Similar to the E-GARCH, it does so by accounting for both a sign and a size effect. Specifically, it models the conditional variance as

$$\sigma_t^2 \;=\; \Big(\omega + \alpha|Z_t| + \gamma\, Z_t + \beta\, \sigma_{t-1}\Big)^2$$

Notice that the GJR-, E-, and T-GARCH all incorporate similar effects, yet their assumption on the functional form of the conditional variance is quite different, namely linear, exponential, and quadratic, respectively.

Nonparametric GARCH

Given the numerous distinct specifications of the volatility dynamics, Bühlmann and McNeil (2002) proposed to learn it from the data. Specifically, they model volatility as a (fixed) general function of the past returns and volatilities, so that

$$\sigma_t^2 \;=\; f(X_{t-1}, \sigma_{t-1}^2)$$

Since $\mathbb{E}\big[X_t^2 \,|\, \mathcal{F}_{t-1}\big] = \sigma_t^2$, f can be estimated by regressing X_t^2 on X_{t-1} and σ_{t-1}^2. By using flexible regression techniques, one can then theoretically capture sign effects, size effects, non-linearities, and various other structures. The only assumption is that the underlying process is fixed as opposed to regime-switching.

As f relies on the unobserved volatilities, the authors employ an iterative estimation procedure, in which initial estimates are obtained from a standard GARCH$(1,1)$ model.

In this analysis, f is estimated using random forest, and the resulting model referred to as *RF-GARCH*. Random forest is a learning algorithm based on bootstrap aggregation of de-correlated regression trees (Breiman, 2001). Tree-based learners can model very quickly varying functions. An advantage in the context of volatility modeling is that they produce predictions that are convex combinations of the training targets, thereby automatically guaranteeing non-negative estimates of the conditional variance.

The standard, higher-order, T- and nonparametric GARCH are all non-stochastic volatility models. Hence, the shape of conditional returns is fixed at that of the innovation distribution. The ones that follow are stochastic volatility models, flexibly incorporating larger standardized moments (see Section 3.3).

Bayesian GARCH

In classical Bayesian GARCH (*Bay-GARCH*) models of the form proposed by Geweke (1993), conditional variance is modeled as

$$\begin{aligned} \sigma_t^2 \;&=\; \omega + \alpha X_{t-1}^2 + \beta \sigma_{t-1}^2 \\ (\omega, \alpha, \beta) \;&\sim\; F_\phi \end{aligned}$$

where F_ϕ is a diffuse independent prior. Specifically, F_ϕ constitutes a truncated bivariate Gaussian prior on (ω, α) and an independent univariate truncated Gaussian prior on β, defined as

$$F_\phi \propto \mathcal{N}\big((\omega,\alpha)^\top; \boldsymbol{\mu}, \boldsymbol{\Sigma}\big)\, \mathcal{N}\big(\beta; \mu, \sigma^2\big) \cdot \mathbb{1}\{(\omega,\alpha,\beta)^\top > \mathbf{0}\}$$

The prior means and covariances are set to 0 and variances to $1'000$. Hence, the prior implies independence between all coefficients. For $(X_{t-1}^2, \sigma_{t-1}^2) = (5,5)$, the prior return kurtosis is approximately 1.24 times that of the innovation distribution. The posterior distribution is approximated using the MCMC sampler implemented by Ardia and Hoogerheide (2010).

Finite Mixture GARCH

Finite mixture GARCH models in the form proposed by Haas, Mittnik, and Paolella (2004b) with $K = 3$ separate standard GARCH regimes are used as a benchmark. The volatility process is then modeled as

$$\begin{aligned} \sigma_t^2 \mid (r_t = j) &= \omega_j + \alpha_j X_{t-1}^2 + \beta_j \sigma_{t-1}^2 \\ \mathbb{P}(r_t = j) &= p_j \end{aligned}$$

with $j \in \{1,2,3\}$. The parameters are estimated both through a MLE and Bayesian MCMC scheme, and the resulting model referred to as *Mix-GARCH* and *BayMix-GARCH*, respectively. In the latter, the prior distributions are diffuse and regime-independent. Model fitting and estimation is performed using the built-in functions of the implementation by Ardia, Bluteau, Boudt, Catania, and Trottier (2019).[7]

Markov-Switching GARCH

Markov-switching GARCH models in the form proposed by Haas et al. (2004a) with $K = 3$ separate standard GARCH regimes are used as a benchmark. The volatility process is then modeled as

$$\begin{aligned} \sigma_t^2 \mid (r_t = j) &= \omega_j + \alpha_j X_{t-1}^2 + \beta_j \sigma_{t-1}^2 \\ \mathbb{P}(r_t = j \mid r_{t-1} = i) &= \boldsymbol{P}_{ij} \end{aligned}$$

where $i, j \in \{1,2,3\}$. The parameters are estimated both through a MLE and Bayesian MCMC scheme, and the resulting models referred to as *MS-GARCH* and *BayMS-GARCH*, respectively. In the latter, the prior distributions are diffuse and regime-independent. Model fitting and estimation of volatility and VaR is performed using the built-in functions of the `R` implementation by Ardia et al. (2019).

4.3 Evaluation Criteria

All models are assessed based on in-sample estimation errors of volatility and VaR. Since the underlying data-generating processes are known, these otherwise latent quantities are available and can hence be directly compared to the models' estimates. Arguably, risk measures such as expected shortfall would better capture tail estimation performance, and out-sample, predictive performance of these quantities would be of higher practical relevance. However,

[7] Their documentation fails to mention how volatility and VaR is estimated in the Bayesian MCMC scheme. All that is mentioned is that "parameter uncertainty is integrated into the estimates" (Ardia et al., 2019, pp. 18–19). It is speculated that the approach is identical to the one laid out for the Bay-GARCH model.

both require significantly larger sample sizes, which are prohibitive given the complex estimation procedures required for nonparametric Bayesian methods.

Model performance is measured by root mean squared error (RMSE) and mean absolute deviation (MAD) of the estimated quantities of interest, defined as

$$\text{RMSE} := \mathbb{E}\Big[\Big(\frac{1}{n}\sum_{t=1}^{n}\big(\sigma_t - \widehat{\sigma}_t\big)^2\Big)^{\frac{1}{2}}\Big], \quad \text{MAD} := \mathbb{E}\Big[\frac{1}{n}\sum_{t=1}^{n}\big|\sigma_t - \widehat{\sigma}_t\big|\Big].$$

for volatility and analogously for VaR. Thereby, it is assumed that the impact of over- and underestimation is identical and unrelated to the scale of the true quantity. For a discussion on the adequacy of different loss functions for volatility and VaR estimation, see e.g. Ardia (2008, Ch. 6.3).

Of each process, $N = 40$ independent samples of length $n = 1'000$ are taken, so that the performance metrics are estimated as

$$\widehat{\text{RMSE}} = \frac{1}{N}\sum_{i=1}^{N}\Big(\frac{1}{n}\sum_{t=1}^{n}\big(\sigma_{i,t} - \widehat{\sigma}_{i,t}\big)^2\Big)^{\frac{1}{2}}, \quad \widehat{\text{MAD}} = \frac{1}{N}\sum_{i=1}^{N}\Big(\frac{1}{n}\sum_{t=1}^{n}\big|\sigma_t - \widehat{\sigma}_t\big|\Big)$$

where $\sigma_{i,t}$ is the time t volatility of the i-th sample, and analogously for VaR. Given independence of the samples, these constitute unbiased estimators for the error metrics. Additionally, estimation errors are assessed on an identical set of samples to ensure a more fair comparison.

It should be noted that relative model performance may vary for different sample sizes, and hence the present analysis is only valid conditional on observations being of length $n = 1'000$.

4.4 Results

This section presents and discusses the results of the simulation study. In-sample estimation performance for volatility and 5%-VaR of the benchmark models is illuminated. Estimation of the DP-GARCH model itself failed to produce reasonable results, which is discussed in Section 2.2 of the appendix.

Tables 1 and 2 report the models' mean volatility and 5%-VaR estimation errors from the simulations.[8] Bias-corrected standard errors are provided for single-sample estimation errors. To obtain standard errors for the error metrics, they can be divided by $N = 40$.

First, the Bayesian GARCH consistently outperformed standard GARCH, albeit sometimes negligibly. On average, the Bayesian GARCH reduced volatility RMSE by 14.5% in the case of Gaussian and by 9.2% in the case of Student-t innovations. A very similar observation is true for VaR. Although the performance benefit of the Bayesian GARCH was expected, the result is unexpected in that (i) performance benefits were smaller when return kurtosis was underspecified as opposed to correctly specified and (ii) they did not concentrate on VaR but were equally present for volatility estimation.

Second, for VaR estimation, the mixture GARCH yielded a significant improvement over the standard GARCH under exponential volatility dynamics with any innovations and the Markov-switching process with Student-t innovations. It even achieved the best overall performance for under exponential and Markov-switching dynamics with Student-t innovations.

[8] Both 1% and 5%-VaR estimation errors were estimated and can be found in the online appendix. Here, results for 5%-VaR are reported as it is thought that these are less affected by the practical limitations discussed at the end of Section 3.3. Indeed, the results for 5%-VaR favor Bayesian models slightly more than those for 1%-VaR, although the opposite could have been expected.

Table 1: Volatility estimation performance

Process	GJR–GARCH		E–GARCH		MS–GARCH	
	$Z \sim \mathcal{N}$	$Z \sim t$	$Z \sim \mathcal{N}$	$Z \sim t$	$Z \sim \mathcal{N}$	$Z \sim t$
RMSE:						
GARCH	0.419	0.549	0.243	0.589	0.373	0.527
	(0.065)	(0.276)	(0.135)	(0.988)	(0.055)	(0.426)
GARCH$(3,3)$	0.427	0.573	0.187*	0.581*	0.373*	0.546
	(0.069)	(0.273)	(0.093)	(1.000)	(0.049)	(0.412)
T–GARCH	**0.258***	**0.401***	0.153*	0.519*	0.368*	0.481*
	(0.060)	(0.225)	(0.097)	(0.854)	(0.054)	(0.349)
RF–GARCH	0.559	0.858	0.378	0.784	0.531	0.704
	(0.130)	(0.432)	(0.021)	(0.845)	(0.088)	(0.380)
Bay–GARCH	0.418*	0.542*	**0.142***	**0.487***	**0.367***	**0.479***
	(0.066)	(0.261)	(0.038)	(0.966)	(0.051)	(0.360)
Mix–GARCH	0.645	1.043	0.234*	1.127	0.486	0.824
	(0.696)	(0.952)	(0.225)	(1.383)	(0.526)	(0.795)
BayMix–GARCH	0.644	0.744	0.385	0.769	0.428	0.699
	(0.516)	(0.482)	(0.969)	(0.875)	(0.188)	(0.885)
MS–GARCH	0.663	1.276	0.308	1.456	0.561	0.929
	(0.559)	(1.338)	(0.411)	(1.536)	(0.437)	(1.040)
BayMS–GARCH	0.559	0.862	0.553	0.782	0.520	0.564
	(0.346)	(1.114)	(0.845)	(1.104)	(0.322)	(0.360)
MAD:						
GARCH	0.273	0.272	0.179	0.231	0.290	0.339
	(0.025)	(0.039)	(0.115)	(0.150)	(0.027)	(0.098)
GARCH$(3,3)$	0.277	0.282	0.131*	0.227*	0.289*	0.346
	(0.026)	(0.042)	(0.075)	(0.150)	(0.020)	(0.092)
T–GARCH	**0.145***	**0.197***	0.113*	0.203*	**0.270***	**0.306***
	(0.023)	(0.050)	(0.078)	(0.145)	(0.020)	(0.058)
RF–GARCH	0.337	0.438	0.288	0.402	0.372	0.440
	(0.038)	(0.085)	(0.013)	(0.088)	(0.023)	(0.052)
Bay–GARCH	0.272*	0.268*	**0.093***	**0.158***	0.288*	0.317*
	(0.026)	(0.034)	(0.019)	(0.116)	(0.020)	(0.050)
Mix–GARCH	0.342	0.325	0.145*	0.184*	0.311	0.330*
	(0.210)	(0.094)	(0.107)	(0.118)	(0.096)	(0.049)
BayMix–GARCH	0.477	0.527	0.292	0.488	0.350	0.542
	(0.510)	(0.489)	(0.668)	(0.540)	(0.188)	(0.855)
MS–GARCH	0.296	0.325	0.165*	0.251	0.323	0.338*
	(0.037)	(0.081)	(0.209)	(0.300)	(0.100)	(0.060)
BayMS–GARCH	0.382	0.626	0.511	0.531	0.441	0.405
	(0.255)	(1.127)	(0.855)	(0.915)	(0.325)	(0.212)

Mean (standard error) of in-sample volatility estimation errors, estimated from $N = 40$ simulations of length $n = 1'000$ of each process. The best score is marked in **bold** and improvements over the standard GARCH with an asterisk (*).

Table 2: Value-at-Risk estimation performance

Process	GJR–GARCH		E–GARCH		MS–GARCH	
	$Z \sim \mathcal{N}$	$Z \sim t$	$Z \sim \mathcal{N}$	$Z \sim t$	$Z \sim \mathcal{N}$	$Z \sim t$
RMSE:						
GARCH	0.687	0.915	0.400	0.978	0.614	0.861
	(0.107)	(0.509)	(0.222)	(1.590)	(0.091)	(0.692)
GARCH$(3,3)$	0.702	0.952	0.308*	0.966*	0.613*	0.892
	(0.113)	(0.502)	(0.152)	(1.606)	(0.080)	(0.666)
T–GARCH	**0.425***	**0.600***	0.252*	0.855*	0.609*	0.770*
	(0.099)	(0.322)	(0.159)	(1.360)	(0.088)	(0.563)
RF–GARCH	0.919	1.355	0.622	1.247	0.874	1.112
	(0.214)	(0.701)	(0.034)	(1.344)	(0.144)	(0.606)
Bay–GARCH	0.691	0.904*	**0.241***	0.835*	**0.604***	0.787*
	(0.108)	(0.481)	(0.062)	(1.550)	(0.085)	(0.582)
Mix–GARCH	0.875	0.959	0.295*	**0.670***	0.673	**0.737***
	(0.396)	(0.559)	(0.194)	(1.119)	(0.196)	(0.465)
BayMix–GARCH	0.860	1.233	0.386*	1.245	0.657	1.110
	(0.322)	(0.885)	(0.675)	(1.339)	(0.167)	(1.516)
MS–GARCH	0.790	1.021	0.336*	0.746*	0.698	0.782*
	(0.149)	(0.636)	(0.339)	(1.120)	(0.213)	(0.457)
BayMS–GARCH	0.813	1.472	0.537	1.198	0.791	0.897
	(0.230)	(1.882)	(0.612)	(1.635)	(0.426)	(0.545)
MAD:						
GARCH	0.448	0.454	0.295	0.415	0.477	0.568
	(0.041)	(0.094)	(0.190)	(0.293)	(0.044)	(0.176)
GARCH$(3,3)$	0.456	0.471	0.216*	0.402*	0.475*	0.580
	(0.042)	(0.097)	(0.123)	(0.288)	(0.034)	(0.164)
T–GARCH	**0.238***	**0.350***	0.186*	0.379*	**0.443***	0.516*
	(0.037)	(0.111)	(0.129)	(0.273)	(0.033)	(0.107)
RF–GARCH	0.554	0.691	0.474	0.619	0.612	0.696
	(0.062)	(0.149)	(0.022)	(0.156)	(0.038)	(0.090)
Bay–GARCH	0.450	0.452*	**0.159***	0.306*	0.475*	0.533*
	(0.043)	(0.086)	(0.030)	(0.243)	(0.034)	(0.100)
Mix–GARCH	0.562	0.428*	0.227*	**0.214***	0.506	**0.483***
	(0.359)	(0.079)	(0.174)	(0.092)	(0.135)	(0.052)
BayMix–GARCH	0.597	0.766	0.332	0.823	0.503	0.868
	(0.351)	(0.816)	(0.681)	(0.996)	(0.099)	(1.485)
MS–GARCH	0.476	0.451*	0.262*	0.268*	0.517	0.497*
	(0.052)	(0.088)	(0.340)	(0.152)	(0.154)	(0.055)
BayMS–GARCH	0.552	1.014	0.491	0.788	0.639	0.634
	(0.217)	(1.888)	(0.622)	(1.335)	(0.416)	(0.350)

Mean (standard error) of in-sample 5%-Value-at-Risk estimation errors, estimated from $N = 40$ simulations of length $n = 1'000$ of each process. The best score is marked in **bold** and improvements over the standard GARCH with an asterisk (*).

Yet, it often performed worse for volatility estimation. An analogous observation can be made for the MS-GARCH. This partially confirms the predictions made in Section 3.3, albeit the result under GJR dynamics are unforeseen.

Third, also Bayesian estimation compared to MLE of both the mixture and Markov-switching GARCH models significantly improved volatility estimation performance under Student-t innovations, reducing RMSE by as much as 25.2% in the case of the mixture and by 39.3% in the case of the Markov-switching model on average. Yet, it worsened it under Gaussian innovations. Moreover, counter to what was theorized in Section 3.3, it considerably worsened VaR predictions, even more so in the case of Student-t innovations.

Fourth, on average, the mixture GARCH outperformed the Markov-switching GARCH in both volatility and VaR estimation. This is true when they were estimated through MLE or Bayesian MCMC and even in the majority of the cases where the true dynamics were actually Markov-switching. This serves as an indication that ignoring regime-dependence does not negatively affect in-sample estimation performance of the conditional return distribution, at least when the model is misspecified in other areas, anyway.

Lastly, the T-GARCH consistently improved the standard GARCH. The benefit was not limited to the asymmetric GJR dynamics, but also the exponential and partly the Markov-switching ones. The GARCH$(3,3)$ had virtually identical performance as the standard GARCH – except under exponential dynamics with Gaussian innovations, where it interestingly outperformed the standard GARCH despite the absence of a higher order in reality. In contrast, the nonparametric GARCH estimation procedure based on random forests did not yield satisfying results in any of the examined settings.

Overall, the only models that were able to yield reliable improvements over the standard GARCH model – at least in the weak sense – were the Bayesian GARCH and T-GARCH. The T-GARCH dominated in the case of asymmetric GJR dynamics, perhaps predictably so. However, the Bayesian GARCH outperformed the T-GARCH under exponential and also Markov-switching dynamics.

It is remarked that estimation of the mixture and Markov-switching models proved to be somewhat unstable, both for MLE and Bayesian MCMC estimation. This can in part be derived from the significantly higher standard errors compared to other models. In fact, estimation of one of the mixture, Markov-switching, and their Bayesian counterparts failed in several attempts.[9] In these cases, which occurred primarily under GJR and exponential dynamics with Student-t innovations, the simulations thus had to be repeated. This may have skewed results in favor of these models. In any case, it may indicate sensitivity to assumptions about the data and perhaps, for Bayesian models, inefficiency discussed at the end of Section 3.3.

Figure 1 visualizes the 5%-VaR prediction errors of the different models for one random sample of the Markov-switching process with Student-t innovations and gives some indication of how they were affected by the misspecifications. In particular, at least two distinct patterns of errors can be observed for all models, indicating that they predictably failed to model the varying dynamics. Additionally, in all models except the T-GARCH, one of the patterns forms a "V"-shape, while the other forms a convex curve, indicating failure to model the asymmetry of the GJR-GARCH and exponential nature of the E-GARCH process, respectively. While, for this sample, the GARCH and Bay-GARCH produced virtually identical VaR estimates, the Mix-GARCH's are arguably systematically closer to the truth. The BayMix-GARCH failed to

[9] For frequentist models, initializing the optimizer with the mean coefficient values of MCMC samples increased but did not guarantee full stability.

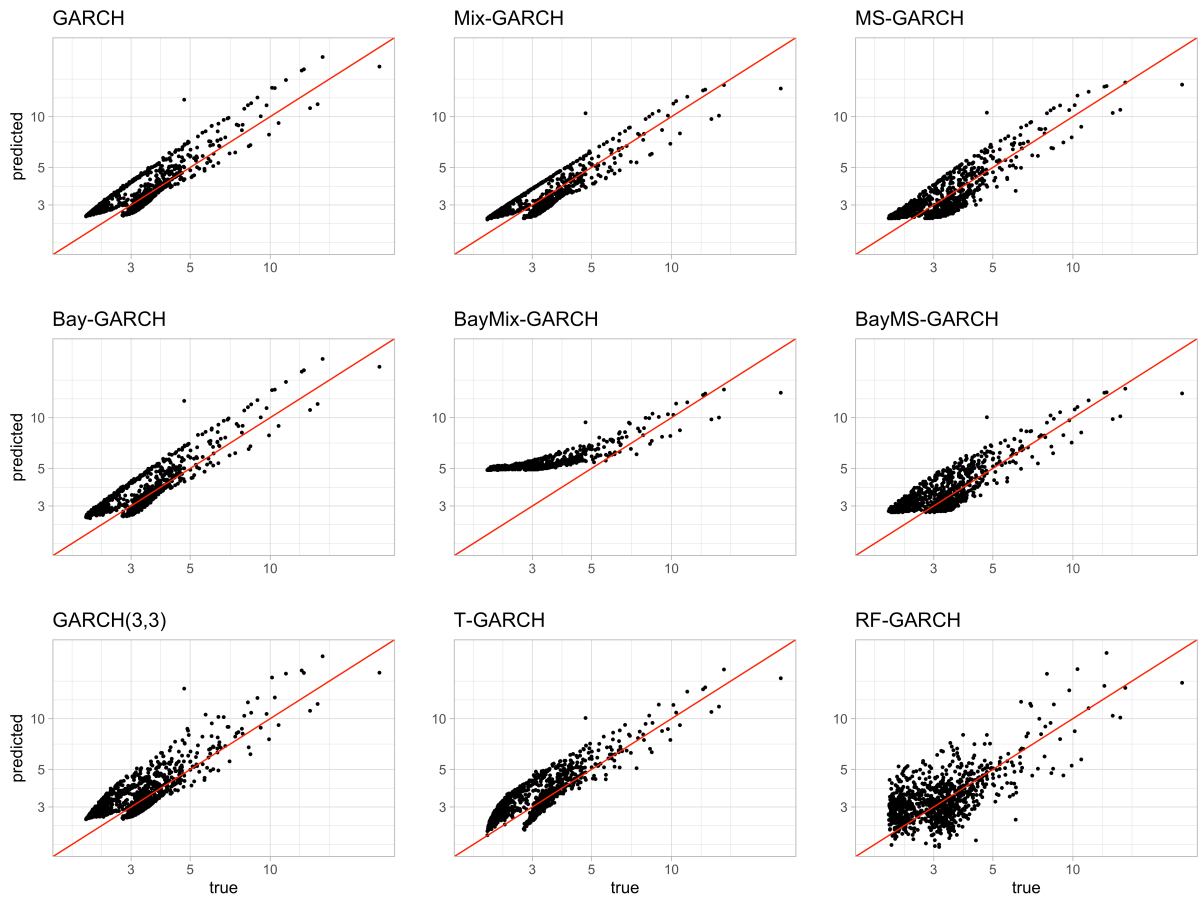

Figure 1: Plots of estimated (y-axes) against true (x-axes) in-sample 5%-Value-at-Risk from data simulated from the Markov-switching GARCH process with Student-t innovations. The red line is the identity function.

produce sufficient variance in its estimates, which may be responsible for the high variance of its estimation performance.

From these findings, one can only speculate how the DP-GARCH would have performed. Both Bayesian estimation and mixtures of GARCH processes seem to have their merits. Yet, these are not as clear-cut as theoretical analyses would suggest, especially when they are combined. On the one hand, the infinite mixture induced by the DP-GARCH might have allowed it to detect an optimal number of components. On the other hand, as estimation of multiple components becomes progressively harder, the resulting mixture might have become more arbitrary.

5 Conclusion

This paper examined the merits and demerits of extensions of the classic GARCH model – Bayesian extensions in general and infinite mixtures through the Dirichlet process in particular – for volatility and Value-at-Risk estimation. It presented arguments why stochastic models, including Bayesian and regime-switching ones, may theoretically yield improved estimation performance under misspecifications in the volatility dynamics and innovation distribution. In particular, it was shown that stochastic modeling of volatility does not only serve as a means to quantify uncertainty but can flexibly adjust the shape of conditional returns,

providing increased robustness. The benefits thereof were theorized to apply primarily to VaR estimation, but also estimation of volatility when modeling non-Gaussian innovations. These hypotheses were partially confirmed in a simulation study, in which the standard GARCH was consistently improved by Bayesian estimation in both aspects and by mixtures for VaR estimation. Less expectedly, Bayesian estimation improved the mixture model for volatility, but not VaR estimation. Despite the richness and appealing theoretical properties of Bayesian regime-switching models, it is speculated that their unsatisfying performance stems from inefficient estimation procedures. The high standard errors and sporadic failure of various regime-switching models – and perhaps also the complete failure of the DP-GARCH estimation algorithm – demonstrate the instability, inefficiency, and sensitivity of current algorithms to assumptions about the data. Finally, as the DP-GARCH estimation algorithm implemented by Lau and Cripps (2012) takes 13–15 hours to generate the 20'000 posterior samples conventionally utilized, the model is highly impractical. Hence, even assuming superior performance of nonparametric Bayesian models, their marginal benefit needs to be weighed against the additional time intensity.

Despite all, Bayesian regime-switching models remain worthy of future investigation. In this paper, some applications emerged as especially promising, including (i) to models with asymmetric innovations to increase robustness against misspecification of not only kurtosis but also skewness, (ii) to mixtures of heterogeneous regimes, such as linear and nonlinear ones, symmetric and asymmetric ones, etc., to increase flexibility of the modeled volatility dynamics, and (iii) to models with filtration-dependent coefficient distributions to allow for consistent modeling of the shape of the conditional return distribution.

Bibliography

Aldous, D. J. (1985). Exchangeability and Related Topics. In L. Bauwens, C. Hafner, and S. Laurent (Eds.), *École d'Été de Probabilités de Saint-Flour XIII — 1983*, Chapter 3, pp. 1–1198. Springer, Berlin.

Ardia, D. (2008). *Financial Risk Management with Bayesian Estimation of GARCH Models: Theory and Applications*, Volume 612. Springer Berlin, Heidelberg.

Ardia, D., K. Bluteau, K. Boudt, L. Catania, and D.-A. Trottier (2019). Markov-Switching GARCH Models in R: The MSGARCH Package. *Journal of Statistical Software 91*(4), 1–38.

Ardia, D. and L. F. Hoogerheide (2010). Bayesian Estimation of the GARCH(1,1) Model with Student-t Innovations. *The R Journal 2*(2), 41–47.

Ausín, M. C. and P. Galeano (2007). Bayesian Estimation of the Gaussian Mixture GARCH Model. *Computational Statistics & Data Analysis 51*, 2636–2652.

Bauwens, L., A. Preminger, and J. Rombouts (2006). Regime Switching GARCH Models. *CORE Discussion Papers*, 1–23.

Blackwell, D. and J. B. MacQueen (1973). Ferguson Distributions Via Polya Urn Schemes. *Annals of Statistics 1*(2), 353–355.

Blei, D. M. and M. I. Jordan (2006). Some Developments of the Blackwell-MacQueen Urn Scheme. *Bayesian Analysis 1*(1), 121–143.

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

Breiman, L. (2001). Random Forests. *Machine Learning 45*, 5–32.

Bühlmann, P. and A. J. McNeil (2002). An Algorithm for Nonparametric GARCH Modelling. *Computational Statistics & Data Analysis 40*(4), 665–683.

Doss, C. R., J. M. Flegal, G. L. Jones, and R. C. Neath (2014). Markov Chain Monte Carlo Estimation of Quantiles. *Electronic Journal of Statistics 8*, 2448–2478.

Ewens, W. J. (1988). Population Genetics Theory – The Past and the Future. In S. Lessard (Ed.), *Mathematical and Statistical Developments of Evolutionary Theory*, pp. 177–227. Montreal: University of Montreal Press.

Ferguson, T. S. (1973). A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics 1*(2), 209–230.

Friedman, J. H. (1991). Multivariate Adaptive Regression Splines. *The Annals of Statistics 19*(1), 1–141.

Gershman, S. J. and D. M. Blei (2012). A Tutorial on Bayesian Nonparametric Models. *Journal of Mathematical Psychology 56*, 1–12.

Geweke, J. (1993). Bayesian Treatment of the Independent Student-t Linear Model. *Journal of Applied Econometrics 8*(1), 19–40.

Ghalanos, A. (2022). *rugarch: Univariate GARCH models*. R package version 1.4-9.

Glosten, L. R., R. Jagannathan, and D. E. Runkle (1993). On the Relation between the Expected Value and the Volatility of the Nominal Excess Return on Stocks. *The Journal of Finance 48*(5), 1779–1801.

Gray, S. F. (1996). Modeling the Conditional Distribution of Interest Rates as a Regime-Switching Process. *Journal of Financial Economics 42*(1), 27–62.

Haas, M., J. Krause, M. S. Paolella, and S. C. Steude (2013). Time-Varying Mixture GARCH Models and Asymmetric Volatility. *North American Journal of Economics and Finance 26*, 602–623.

Haas, M., S. Mittnik, and M. S. Paolella (2004a). A New Approach to Markov-Switching GARCH Models. *Journal of Financial Econometrics 2*(4), 493–530.

Haas, M., S. Mittnik, and M. S. Paolella (2004b). Mixed Normal Conditional Heteroskedasticity. *Journal of Financial Econometrics 2*(2), 211–250.

Haas, M. and M. S. Paolella (2012). Mixture and Regime-Switching GARCH Models. In L. Bauwens, C. Hafner, and S. Laurent (Eds.), *Handbook of Volatility Models and Their Applications*, Chapter 3, pp. 71–102. John Wiley & Sons Inc.

Hastie, T. and R. Tibshirani (1986). Generalized Additive Models. *Statistical Science 1*(3), 297–318.

Lau, J. W. (2023, June). personal e-mail communication.

Lau, J. W. and E. Cripps (2012). Bayesian Non-Parametric Mixtures of GARCH(1,1) Models. *Journal of Probability and Statistics 2012*, 1–16.

Lau, J. W. and M. K. P. So (2009). A Monte Carlo Markov Chain Algorithm for a Class of Mixture Time Series Models. *Statistics and Computing 21*, 69–81.

Li, D., A. Clements, and C. Drovandi (2021). Efficient Bayesian estimation for GARCH-type models via Sequential Monte Carlo. *Econometrics and Statistics 19*, 22–46.

Liaw, A. and M. Wiener (2002). Classification and Regression by randomForest. *R News 2*(3), 18–22.

Nelson, D. B. (1991). Conditional Heteroskedasticity in Asset Returns: A New Approach. *Econometrica 59*(2), 347–370.

Pitman, J. (1996). Some Developments of the Blackwell-MacQueen Urn Scheme. *Statistics, Probability and Game Theory 30*, 245–267.

Savitzky, A. and M. J. E. Golay (1964). Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry 36*(8), 1627–1639.

Appendix

1 Proofs

Proof of Lemma 3.2. A random variable σ_t's k-th standard moment is

$$\begin{aligned}\tilde{\mu}_k(\sigma_t) &:= \mathbb{E}\left[\left(\frac{\sigma_t - \mathbb{E}[\sigma_t]}{\mathrm{sd}(\sigma_t)}\right)^k\right] \\ &= \frac{\mathbb{E}\left[\left(\sigma_t - \mathbb{E}[\sigma_t]\right)^k\right]}{\mathbb{V}(\sigma_t)^{k/2}}\end{aligned}$$

In the standard GARCH model, $\sigma_t = \boldsymbol{\psi}_t^\top \boldsymbol{\theta}_t$, where $\boldsymbol{\psi}_t = \left(1, X_{t-1}^2, \sigma_{t-1}^2\right)$. Hence,

$$\begin{aligned}\widetilde{\mu}_k\left(X_t \mid \mathcal{F}_{t-1}\right) &= \frac{\mathbb{E}\left[\left(\boldsymbol{\psi}_t^\top \boldsymbol{\theta}_t - \mathbb{E}[\boldsymbol{\psi}_t^\top \boldsymbol{\theta}_t]\right)^k\right]}{\mathbb{V}\left(\boldsymbol{\psi}_t^\top \boldsymbol{\theta}_t\right)^{k/2}} \\ &= \frac{\mathbb{E}\left[\left(\boldsymbol{\psi}_t^\top \left(\boldsymbol{\theta}_t - \mathbb{E}[\boldsymbol{\theta}_t]\right)\right)^k\right]}{\left(\boldsymbol{\psi}_t^\top \mathbb{C}ov(\boldsymbol{\theta}_t) \boldsymbol{\psi}_t\right)^{k/2}}\end{aligned}$$

□

Proof of Lemma 3.3. For standardized random variables Z_t, i.e. $\mathbb{E}[Z_t] = 0$ and $\mathbb{E}[Z_t^2]$, standard moments are

$$\tilde{\mu}_k(Z_t) = \mathbb{E}[Z_t^k]$$

For $X_t = \sigma_t Z_t$ with $\sigma_t \perp\!\!\!\perp Z_t$, they are

$$\begin{aligned}\tilde{\mu}_k\left(X_t \mid \mathcal{F}_{t-1}\right) &= \frac{\mathbb{E}\left[\left(\sigma_t Z_t - \mathbb{E}[\sigma_t Z_t]\right)^k\right]}{\mathbb{V}\left(\sigma_t Z_t\right)^{k/2}} \\ &= \frac{\mathbb{E}\left[\left(\sigma_t Z_t - \mathbb{E}[\sigma_t]\mathbb{E}[Z_t]\right)^k\right]}{\left(\mathbb{E}[\sigma_t^2]\mathbb{E}[Z_t^2] - \mathbb{E}[\sigma_t]^2\mathbb{E}[Z_t]^2\right)^{k/2}} \\ &= \frac{\mathbb{E}\left[\sigma_t^k\right]}{\mathbb{E}\left[\sigma_t^2\right]^{k/2}} \mathbb{E}[Z_t^k] \\ &= \frac{\mathbb{E}\left[\left(\boldsymbol{\psi}_t^\top \boldsymbol{\theta}_t\right)^k\right]}{\mathbb{E}\left[\left(\boldsymbol{\psi}_t^\top \boldsymbol{\theta}_t\right)^2\right]^{k/2}} \widetilde{\mu}_k(Z_t)\end{aligned}$$

Kurtosis is defined as the 4th standard moment. Setting $k = 4$ yields

$$\begin{aligned}
\widetilde{\mu}_4\big(X_t \mid \mathcal{F}_{t-1}\big) &= \frac{\mathbb{E}\big[\sigma_t^4\big]}{\mathbb{E}\big[\sigma_t^2\big]^2}\,\widetilde{\mu}_4(Z_t) \\
&= \frac{\mathbb{V}\big(\sigma_t^2\big) + \mathbb{E}\big[\sigma_t^2\big]^2}{\mathbb{E}\big[\sigma_t^2\big]^2}\,\widetilde{\mu}_4(Z_t) \\
&= \left(1 + \frac{\mathbb{V}\big(\sigma_t^2\big)}{\mathbb{E}\big[\sigma_t^2\big]^2}\right)\,\widetilde{\mu}_4(Z_t) \\
&= \left(1 + \frac{\boldsymbol{\psi}_t^\top \mathbb{C}ov(\boldsymbol{\theta}_t)\boldsymbol{\psi}_t}{\big(\boldsymbol{\psi}_t^\top \mathbb{E}[\boldsymbol{\theta}_t]\big)^2}\right)\,\widetilde{\mu}_4(Z_t)
\end{aligned}$$

□

Note that all expectations and (co)variances are taken w.r.t. $\boldsymbol{\theta}_t \sim \mathcal{Q}_t$, where $\mathcal{Q}_t$ is the posterior measure of $\boldsymbol{\theta}_t$.

2 Simulation Study

2.1 Implementation Details

The simulation study is based on code written in `R`, which was run on the ETH-Zurich Euler cluster. All code and data produced thereby is available in the online appenxix. In the following, the specific implementation of the simulation study is detailed.

Data generation: The data was generated using the simulation methods from the `rugarch` (Ghalanos, 2022) and `MSGARCH` (Ardia et al., 2019) package for the GJR- and E-GARCH resp. the MS-GARCH processes. Each sample was generated by simulating a time series of length $1'500$, of which only the last $1'000$ datapoints were used.

GARCH: Model fitting and estimation of volatility and VaR was done using the fitting methods from `rugarch` and `msgarch` for the standard GARCH and E-GARCH model resp. for the MS-GARCH models. A positivity but no stationarity constraint was imposed.

NP-GARCH: The NP-GARCH model was estimated through the algorithm outlined in Algorithm 3, which is an adapted version of the algorithm proposed by Bühlmann and McNeil (2002, p. 667). Their version additionally uses sample weights and performs a final smoothing step, in which $\hat{f}$ is estimated once more after averaging the volatilities over the last K iterations. However, the use of sample weights lead to eventual weights of 0 resp. ∞, while

Algorithm 3 Estimation of nonparametric GARCH model (Bühlmann and McNeil, 2002)

1: Fit a GARCH$(1,1)$ model to the data $X_{1:n}$
2: Set $\widehat{\sigma}_{0,t}$ to the estimated volatilities
3: **for** $m = 1$ to M **do**
4: Estimate $\hat{f}_m$ by regressing $X_{2:n}^2$ on $X_{1:n-1}$ and $\widehat{\sigma}_{m-1,1:n-1}^2$
5: Set $\widehat{\sigma}_{m,t}^2 \equiv \hat{f}_m(X_t, \widehat{\sigma}_{m-1,t})$
6: **end for**
7: Set $\widehat{\sigma}_t \equiv \frac{1}{K}\sum_{k=1}^{K} \widehat{\sigma}_{M-k+1,t}$

the final smoothing step significantly reduced performance. The hyperparameters were set to $M \equiv 20$ and $K \equiv 5$ in all cases. The chosen number of iterations M is considerably higher than suggested sufficient by the original authors (Bühlmann and McNeil, 2002, p. 671) since continued performance gains were observed during testing. Regressions were performed using `randomForest` based on the implementation by Liaw and Wiener (2002). The only non-default specifications were 2′000 base estimators. Also generalized additive models (Hastie and Tibshirani, 1986), mulivariate adaptive regression splines (Friedman, 1991), and local polynomial regression (Savitzky and Golay, 1964) were tested as regression methods, yet yielded worse results on average. The different behavior w.r.t. the use of weights, the final smoothing step, and the number of iterations might be explained by the fact that the original authors relied on yet another regression method, namely kernel regression.

Bay-GARCH: The Bayesian GARCH models were estimated using the `bayesGARCH` package by Ardia and Hoogerheide (2010). 20′000 samples of the model coefficients were generated from the posterior distribution. Thereof, the first 10′000 were discarded to burn in the Markov chain, yielding an effective sample size of $M = 10'000$. Volatility was computed across the full length time series for each draw and the result averaged over the posterior samples. That is, given the utilized draws $\boldsymbol{\theta}_i \sim F_{\widetilde{\phi}}$, $i = 1, ..., M$, of the GARCH coefficients, volatility estimates were computed as

$$\widehat{\sigma}_t \;=\; \frac{1}{M}\sum_{i=1}^{M}\sqrt{\omega_i + \alpha_i X_{t-1}^2 + \beta_i \widehat{\sigma}_{t-1}^2(\boldsymbol{\theta}_i)}$$

where $\widehat{\sigma}_t^2(\boldsymbol{\theta}_i) = \omega_i + \alpha_i X_{t-1}^2 + \beta_i \widehat{\sigma}_{t-1}^2(\boldsymbol{\theta}_i)$ is determined recursively with the initial value defined as $\widehat{\sigma}_1^2(\boldsymbol{\theta}_i) \equiv \omega_i$. To estimate VaR, for each sampled volatility level, one draw was taken from the corresponding return distribution, i.e. $\widehat{x}_{i,t} \sim \mathcal{N}\big(0, \widehat{\sigma}_t^2(\boldsymbol{\theta}_i)\big)$, and the empirical α-quantile of the resulting sample $\{\widehat{x}_{i,t}\}_{i=1}^{M}$ taken using the default algorithm of the `quantile` function. Since `bayesGARCH` is implemented for Student-t innovations only, the Gaussian innovations were approximated by fixing the degrees of freedom at 500.

(Bay) Mix- & MS-GARCH: All mixture and Markov-switching GARCH models, frequentist and Bayesian, were estimated using the `MSGARCH` package by Ardia et al. (2019). To estimate volatility and VaR, the built-in functions were used. For the Bayesian models, the same length burn-in period of 10′000 and number of posterior samples of 10′000 was used as for the Bay-GARCH. For the frequentist models, the mean parameter values of the same MCMC samples was used as initial values for the optimizer, with the purpose of increasing numerical stability.

DP-GARCH: Estimation of the DP-GARCH model was attempted using the implementation used in Lau and Cripps (2012), courteously provided by the authors.

2.2 Failure of the DP-GARCH Model

Here, the lack of reported results for the DP-GARCH model is elaborated.

Estimation of the model was attempted utilizing the implementation by Lau and Cripps (2012), which takes a time series $X_{1:n}$ as input and outputs posterior samples of GARCH coefficients at each time step, $\{(\omega_{i,t}, \alpha_{i,t}, \beta_{i,t})\}$, $t = 1, ..., n$, $i = 1, ...M$. The implementation was applied to the simulated data in the same way as the benchmark models, and no major modifications or re-specifications were made. In all attempts, the code ran error-free. However, the posterior samples obtained were highly unreasonable. Specifically, samples $\beta_{i,t}$ were consistently well above 1, especially for $t > 50$, giving rise to exploding volatility and hence VaR estimates. This was the case for early samples of the Markov chain and also samples as late as the $i \to 20'000^{th}$ iteration.

In response to this occurrence, several attempts were made to obtain reasonable estimates. First, the initialization of the partitioning of time indices were changed from one separating all time steps, i.e. $\boldsymbol{p} = \{\{1\}, \{2\}, ..., \{n\}\}$, to one separating only the first half from the second half of the time series, i.e. $\boldsymbol{p} = \{\{1, ..., \lceil \frac{n}{2} \rceil\}, \{\lceil \frac{n}{2} \rceil + 1, ..., n\}\}$. Theoretically, results should be robust against partition initialization (Lau, 2023). Second, the authors' implementation of other Bayesian nonparametric laws, including the Poisson-Dirichlet and the Normalized Generalized Gamma process, were used. Changing the initial partition only had the effect that MCMC coefficient samples for time steps after approximately $t = 20$ remained unchanged from their initializations over the entire length of the Markov chain, while the use of different laws had no noticeable effect. This might suggest extremely low acceptance probabilities within the Metropolis-Hastings algorithm. The results when using the Dirichlet, Poisson-Dirichlet, and Normalized Generalized Gamma processes were equally unreasonable.

The implementation was further applied to the same empirical data used by Lau and Cripps (2012) in their evaluation of nonparametric Bayesian GARCH models, which was also provided by the authors. They reported superior performance of these models compared to classical Bayesian GARCH models in terms of marginal likelihood (pp. 14–15). Although no explicit assessment of marginal likelihood was made, the posterior MCMC samples were again highly unreasonable and thus unlikely to have yielded useful results. Hence, the attempt failed to reproduce their findings. Some coefficient samples from the post-burning stage from both the attempt on simulated and the empirical data can be found in the online appendix.

Several conceivable explanations for this occurrence exist. First, according to one of the creators of the implementation, there might be an inefficiency in the proposal mechanism within the Metropolis-Hastings step (Lau, 2023). However, this does not explain the discrepancy between the observations made here and the results reported in Lau and Cripps (2012, pp. 14–15). Second, it is possible that in the time between the development of the implementation and the estimation attempt – a span of over 11 years – some mechanisms, such as utilized packages, have changed. Yet, such changes would have to be so subtle as to not produce any error messages and halt execution of the program. Third, to import the functions written in `C` into `R`, generation of a dynamic library file was necessary. The authors of Lau and Cripps (2012) only provided such a file compatible with `Windows` operating systems. To run the implementation on the `Linux`-based ETH Euler cluster, a `Unix`-compatible version thereof was created. This constitutes another potential point of complications. Yet, this step was performed with the help of a member of the ETH Cluster Support Team.

3 Online Appendix

The online appendix can be accessed here (https://github.com/kochnn/DP-GARCH). In particular, it includes

- data generated in the simulations and used for the numerical analysis, including some post-burnin posterior coefficient samples from the DP-GARCH;
- residuals plots of all models for all processes and metrics; and
- model implementations in `R` resp. `C`.

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of originality

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor.

Title of work (in block letters):

Dirichlet Process GARCH Mixtures for Volatility & Value-at-Risk Estimation

Authored by (in block letters):

For papers written by groups the names of all authors are required.

Name(s):	**First name(s):**
Koch	Nicolas Noël

With my signature I confirm that

- I have committed none of the forms of plagiarism described in the 'Citation etiquette' information sheet.
- I have documented all methods, data and processes truthfully.
- I have not manipulated any data.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for plagiarism.

Place, date	**Signature(s)**
Zurich, 26/06/2023	N. Koch

For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.