

Variational Inference for GARCH-family Models

Martin Magris, Alexandros Iosifidis
Department of Electrical & Computer Engineering
Aarhus University
Aarhus, Denmark
{magris, ai}@ece.au.dk

Abstract—The Bayesian estimation of GARCH-family models has been typically addressed through Monte Carlo sampling. Variational Inference is gaining popularity and attention as a robust approach for Bayesian inference in complex machine learning models; however, its adoption in econometrics and finance is limited. This paper discusses the extent to which Variational Inference constitutes a reliable and feasible alternative to Monte Carlo sampling for Bayesian inference in GARCH-like models. Through a large-scale experiment involving the constituents of the S&P 500 index, several Variational Inference optimizers, a variety of volatility models, and a case study, we show that Variational Inference is an attractive, remarkably well-calibrated, and competitive method for Bayesian learning.

Index Terms—Variational inference, Volatility, GARCH, Bayes

I. INTRODUCTION

The classical estimation procedures for GARCH-family models are the frequentist maximum likelihood, quasi maximum likelihood, and the generalized method of moments approaches [6]. Recently, there has been a growing interest in using Bayesian estimation techniques, as they offer several advantages over the traditional approaches [7]. For instance, in the frequentist approach, models are compared with no other means than their likelihood, whereas Bayes factors and marginal likelihood allow comparisons of non-nested models. Bayesian estimation provides reliable results in finite samples, and, e.g., can uncover the full value-at-risk density. Maximum likelihood estimators present some limitations when the errors are heavy-tailed and may not be asymptotically Gaussian [11], and positivity of the conditional variance and stationarity requirements can lead to complex non-linear inequalities, making constrained optimization cumbersome. For the Bayesian estimation of GARCH-family models, Monte Carlo (MC) sampling is the predominant approach [e.g., 11, 1, 28]. Indeed the recursive nature of the conditional variance makes the joint posterior of unknown parametric form, and the choice of the sampling algorithm is crucial. The Griddy-Gibbs sampler has extensively been used in this context [e.g., 5, 3], along with importance sampling [e.g., 9, 13], and the Metropolis-Hastings (MH) [18, 8]. Different extensions of the MH algorithm have been proposed [e.g., 19], along with the use of alternative methods [e.g., 2]. For a broader overview, see, e.g., the surveys [7, 28], or the textbook [1].

The ability of Bayesian methods to address uncertainty via posterior distribution gained much attention in Machine Learning (ML). In the last decade, sophisticated Bayesian

methods have been advanced for training high-dimensional ML models, and the theory of Bayesian neural networks has been extensively developed [see, e.g., 15, for a review]. Under the complexity of typical ML models, sampling methods do not scale up in high dimensions and difficult to apply. Variational Inference (VI) stands as a successful and feasible alternative, widely exploited in ML applications [12, 25, 14]. VI reduces the typical Bayesian integration problem to a simpler optimization problem aimed at finding the “best” approximation of the true posterior distribution in the sense described in Sec. II-B. Despite its consolidated use in ML, VI has not received much attention in econometrics and finance as a feasible Bayesian alternative to MC sampling.

In particular, the use of VI as a tool for the Bayesian estimation of GARCH-family models remains unaddressed. Though VI has been used in volatility forecasting with ML models [22, 21], there have been few self-contained VI applications concerning GARCH models [17, 16, 27]. This paper fills the this gap and addresses the feasibility and appropriateness of VI as an alternative to MC sampling and maximum likelihood estimation by conducting extensive experiments on the constituents of the S&P500 index. We show how Gaussian VI can be implemented and applied at a large scale as an unconstrained optimization problem through appropriate parameter transforms. We validate the results over several in-sample and out-of-sample performance metrics. Along with a focused study on the Microsoft Inc. stock data, with different optimization algorithms, we show VI is an effective, robust, and competitive approach for the Bayesian estimation of various GARCH-family models.

II. METHODS

A. GARCH models

A major task of financial econometrics is modeling volatility in asset returns. Volatility is considered a measure of risk for which investors demand a premium for investing in risky assets. Empirical observations of financial returns reveal some stylized facts. Whereas returns are nearly uncorrelated, they contain higher-order dependences. The correlation of absolute and squared returns is positive and persistent. Autocorrelated daily volatility thus appears to be predictable, and Autoregressive Conditional Heteroskedasticity (ARCH) models provide the basis for the most popular parameterizations of this dependence.

Let ε_t be a random variable that has mean and variance conditionally on the information set \mathcal{F}_{t-1} (σ -algebra generated by $\varepsilon_{t-j}, j \geq 1$). For the ARCH model, $\mathbb{E}(\varepsilon_t|\mathcal{F}_{t-1}) = 0$ and the conditional variance $h_t = \mathbb{E}(\varepsilon_t^2|\mathcal{F}_t)$ is a parametric function on \mathcal{F}_{t-1} [26]. The sequence $\{\varepsilon_t\}$ may be observed or be an error innovation sequence of an econometric model: $\varepsilon_t = r_t - \mu_t(r_t)$ with r_t and observable random variable (e.g., a daily return) and $\mu_t(r_t)$ the conditional expectation of r_t given \mathcal{F}_{t-1} . The parametric form of the ARCH model reads:

$$\begin{aligned} r_t &= \mu + \varepsilon_t, \\ \varepsilon_t &= h_t^{1/2} z_t, \quad z_t \sim iid \mathcal{N}(0, 1), \\ h_t &= \omega + \alpha_1 \varepsilon_{t-1}^2, \end{aligned}$$

with $t = 1, \dots, T$ and, to ensure $h_t > 0$ and identification, $\omega > 0$, $0 < \alpha_1 < 1$. We shall assume $\mu \equiv 0$. h_t is the conditional and time-dependent volatility of ε_t . The way ε_t is defined guarantees white noise properties, since z_t is a sequence of iid variables. Normality is a typical assumption for the iid sequence $\{z_t\}$, but leptokurtic alternatives are also used. The iid assumption guarantees the white noise property of $\{\varepsilon_t\}$. In the ARCH equation defining the parametric form for the conditional variance, the linear function of the squared innovation at $t - 1$ can be generalized to a higher-order ARCH(q):

$$h_t = \omega + \sum_{j=1}^q \alpha_j \varepsilon_{t-j}^2,$$

where $\omega > 0$, $\alpha_j \geq 0$, with at least an $\alpha_j > 0$. Note that the volatility, the object of modeling, is not observed: using ε_t^2 is an immediate solution, but alternatives exist if, e.g., the data is available at intraday frequencies.

For the ARCH family, the decay rate in the autocorrelation of ε_t^2 is too rapid compared to the observed time series: the so-called Generalized ARCH (GARCH) is a predominant alternative. In a GARCH(p, q) model, the conditional variance is not only a function of the lagged innovations but also of its lags:

$$h_t = \omega + \sum_{j=1}^q \alpha_j \varepsilon_{t-j}^2 + \sum_{j=1}^p \beta_j h_{t-j}.$$

The overwhelming model has been the GARCH(1, 1). Sufficient conditions for the positivity of conditional variances are $\omega > 0$, $\alpha_j \geq 0$, $j = 1, \dots, q$ and $\beta_j \geq 0$, $j = 1, \dots, p$. Identifiability requires at least one $\beta_j > 0$ and one $\alpha_j > 0$, and for stationarity $\sum \alpha_j + \sum \beta_j < 1$.

GARCH models have been extended and generalized in many different directions. Among these, the empirical evidence of asymmetry in volatility clustering motivates the GJR-GARCH [10] model, which assumes the response of the variance to a shock not to be independent of its sign:

$$h_t = \omega + \sum_{j=1}^q \alpha_j + \sum_{j=1}^o \gamma_j I(\varepsilon_{t-j} > 0) \varepsilon_{t-j}^2 + \sum_{j=1}^p \beta_j h_{t-j},$$

with $I(\cdot)$ an indicator function, defines the GJR-GARCH(p, o, q) (with $o = 0$ we simply write GJR-GARCH(p, q)). It must hold $\omega > 0$, $\alpha_j \geq 0$, $\beta_j \geq 0$, $\gamma_j \geq 0$, $\sum \alpha_j + \gamma_j \geq 0$, and $\sum \alpha_j + 1/2 \sum \gamma_j + \sum \beta_j < 1$.

The Exponential GARCH (EGARCH) model is another popular extension. The family of EGARCH(p, q) models can be defined with

$$\log h_t = \omega + \sum_{j=1}^q g_j(z_{t-j}) + \sum_{j=1}^p \beta_j \log h_{t-j}.$$

In our analyses we adopt the version of [20] where $g_j(z_{t-j}) = \alpha_j z_{t-j} + \psi_j(|z_{t-j}| - \mathbb{E}(|z_{t-j}|))$. The model does not impose any restriction on the parameters because, since the equation is on log variance instead of variance itself, the positivity of the variance is automatically satisfied. This is a big advantage in model estimation. For a concise presentation of the advantages and limitations of the EGARCH model, refer, e.g., to [26]. By including $\gamma_j I(\varepsilon_{t-j} > 0) \varepsilon_{t-j}^2$, $j = 1, \dots, o$ terms in the above conditional variance equation, one defines the GJR-EGARCH(p, o, q).

In our analyses, we furthermore adopt the Fractionally Integrated GARCH (FIGARCH). The FIGARCH model [4] conveniently explains the slow decay in autocorrelation functions of squared observations of typical daily return series. With the FIGARCH, the effect of the lagged ε_t^2 on h_t decays hyperbolically as a function of the lag length. The FIGHARCH(p, d, m) process is defined as:

$$(1 - L)^d \phi(L) \varepsilon_t^2 = \bar{\omega} + (1 - \beta(L)) v_t,$$

where L is the lag operator, $\phi(L) = \sum_{j=1}^{m-1} \phi_j L^j$, $\beta(L) = \sum_{j=1}^p \beta_j L^j$, $v_t = \varepsilon_t^2 - h_t^2$, and d is the order of fractional differencing that guides the long-memory properties of the process [26]. Of relevance for estimation is its equivalent ARCH(∞) representation of the model:

$$h_t = \omega + \sum_{j=1}^{\infty} \lambda_j \varepsilon_{t-j}^2, \quad (1)$$

where $\omega > 0$, and $\lambda_k \geq 0$ are recursively defined. For the FIGARCH (1, d , 1), $\delta_1 = d$, $\lambda_1 = \phi - \beta + d$, $\delta_k = (k - 1 - d)/k \delta_{k-1}$, $\lambda_k = \beta \lambda_{k-1} + \delta_k - \phi \delta_{k-1}$, with the constraints $\omega > 0$, $0 \leq d \leq 1$, $0 \leq \phi \leq (1 - d)/2$, $0 \leq \beta \leq d + \phi$, sufficient to ensure the positivity of the conditional variance [4].

1) *Estimation:* With a possibly misspecified but convenient standard likelihood function and the assumption that the dynamic of the volatility process is correctly specified, the models described earlier are generally estimated via Quasi Maximum Likelihood (QML). Under a Gaussian likelihood, the QML objective generally reads:

$$\ell(\boldsymbol{\nu}) = \sum_{t=1}^T \left(\log h_t(\boldsymbol{\nu}) + \frac{\varepsilon_t^2}{h_t(\boldsymbol{\nu})} \right), \quad (2)$$

where $\boldsymbol{\nu}$ collects all the relevant parameters, e.g., for the GARCH(1, 1), $\boldsymbol{\nu} = (\omega, \alpha_1, \beta_1)$, and the dependence of the

conditional variance on it, is made explicit. Constrained gradient descent procedures are effective for minimizing (2). Sec. II-C discusses using parameter transforms to perform unconstrained optimization. Eq. (2) implies a recursive relation whose implementation is expensive. For initialization, it is common to back-cast $\max\{p, o, q\}$ values with the average value of $\{r_t^2\}$.

B. Variational Inference

1) *General principle:* Let y denote the data and $p(y|\theta)$ the likelihood of the data based on a postulated model with $\theta \in \Theta$ a d -dimensional vector of model parameters. Let $p(\theta)$ be the prior distribution on θ . The goal of Bayesian inference is the posterior distribution $p(\theta|y) = p(y, \theta)/p(y)$, where $p(y) = \int_{\Theta} p(y|\theta)p(\theta)d\theta$. Bayesian inference is generally difficult since the marginal likelihood $p(y)$ is often intractable and of unknown form, and Variational Inference (VI) is an attractive alternative.

VI consists of an approximate method approximating the posterior distribution by a probability density $q(\theta)$ (called variational distribution) belonging to some tractable class of distributions \mathcal{Q} . VI thus turns the Bayesian inference problem into that of finding the best approximation $q^*(\theta) \in \mathcal{Q}$ to $p(\theta|y)$ by minimizing the Kullback-Leibler (KL) divergence from $q(\theta)$ to $p(\theta|y)$:

$$q^* = \arg \min_{q \in \mathcal{Q}} \text{KL}(q||p(\theta|y)) = \arg \min_{q \in \mathcal{Q}} \int q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta.$$

It can be shown the KL minimization is equivalent to the maximization of the so-called Lower Bound (LB) on $\log p(y)$, [e.g. 27]:

$$\mathcal{L}(q) := \int q(\theta) \log \frac{p(y|\theta)p(\theta)}{q(\theta)} d\theta = \mathbb{E}_q[h(\theta)], \quad (3)$$

with $h_{\zeta}(\theta) = \log p(y|\theta) + \log p(\theta) - \log q_{\zeta}(\theta)$. In fixed-form VI, the parametric form of the variational posterior is set. Typically, the target is a Gaussian distribution of mean μ and covariance Σ , and q_{ζ} in the set \mathcal{Q} of Gaussian distributions, with $\zeta = \{\mu, \text{vec}(\Sigma)\}$ a vector of parameters. VI seeks the parameter ζ^* optimizing (3).

The standard approach for maximizing the LB is based on a stochastic gradient descent update, whose basic form is

$$\zeta_{t+1} = \zeta_t + \delta \left[\mathcal{I}_{\zeta}^{-1} \hat{\nabla}_{\zeta} \mathcal{L}(q_{\zeta}) \right] \Big|_{\zeta=\zeta_t}, \quad (4)$$

where t denotes the iteration, δ a the step size, and $\hat{\nabla}_{\zeta} \mathcal{L}(q_{\zeta})$ a stochastic estimate of the Euclidean gradient. In place of Euclidean gradients, the recent literature adopts natural gradients leading to improved step directions by accounting for the information geometry of the variational distribution [see, e.g., 15]. With natural gradients, \mathcal{I}_{ζ}^{-1} is the corresponding Fisher Information Matrix, otherwise is the identity matrix I , of size equal to the number of trainable parameters d .

2) *Algorithms:* A major aspect of implementing (4) is the gradient computation. Methods requiring the actual computation of the gradients of the loss, such as the reparametrization trick [12], are expensive to implement at a large scale within the recurrent form of the likelihood (2). Furthermore, the use of automatic differentiation is not a widespread practice in econometrics and finance, largely adopting numerical differentiation. The approaches discussed here rely on the use of the log-derivative trick for evaluating the gradient of the expectation $\mathbb{E}_{q_{\zeta}}[h_{\zeta}(\theta)]$ as an expectation of a gradient:

$$\tilde{\nabla}_{\zeta} \mathcal{L}(q_{\zeta}) = \mathcal{I}_{\zeta}^{-1} \mathbb{E}_{q_{\zeta}}[\nabla_{\zeta} [\log q_{\zeta}(\theta)] h_{\zeta}(\theta)]. \quad (5)$$

Algorithm 1 sketches the gradient-free optimization approach. Note that at each iteration, the expectation in (5) is approximated with S samples from the posterior q_{ζ_t} . Different optimization algorithms differ in how ζ is defined (e.g., it updating a natural parameter), in how natural-gradient computations are performed, and in the adoption of alternatives forms for (4) (e.g., using a retraction in manifold optimization). ML research widely adopts a Gaussian prior of zero mean and covariance matrix τI , with $\tau > 0$.

Algorithm 1 General form of a gradient-free VI optimizer

Set hyperparameters (here β, S, τ), $t = 0$

Set initial values ζ_0

repeat

 Simulate $\theta_s \sim q_{\zeta_t}$, for $s = 1, \dots, S$

$h_{\zeta_t}(\theta) = \log p(\theta) + \log p(y|\theta) - \log q_{\zeta_t}(\theta)$

$\tilde{\nabla}_{\zeta_t} \mathcal{L} = \frac{1}{S} \sum_{s=1}^S \nabla_{\zeta} \log q_{\zeta}(\theta_s) \Big|_{\zeta=\zeta_t} \times h_{\zeta_t}(\theta_s)$

$\zeta_{t+1} = \zeta_t + \beta \tilde{\nabla}_{\zeta_t} \mathcal{L}$

$t = t + 1$

until stopping criterion is met

We briefly introduce the three state-of-the-art optimizers adopted in the empirical analysis. The gradient in (5) is often called a black-box gradient. Despite the terminology, the black-box approach has not to be intended as an opaque mechanism, but as a transparent and accessible solution for computing lower bound's derivatives without explicitly requiring model's derivatives. The expectation in (5), as of Algorithm 1, is computed as an average of products between the easy-to-derive gradients of the variational loglikelihood $\nabla_{\zeta} \log q_{\zeta}(\theta_s)$ computed in $\zeta = \zeta_t$ and the h -function $h_{\zeta_t}(\theta_s)$, so that the computation of $\tilde{\nabla}_{\zeta_t} \mathcal{L}$ involves only h -function's queries, and not its gradients w.r.t. ζ .

Black-box VI (BBVI), [25] uses the rule (4) applied to Euclidean black-box gradients, computed as in (5). Quasi-Black box VI (QBVI) [17] extends BBVI using natural gradients. QBVI relies on a natural-parameter parametrization of the variational posterior enabling natural gradient updates without requiring the explicit computation and inversion of the Fisher matrix. This is a relevant computational advantage. BBVI and QBVI are broadly applied under a diagonal covariance matrix specification and a log-variance parametrization, as they cannot guarantee the positive definiteness of the variational covariance matrix. Conversely, the two are of low complexity

as matrix operations (especially inversion) are straightforward. Manifold Gaussian Variational Bayes (MGVB) is a black-box approach, boosted by natural gradients, relying on manifold optimization to grant the positive definiteness of the full covariance matrix [27]. MGVB solves the positive definiteness issue while allowing for additional modeling flexibility provided by its full covariance specification. Certain theoretical issues and some approximations that MGVB relies upon are resolved by the Exact Manifold Gaussian Variational Bayes (EMGVB) approach, that further improves the computation of the natural gradients [16].

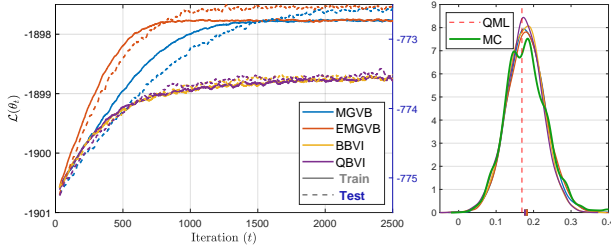


Fig. 1. Lower bound optimization (left) and posterior distribution of for the γ parameter (right). GJR-GARCH(1,1), Microsoft Inc. data.

C. Transformations

It is clear that the application of Gaussian VI is problematic for the heavily constrained volatility models of Sec. II-A. For instance, a Gaussian posterior is incompatible with the $\omega > 0$ or the $0 < d < 1$ requirements and plausibly with a Gaussian covariance structure (Gaussian copula). Moreover, the adoption of Gaussian priors is also inadequate. These issues can be fixed with appropriate parameter transforms.

In Algorithm 1, two steps are critical: (i) sampling from the Gaussian variational posterior and (ii) evaluating the model log-likelihood. By adopting the VI Gaussian framework, the unconstrained components of a sample θ from $q_{\zeta} = \{\mu, \text{vec}(\Sigma)\}$ need to be appropriately transformed into the valid constrained space for evaluating the log-likelihood. This is done, e.g., in [27] for the GARCH(1,1), and aligned with the well-adopted rationale for VI in medium-scale ML models of [14].

Let ν denote a d -dimensional parameter parametrizing a GARCH-family model m , and \mathcal{C}_m the constrained parameter space where ν lives. Be $\psi_m : \mathcal{C}_m \mapsto \mathbb{R}^d$ a transform that maps $\nu \in \mathcal{C}_m$ to $\theta \in \mathbb{R}^d$. Let ψ_m^{-1} denote the corresponding inverse transform, i.e., $\nu = \psi_m^{-1}(\theta)$. This is the relevant transform in VI; we will show that such ψ_m^{-1} exists and is of simple form for the models in Sec. II-A. Through ψ_m^{-1} we can apply the update (4), map posterior samples $\theta \sim q_{\zeta}$ into \mathcal{C}_m as $\psi_m^{-1}(\theta)$, evaluate the likelihood, and approximate the expectation in (5). Similarly, we can, e.g., compare the QML estimates with the mean of transformed posterior samples, interpretable as approximations of the transformed posterior living in \mathcal{C}_m :

$$\mathbb{E}_{q_{\zeta^*}}[\psi_m^{-1}(\theta^*)] \approx \frac{1}{S} \sum_{s=1}^S \psi_m^{-1}(\theta_s^*), \quad \theta_s^* \sim q_{\zeta^*}. \quad (6)$$

In principle, QML could optimize $\ell(\psi_m^{-1}(\theta))$, yet the use of constrained optimization is preponderant (e.g., in Python's `arch` and R's `rugarch` packages), so that parameter transforms are not relevant for standard QML estimation. Indeed, we are unaware of any work presenting such transformations. As fundamental for applying VI, and for future reference, we summarize them in the following propositions. For an element θ_{λ} of the vector θ representing a certain parameter λ of the model m , be θ_{λ} its corresponding element in θ . Let f denote the logistic function.

Proposition 1 (Inverse transforms for the FIGARCH): The FIGARCH constraints $\omega > 0$, $0 \leq d \leq 1$, $0 \leq \phi \leq (1 - d)/2$, $0 \leq \beta \leq d + \phi$, are satisfied by the following inverse transforms:

$$\begin{aligned} \omega &= \exp(\theta_{\omega}), & d &= f(\theta_d), \\ \phi &= f(\theta_{\phi})(1 - d)/2, & \beta &= f(\theta_{\beta})(\phi + d). \end{aligned}$$

Proof. The result follows immediately from [4, footnote 19].

Proposition 2 (Inverse transforms for the GJR-GARCH): The GJR-GARCH(p, o, q) constraints $\omega > 0$, $\alpha_i \geq 0$, $\sum \gamma_k + \sum \alpha_i \geq 0$, $\beta_j \geq 0$, $\sum (\alpha_i + 1/2\gamma_k + \beta_j) < 1$, for $j = 1, \dots, p$, $k = 1, \dots, o$, $i = 1, \dots, q$, for the GJR-GARCH(1,1) model are satisfied by the following inverse transforms:

$$\begin{aligned} \omega &= \exp(\theta_{\omega}), & \alpha &= f(\theta_{\alpha}), \\ \gamma &= f(\theta_{\gamma})(2(1 - \alpha) + \alpha) - \alpha, & \beta &= f(\theta_{\beta})(1 - \alpha - 1/2\gamma). \end{aligned}$$

For the general GJR-GARCH(p, o, q) case, inverse transforms can be computed as for Algorithm 2.

Proof. The transform for θ_{ω} is obvious. As γ and β are still to be determined, the constraints imply that α can lay anywhere in $[0, 1]$, so $\alpha = f(\theta_{\alpha})$. It is required that $\gamma + \alpha > 0$, and $\alpha + \beta + 1/2\gamma < 1$. Yet β is to be determined, so $1/2\gamma < 1 - \alpha$. The two give $-\alpha < \gamma < 2(1 - \alpha)$: we first map γ to $[2(1 - \alpha) + \alpha]$ and then shift the interval by $-\alpha$. I.e., $\gamma = f(\theta_{\gamma})[2(1 - \alpha) + \alpha] - \alpha$. Now map θ_{β} in $1 - \alpha - 1/2\gamma$, that is $\beta = f(\theta_{\beta})(1 - \alpha - 1/2\gamma)$. With $p \geq 0, o \geq 0, q \geq 0$, the same interval-partitioning reasoning is sequentially repeated.

The last proposition applies to the ARCH ($o = p = 0$) and GARCH models ($o = 0$) as a special case. Similarly, one can transform the possibly constrained trainable parameters of any postulated distribution for the iid $\{z_i\}$ innovations. For example, for a $GED(\lambda)$ distribution $\lambda = 2 + \theta_{\lambda}^2 + \epsilon$, where $\epsilon > 0$ is a pedestal to grant $\lambda > 2$ holds strictly.

Algorithm 2 Inverse transformation for the GJR(p, o, q)

$\omega = \exp(\theta_{\omega})$	α_i
$s = 1 - \epsilon$	else
for $i = 1, \dots, p$ do	$\gamma_i = 2s\gamma_i$
$\alpha_i = f(\theta_{\alpha_i})s$	end if
$s = 1 - s$	$s = 1 - 0.5s$
end for	end for
for $i = 1, \dots, o$ do	for $i = 1, \dots, q$ do
$\gamma_i = f(\theta_{\gamma_i})$	$\beta_i = f(\theta_{\beta_i})$
if $p \geq i$ then	$s = 1 - s$
$\gamma_i = (2s + \alpha_i)\gamma_i -$	end for

III. EXPERIMENTS

A. Data

For our empirical analyses, we use daily close-to-close log returns for the constituents of the S&P500 index. Our data covers 1383 trading days, from 1 January 2018 to 30 June 2023, divided into train and test sets with a 75%-25% split following chronological order. We use 488 stocks, since some constituents changed and their time series are incomplete.

B. Models, and optimization

To assess how satisfactory VI is in volatility modeling, we adopt the following volatility models: ARCH(1), GARCH(1,1), GJR-GARCH(1,1), EGARCH(0,1), EGARCH(1,1), GJR-EGARCH(1,1), FIGARCH(1,d,1). The case study on the Microsoft Corp. data additionally includes the GARCH(2,1), EGARCH(2,1), and the FIGARCH(0,d,1). The analyses adopt the BBVI, QBVI, MGVB, and EMGVB optimizers for VI (the first two under a diagonal variational covariance matrix). For the FIGARCH models, we implement (1) by the method of [23]. As a baseline for comparison, we adopt QML estimates and a Monte Carlo Markov Chain Sampler (MC). An MC reference for VI is advisable as it provides a benchmark for highlighting biases and assessing the quality of the Gaussian variational approximation.

For consistency, in VI, we adopt the same set of hyperparameters for all the experiments and optimizers. In particular, we use a learning rate of 0.005, 50 MC draws for approximating the expectation (5), a diagonal normal prior of unit variance and initial values $\mu_0 = 0$, $\Sigma_0 = 0.1I$. To increase the stability of the learning process, we update the gradients with a momentum factor of 0.4. Both MC and VI algorithms are run for a longer-than-required number of iterations. This avoids tuning the parameter on a case-by-case basis (which is unfeasible with hundreds of stocks) and provides reasonable guarantees that the algorithms converged. For VI, we observe that typically 1500 iterations are sufficient for the LB to reach a plateau (Fig. 1), yet we terminate the training after 2500 iterations. Opposed to ML, in statistics overfitting is a major concern in model selection. For example, in maximum likelihood estimation, the dynamic of the likelihood on a (usually nonexistent) test set is ignored. Early-stopping criteria for MC/VI based on the test loss would lead to non-comparability with the fully-in-sample-optimized maximum likelihood estimates. The relevant data and codes for the experiments are available at github.com/mmagris/GARCHVI.

C. Results

1) *General results:* To assess the effectiveness of VI as a Bayesian procedure, we adopt four performance metrics on the training and test sets. For VI, performances are computed as averages of 7,000 inversely-transformed samples from the estimated variational posterior $q_{\zeta^*}(\theta)$. For MC, the last 7,000 samples of the Markov chain. The metrics are the Negative Log-Likelihood (NLL), the Root Mean Squared Error (RMSE), the Mean Absolute Deviation (MAD), and the Qlik

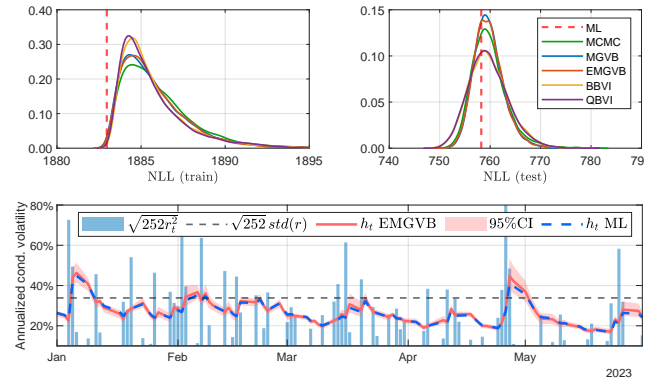


Fig. 2. Distribution of the train (top left) and test (top right) NLL, and confidence bounds for the conditional volatility (bottom). GJR-GARCH(1,1), Microsoft Inc. data.

loss [24]. For the last three, as proxies for the observed conditional variances, we adopt squared returns [24].

Table I presents the overall estimation results for the 488 stocks. For a performance metric M_E^x computed on a data subsample x as a result of an estimation procedure E , the entries in the Table correspond to mean performance deviations from the QML benchmark expressed as percentages, i.e., $100(M_E^x/M_{QML}^x - 1)$, and their standard deviations for the S&P 500 stocks. We comment on the predominance of positive signs indicating that QML is, overall, the preferred estimation approach from a merely quantitative perspective. However, VI methods can sometimes outperform QML in certain model/loss combinations. Clearly, test NLL values are always positive, confirming that QML estimates are optimal in this sense. At a general level, the typical magnitude of the ratios is in the sub-1% order, indicating that MC/VI estimation procedures are indeed effective w.r.t. QML and each other. Among the VI optimizers, we do not observe patterns indicating the dominance of one optimizer over another, implying they all determine a comparable variational approximation and, thus, performance. At the same time, the homogeneity in the results indicates that VI is robust with respect to the choice of the optimizer; all the optimizers appear adequate for the problems analyzed. By applying Chebyshev's inequality with, e.g., a margin of 3 standard deviations, the differences are broadly insignificant, indicating that the Gaussian variational approximation for the unconstrained parameters is plausible, at least for capturing the first two moments of the margins.

Regardless of the VI optimizer and volatility model, these results show that VI firmly stands as a valid estimation alternative to MC sampling. The in-sample and out-of-sample loss in performance w.r.t. QML is negligible, while the Bayesian framework enables several advantages, as for Sec. I.

2) *Case study:* We provide further analyses for Microsoft Corp. data. Table II summarizes the results regarding the training processes and the parameters' estimates.

The takeaway, justified blow, is that differences in performance metrics are broadly negligible, and the differences in the estimated parameters are minor. The impact of the

TABLE I

MAIN ESTIMATION RESULTS. AVERAGES AND RESPECTIVE STANDARD DEVIATIONS ACROSS THE S&P 500 STOCKS, OF THE DIFFERENCES BETWEEN THE INDIVIDUAL ESTIMATORS' PERFORMANCE WITH RESPECT TO THE QML PERFORMANCE, EXPRESSED AS A PERCENTAGE OF THE QML PERFORMANCE. PERFORMANCES ARE EVALUATED IN THE TRANSFORMED POSTERIOR MEAN (6).

	Train				Test			
	RMSE	MAD	QLIK	NLL	RMSE	MAD	QLIK	NLL
ARCH(1)								
MCMC	+0.257±35.820	+0.488±42.559	+0.105±1.306	+0.045±0.750	+0.422±40.208	+0.562±51.238	+0.128±18.353	+0.048±6.264
MGVB	+0.308±39.794	+0.555±41.888	+0.113±1.476	+0.048±0.745	+0.525±41.932	+0.663±52.409	+0.175±20.910	+0.066±7.305
EMGVB	+0.313±39.756	+0.562±41.596	+0.113±1.474	+0.048±0.745	+0.532±41.673	+0.670±52.077	+0.176±20.927	+0.066±7.304
BBVI	+0.313±39.755	+0.562±41.594	+0.113±1.474	+0.048±0.745	+0.532±41.671	+0.670±52.075	+0.176±20.927	+0.066±7.304
QBVI	+0.277±37.496	+0.532±40.190	+0.116±1.464	+0.049±0.744	+0.502±37.718	+0.639±50.450	+0.181±19.112	+0.067±7.027
GARCH(1,1)								
MCMC	-0.008±57.547	+0.033±126.751	+0.279±8.379	+0.111±3.371	+0.378±45.971	+0.320±117.589	+0.218±77.296	+0.080±30.211
MGVB	-0.019±56.563	+0.118±122.241	+0.265±7.860	+0.106±3.149	+0.394±44.033	+0.431±112.530	+0.221±73.356	+0.081±28.555
EMGVB	-0.016±56.838	+0.129±123.087	+0.268±7.888	+0.107±3.170	+0.399±44.328	+0.441±113.424	+0.223±73.864	+0.082±28.690
BBVI	-0.016±56.843	+0.129±123.097	+0.268±7.889	+0.107±3.171	+0.399±44.333	+0.441±113.433	+0.223±73.869	+0.082±28.692
QBVI	-0.098±40.651	+0.048±107.369	+0.225±5.025	+0.090±1.882	+0.270±32.590	+0.355±97.150	+0.244±47.097	+0.091±18.429
GJR-GARCH(1,1)								
MCMC	+0.216±136.296	+1.076±188.896	+0.498±19.800	+0.198±8.189	+1.023±110.649	+1.350±206.975	+0.207±155.432	+0.079±60.732
MGVB	+0.191±131.465	+0.889±186.580	+0.509±19.616	+0.202±8.006	+0.959±105.886	+1.140±204.363	+0.226±154.253	+0.084±59.804
EMGVB	+0.173±119.290	+0.903±189.253	+0.513±20.138	+0.204±8.280	+0.972±109.063	+1.151±207.438	+0.235±156.458	+0.088±60.738
BBVI	+0.173±119.297	+0.903±189.261	+0.513±20.140	+0.204±8.281	+0.972±109.076	+1.151±207.446	+0.235±156.471	+0.088±60.744
QBVI	-0.041±83.022	+0.727±154.161	+0.408±12.397	+0.162±4.942	+0.580±66.795	+0.998±163.206	+0.169±103.345	+0.066±40.164
EGARCH(0,1)								
MCMC	+0.591±100.969	+0.447±47.094	+0.099±0.889	+0.043±0.509	+0.277±35.295	+0.278±26.375	+0.079±4.951	+0.029±1.776
MGVB	+0.704±119.838	+0.418±49.920	+0.111±0.935	+0.048±0.540	+0.335±49.310	+0.234±31.721	+0.111±9.794	+0.041±3.327
EMGVB	+0.667±113.577	+0.400±50.086	+0.107±0.870	+0.046±0.524	+0.320±47.369	+0.216±30.581	+0.108±9.811	+0.040±3.336
BBVI	+0.667±113.588	+0.400±50.089	+0.107±0.870	+0.046±0.524	+0.320±47.373	+0.216±30.584	+0.108±9.811	+0.040±3.336
QBVI	+0.591±99.846	+0.338±46.964	+0.106±0.864	+0.046±0.518	+0.258±38.595	+0.163±29.894	+0.104±10.196	+0.039±3.497
EGARCH(1,1)								
MCMC	+0.174±28.495	+0.450±33.825	+0.199±1.921	+0.080±1.076	+0.244±28.261	+0.501±57.923	+0.216±33.674	+0.079±12.473
MGVB	+0.131±30.125	+0.533±31.338	+0.198±3.107	+0.079±1.237	+0.276±31.356	+0.609±54.202	+0.237±33.555	+0.085±12.823
EMGVB	+0.455±68.170	+0.708±83.257	+0.622±40.371	+0.240±13.503	+0.571±63.156	+0.809±169.018	+0.625±97.310	+0.224±35.607
BBVI	+0.456±68.614	+0.724±83.494	+0.623±40.489	+0.241±13.556	+0.577±63.868	+0.807±164.302	+0.626±97.604	+0.225±35.741
QBVI	+0.397±76.614	+0.998±99.027	+0.748±49.880	+0.288±16.616	+0.745±84.246	+1.375±178.473	+0.749±115.469	+0.261±39.564
GJR-EGARCH(1,1)								
MCMC	+0.165±38.646	+0.785±44.847	+0.259±2.923	+0.103±1.644	+0.406±47.171	+0.849±83.595	+0.259±41.236	+0.094±14.483
MGVB	+0.202±56.358	+0.819±54.540	+0.310±9.159	+0.127±4.050	+0.465±53.650	+0.929±112.795	+0.310±60.818	+0.114±21.561
EMGVB	+0.846±107.619	+1.214±120.063	+1.026±71.257	+0.394±24.872	+1.043±146.509	+1.124±272.713	+0.781±161.688	+0.276±55.370
BBVI	+0.847±108.432	+1.235±120.755	+1.028±71.166	+0.394±24.853	+1.047±146.676	+1.150±273.753	+0.783±162.340	+0.277±55.636
QBVI	+0.828±139.379	+1.831±153.957	+1.170±79.713	+0.449±27.667	+1.412±187.324	+2.177±324.510	+0.897±175.608	+0.326±64.718
FIGARCH(1,d,1)								
MCMC	+0.090±54.384	+0.845±136.805	+0.258±15.436	+0.103±6.538	+0.204±48.465	+1.007±205.486	-0.028±64.930	-0.018±25.460
MGVB	+0.146±59.358	+2.381±217.490	+0.149±21.112	+0.061±8.566	+0.475±67.696	+2.620±277.691	-0.587±101.008	-0.213±36.897
EMGVB	+0.172±68.563	+2.477±240.598	+0.153±21.398	+0.063±8.714	+0.514±76.643	+2.762±307.879	-0.569±105.048	-0.206±38.216
BBVI	+0.172±68.730	+2.478±240.930	+0.153±21.404	+0.063±8.717	+0.514±76.825	+2.764±308.238	-0.569±105.073	-0.206±38.222
QBVI	+0.145±58.516	+2.427±217.864	+0.168±20.448	+0.069±8.299	+0.493±67.799	+2.716±301.438	-0.577±99.190	-0.209±36.377

chosen estimation method is secondary, promoting VI as a solid alternative to MC and QML. Performance metrics are broadly overlapping, and differences are not statistically different except for the train NLL, consistently minimized by QML.

It is instructive to look at the performance metrics achieved by the different optimizers. For all the models, the MSFT findings align with the percentage reported in Table I. On both the training and test sets, the values of the performance metrics are remarkably aligned between QML and the Bayesian estimators, also for the additional models. Switching to a Bayesian framework does not harm with respect to QML performance on both sets. The estimated variance indicates broad non-significance in the difference across the Bayesian estimates, if not for NLL^{train} . We do not include the additional table with all the cross-testing results but rather visually discuss this case in Figure 2. The top-right panel of Figure 2 shows that the hypothesis $NLL_{QML}^{\text{train}} = NLL_{VI}^{\text{train}}$ cannot be rejected. Conversely, its rejection on the training set validates the above.

Extending the analysis to the value of the optimized LB, we observe that all the VI optimizers are rather equivalent for Bayesian inference, targeting a similar optimum. Clearly, the

diagonal BBVI and QBVI ones do not reach the same LB optimum that MGVB and EMGVB do (see. e.g., the GJR-GARCH(1,1) case in Table II and Figure 1), yet the differences are well within 1%, both on the training and test sets. The differences in the LB correspond to differences in the posterior estimates, explaining differences in the estimates posterior means of Table II. Yet, performance metrics are practically analogous; all the reported estimates can be considered equally effective, especially for out-of-sample forecasting.

In this regard, we observe a remarkable alignment between the MC and MGVB/EMGVB estimates and the standard deviations, suggesting that the full-covariance Gaussian specification appears feasible, at least for capturing the first two moments of the marginal distribution of the true posterior (approximated by the MC sampler). Figure 1 includes the posterior means for the GJR-GARCH(1,1) model in the constrained parameter space. The plot highlights the importance of allowing for a full-covariance specification, and the MGVB/EMGVB overlap to the MC density supports the Gaussian variational framework. Observing the QML estimates within the region of high density further validates the overall VI calibration with respect to QML.

distribution but rather to the robustness of the VI setup in this setting.

Small prior variance does affect the posterior estimation by keeping the posterior mean rather away from the QML estimates (see Table IV). This is certainly positive if one has motivated prior beliefs on the parameter in the unconstrained space (though unlikely). On the other hand, prior variances greater than one already deliver similar estimates (further aligned with QML). Encoding prior lack of knowledge appears to be relatively smooth; a prior variance τI with $\tau > 1$ is effective in this regard.

TABLE III

ESTIMATION RESULTS UNDER DIFFERENT PARAMETRIC FORMS OF THE IID INNOVATIONS. MICROSOFT INC. DATA, GJR-GARCH(1,1).

Error	Optimizer	ω	α	γ	β	ν	λ
Normal	ML	0.169	0.089	0.169	0.772		
	MCMC	0.218	0.111	0.182	0.731		
	EMGVB	0.218	0.111	0.182	0.730		
Student-t	ML	0.148	0.067	0.203	0.788	6.991	
	MCMC	0.236	0.106	0.226	0.720	6.417	
	EMGVB	0.233	0.106	0.225	0.721	6.394	
GED	ML	0.155	0.076	0.185	0.783	1.408	
	MCMC	0.221	0.109	0.198	0.724	1.462	
	EMGVB	0.223	0.109	0.199	0.723	1.458	
Skew-t	ML	0.151	0.061	0.205	0.790	7.402	-0.130
	MCMC	0.230	0.102	0.227	0.725	6.521	-0.120
	EMGVB	0.229	0.099	0.225	0.727	6.632	-0.124

TABLE IV

EFFECT OF DIFFERENT PRIOR VARIANCES. MICROSOFT INC. DATA, GJR-GARCH(1,1).

τ	Optimizer	ω	α	γ	β
0.01	MCMC	1.033	0.477	0.242	0.210
	EMGVB	1.034	0.475	0.244	0.210
0.1	MCMC	0.646	0.286	0.207	0.440
	EMGVB	0.650	0.285	0.212	0.439
1	MCMC	0.218	0.111	0.182	0.731
	EMGVB	0.218	0.111	0.182	0.730
5	MCMC	0.168	0.092	0.176	0.773
	EMGVB	0.173	0.092	0.176	0.769
10	MCMC	0.164	0.091	0.176	0.776
	EMGVB	0.166	0.089	0.175	0.775
20	MCMC	0.160	0.092	0.174	0.780
	EMGVB	0.163	0.088	0.175	0.778
	ML	0.169	0.089	0.169	0.772

IV. CONCLUSION

This paper documents the validity of Variational Inference (VI) as a tool for the Bayesian estimation for common volatility models of the GARCH family. We show that within a Gaussian variational framework, VI gradient-free black-box methods are robust and aligned with both the estimates obtained via Monte Carlo sampling and traditional Quasi Maximum Likelihood (QML). In this setting, we show how to adopt parameter transforms to enable VI principles and provide valuable insights on VI by the use of extensive performance statistics calculated from the individual time series of the S&P500 constituents. Along with a case study and different robustness analyses, we conclude that VI stands as a reliable, adequate, and suitable alternative to MC sampling and QML. The differences in training and test performance

metrics with respect to QML and MCMC are typically within the order of 1%. Despite our evidence on the validity of Gaussian variational margins, future research may investigate the appropriateness of the Gaussian copula and the use of alternative dependence structures. More in general, the VI framework could be applied to other domains, such as, e.g., stochastic volatility models or derivative pricing. We hope that our results will promote the deployment of VI in econometric and financial applications, encouraging the use of further toolsets and results from the Machine Learning research.

REFERENCES

- [1] David Ardia. *Financial Risk Management with Bayesian Estimation of GARCH Models Theory and Applications*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.
- [2] David Ardia. Bayesian estimation of a markov-switching threshold asymmetric garch model with student-t innovations. *The Econometrics Journal*, 12(1):105–126, 2009.
- [3] María Concepción Ausín and Pedro Galeano. Bayesian estimation of the gaussian mixture garch model. *Computational Statistics & Data Analysis*, 51(5):2636–2652, 2007.
- [4] Richard T. Baillie, Tim Bollerslev, and Hans Ole Mikkelsen. Fractionally integrated generalized autoregressive conditional heteroskedasticity. *Journal of econometrics*, 74(1):3–30, 1996.
- [5] Luc Bauwens and Michel Lubrano. Bayesian inference on garch models using the gibbs sampler. *The Econometrics Journal*, 1(1):C23–C46, 1998.
- [6] Tim Bollerslev, Ray Y. Chou, and Kenneth F Kroner. Arch modeling in finance: A review of the theory and empirical evidence. *Journal of econometrics*, 52(1-2): 5–59, 1992.
- [7] Klaus Böcker, editor. *Rethinking Risk Measurement and Reporting*, volume 2, chapter 1, pages 1–22. Risk Books, London, 2nd. edition, 2010.
- [8] Richard Gerlach and Cathy W.S. Chen. Bayesian inference and model comparison for asymmetric smooth transition heteroskedastic models. *Statistics and Computing*, 18:391–408, 2008.
- [9] John Geweke. Bayesian inference in econometric models using monte carlo integration. *Econometrica*, 57(6): 1317–1339, 1989.
- [10] Lawrence R. Glosten, Ravi Jagannathan, and David E Runkle. 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, 1993.
- [11] Peter Hall and Qiwei Yao. Inference in arch and garch models with heavy-tailed errors. *Econometrica*, 71(1): 285–317, 2003.
- [12] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes, 2013.
- [13] Frank Kleibergen and Herman K. Van Dijk. Non-stationarity in garch models: A bayesian analysis. *Journal of Applied Econometrics*, 8(S1):S41–S61, 1993.

- [14] Alp Kucukelbir, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David M Blei. Automatic differentiation variational inference. *Journal of machine learning research*, 18(14):1–45, 2017.
- [15] Martin Magris and Alexandros Iosifidis. Bayesian learning for neural networks: an algorithmic survey. *Artificial Intelligence Review*, 56(10):11773–11823, 2023.
- [16] Martin Magris, Mostafa Shabani, and Alexandros Iosifidis. Exact manifold gaussian variational bayes, 2022.
- [17] Martin Magris, Mostafa Shabani, and Alexandros Iosifidis. Quasi black-box variational inference with natural gradients for bayesian learning, 2022.
- [18] Peter Müller and Andy Pole. Monte carlo posterior integration in garch models. *Sankhyā: The Indian Journal of Statistics, Series B (1960-2002)*, 60:127–144, 1998.
- [19] Teruo Nakatsuma. A markov-chain sampling algorithm for garch models. *Studies in Nonlinear Dynamics & Econometrics*, 3(2), 1998.
- [20] Daniel B. Nelson. Conditional heteroskedasticity in asset returns: A new approach. *Econometrica*, 59(2):347–370, 1991.
- [21] Nghia Nguyen, Minh-Ngoc Tran, David Gunawan, and Robert Kohn. A statistical recurrent stochastic volatility model for stock markets. *Journal of Business & Economic Statistics*, 41:414–428, 2022.
- [22] Trong-Nghia Nguyen, Minh-Ngoc Tran, and Robert Kohn. Recurrent conditional heteroskedasticity. *Journal of Applied Econometrics*, 37(5):1031–1054, 2022.
- [23] Morten Ørregaard Nielsen and Antoine L. Noël. To infinity and beyond: Efficient computation of arch (∞) models. *Journal of Time Series Analysis*, 42(3):338–354, 2021.
- [24] Andrew J. Patton. Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, 160(1):246–256, 2011.
- [25] Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In Samuel Kaski and Jukka Corander, editors, *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33 of *Proceedings of Machine Learning Research*, pages 814–822, Reykjavik, Iceland, 22–25 Apr 2014. PMLR.
- [26] Timo Teräsvirta. An introduction to univariate garch models. In *Handbook of financial time series*, pages 17–42. Springer, Berlin, Heidelberg, 2009.
- [27] Minh-Ngoc Tran, Dang H Nguyen, and Duy Nguyen. Variational bayes on manifolds. *Statistics and Computing*, 31:1–17, 2021.
- [28] Audrone Virbickaite, M. Concepción Ausín, and Pedro Galeano. Bayesian inference methods for univariate and multivariate garch models: A survey. *Journal of Economic Surveys*, 29(1):76–96, 2015.