Maximum Likelihood Estimation of Gaussian and Student's t GARCH: A Unified Penalty Method

Chenyu Gao, *Graduate Student Member, IEEE*, Ziping Zhao, *Member, IEEE*, and Daniel P. Palomar, *Fellow, IEEE*

***Abstract*—The generalized autoregressive conditional heteroskedasticity (GARCH) model is widely used to characterize time-varying conditional volatility in time series analysis. This paper studies the maximum likelihood (ML) estimation of GARCH model parameters under the assumption that the conditional distribution of the innovations follows either a Gaussian or a Student's t distribution. The estimation problems are challenging due to the non-convex and recursive coupling nature of the model parameters, which often leads to convergence issues in existing ML estimation methods. Moreover, many existing approaches fail to explicitly enforce the stationarity constraint, a desirable property for GARCH models. To address these challenges, we propose a unified penalty method for ML estimation of GARCH models that effectively handles the parameter coupling and allows flexible incorporation of stationarity constraints. We develop a convergent estimation algorithm based on the block majorization-minimization (BMM) framework, which efficiently exploits the problem structure to update the model parameters. We prove that the sequence generated by the BMM algorithm converges to the set of Karush-Kuhn-Tucker points. Notably, for the Student's t case, we provide the first theoretical characterization of the convexity of the negative conditional log-likelihood function with respect to the shape parameter. Furthermore, our algorithm naturally extends to M-estimation of GARCH, estimation of GARCH variants, and joint estimation of GARCH and conditional mean models. Numerical experiments on synthetic data demonstrate that the proposed algorithm outperforms existing methods in terms of parameter estimation accuracy and objective value. Its effectiveness is further validated on real-world datasets from applications including financial volatility modeling and radar target detection.**

***Index Terms*—Time series analysis, GARCH, conditional volatility, parameter estimation, Student's t, non-convex optimization, majorization-minimization.**

I. Introduction

In time series analysis, a time series is often viewed as a collection of random variables ordered in time or as manifestations of a random process [2], [3]. Traditional models often assume that the conditional volatility, that is, the conditional standard deviation of a time series, remains constant over time. However, empirical evidence from various real-world applications contradicts this assumption, revealing substantial variability. This discrepancy highlights the necessity of modeling conditional volatility as non-constant. Engle [4] was the first to highlight the non-constant nature of conditional volatility in economic time series, leading him to propose the autoregressive conditional heteroskedasticity (ARCH) model. This model describes the volatility dynamics of a time series based on its past residuals, defined as deviations from the conditional mean. Subsequently, the generalized autoregressive conditional heteroskedasticity (GARCH) model was introduced by Bollerslev [5]. The GARCH model extends the ARCH model by modeling the conditional volatility as serially correlated, using both past residuals and past volatility through an autoregressive moving average structure to forecast future volatility. As a general framework, GARCH encompasses both the constant volatility model and the ARCH model as special cases. Over time, numerous variants of the GARCH model have been developed [6], [7], [8], [9], [10]. For comprehensive reviews of GARCH models, readers may refer to [11], [12], [13].

The GARCH model was originally developed in econometrics to analyze economic time series such as inflation [4], [5] and financial time series such as asset returns [14]. It is particularly effective in capturing volatility clustering and accommodating heavy-tailed distributions [15]. In Fig. 1, we plot the daily log-return residuals of HSBC Holdings plc (HSBC) from the New York Stock Exchange, covering the period from January 1st, 2022 to December 31st, 2024, along with the corresponding volatility envelopes estimated using the Student's t GARCH and integrated GARCH [6] models (the data to the left of the vertical dotted line are used for model estimation). We can observe the volatility clustering phenomenon from the log-return residuals, where intervals of high volatility tend to cluster together, and low-volatility intervals also

Received 12 March 2025; revised 24 July 2025; accepted 17 August 2025. Date of publication 1 September 2025; date of current version 16 December 2025. This work was supported in part by the National Nature Science Foundation of China (NSFC) under Grant 62001295 and in part by the Hong Kong Research Grants Council (RGC) under Grant 16206123. An earlier version of this paper was presented at the 22nd IEEE Statistical Signal Processing Workshop (SSP), Hanoi, Vietnam, 2–5 July 2023 [DOI: 10.1109/SSP53291.2023.10208065]. The associate editor coordinating the review of this article and approving it for publication was Herwig Wendt. *(Corresponding author: Ziping Zhao.)*

Chenyu Gao and Ziping Zhao are with the School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China (e-mail: gaochy@shanghaitech.edu.cn; zipingzhao@shanghaitech.edu.cn).

Daniel P. Palomar is with the Department of Electronic and Computer Engineering and the Department of Industrial Engineering and Decision Analytics, Hong Kong University of Science and Technology, Kowloon, Hong Kong (e-mail: palomar@ust.hk).

This article has supplementary downloadable material available at https://doi.org/10.1109/TSP.2025.3604632, provided by the authors.

Digital Object Identifier 10.1109/TSP.2025.3604632

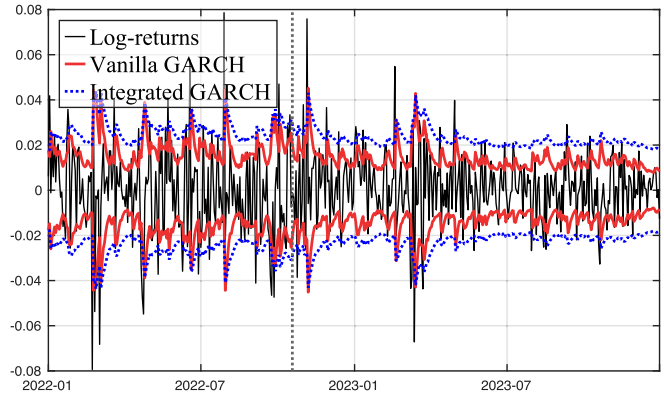


Fig. 1. The log-return residuals of HSBC and the estimated GARCH models.

tend to group. The estimated GARCH and integrated GARCH models effectively capture this phenomenon. Beyond econometrics, GARCH has found extensive applications in various domains, including signal processing and machine learning. For example, in speech signal processing, GARCH is commonly employed for speech enhancement, as speech signals in the short-time Fourier transform domain exhibit both volatility clustering and heavy-tailed behavior [16], [17], [18], [19], [20], [21]. In array signal processing, radar clutter can be described using the GARCH process because of its volatility clustering characteristics [22], [23]. Additionally, GARCH models have been applied in several other areas, such as anomaly detection [24], [25], [26], covariance recovery [27], biosignal processing [28], image processing [29], [30], graph signal processing [31], electricity price forecasting [32], [33], wind power prediction [34], sentiment analysis [35], and traffic prediction [36], [37], [38].

Despite its strong modeling capacity and wide range of applications, estimating the parameters of a GARCH model is a highly challenging task. Methods such as the generalized method of moments [39] and ordinary least squares estimation [40] have been proposed. While these methods are easy to implement, they are not asymptotically efficient and perform poorly in small samples [41]. Maximum likelihood (ML) estimation remains the most widely used approach for GARCH models and can be performed under various assumptions on the conditional distribution of the innovations. The Gaussian distribution assumption was first considered for ML estimation in [5]. However, in fields such as finance, the observed residuals often exhibit heavier tails than those that fit a Gaussian GARCH model [42], [44]. Consequently, the Student's t distribution is often regarded as a more suitable alternative, as it offers greater flexibility in modeling tail behavior [14], [45], [46]. However, solving ML estimation problems for GARCH parameters is inherently challenging. First, the estimation must respect certain constraints. The positivity constraint ensures that the estimated variance remains nonnegative, which is essential for the validity of the volatility model. The stationarity constraint is typically imposed to ensure that the volatility process remains stationary over time, a desirable property in GARCH modeling [4], [5], [6]. In addition, the volatility dynamics are governed by recursive equations that couple all model parameters, which makes the optimization problem highly non-convex and difficult to solve.

In the literature, various algorithms have been proposed for the ML estimation of GARCH model parameters. Many of these methods are implemented in publicly available software packages, including the R packages tseries [47], fGarch [48], and rugarch [49], the Python package arch [50], as well as the commercial MATLAB toolbox Econometrics [51]. A quasi-Newton method called Berndt-Hall-Hall-Hausman (BHHH) was introduced in the seminal works on Gaussian GARCH [5] and Student's t GARCH [14]. Other quasi-Newton methods, such as the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm and the limited-memory BFGS (L-BFGS) algorithm, are incorporated in software packages including the R packages tseries, fGarch, and rugarch, with BFGS used in tseries and L-BFGS implemented in fGarch and rugarch. However, these quasi-Newton methods do not enforce the stationarity constraint during estimation [52], [53], [54]. In addition to quasi-Newton methods, sequential quadratic programming (SQP) and interior-point method (IPM) have also been proposed to address the constrained ML estimation problem. The SQP method is implemented in the R package rugarch, the Python package arch, and the MATLAB toolbox Econometrics, while the IPM method is available only in Econometrics. Although SQP and IPM explicitly incorporate the stationarity constraints, they do not guarantee that the iterates remain feasible at every iteration. To handle the volatility equations, existing algorithms typically substitute them into the objective function. However, the calculation of the gradient of the conditional log-likelihood function remains computationally intensive due to the recursive coupling of the model parameters. This computational burden increases the risk of round-off errors and may result in convergence failures for many algorithms.

Recognizing these limitations, we propose a novel, efficient, and provably convergent algorithm based on the block majorization-minimization (BMM) framework [55] to solve the ML estimation problems, under the assumption that the conditional distribution of innovations follows either a Gaussian or a Student's t distribution. It is important to emphasize that the contribution of this work does not lie in introducing a new GARCH model, but rather in developing efficient algorithms for parameter estimation of existing GARCH models. Unlike existing general-purpose algorithms, the proposed BMM algorithm offers several key advantages. First, to address the recursively coupled volatility equations inherent in GARCH models, we introduce a unified penalty method for the ML estimation problems. This method enables the volatility equations to be handled in a decoupled manner, which significantly reduces the computational complexity at each iteration. Second, the BMM algorithm explicitly incorporates the stationarity constraint while ensuring that all iterates remain within the feasible set. This property, often lacking in existing methods, is critical for GARCH estimation as it guarantees the stationarity of the

model. Third, the algorithm leverages the block structure of the variables to simplify subproblem updates, which often admit closed-form solutions. In addition, the algorithm guarantees convergence to a stationary point.

The contributions of this paper are summarized as follows:

- We study the ML estimation problems for GARCH models of arbitrary order, considering both Gaussian and Student's t conditional distributions of the innovations. We introduce a unified penalty method for these estimation problems, which effectively handles the recursively coupled volatility equations while explicitly incorporating the stationarity constraint.
- We propose a BMM algorithm that enables efficient variable updates, often with closed-form solutions, while ensuring that all iterates remain within the feasible set.
- For the Student's t ML estimation problem, we provide the first known theoretical characterization of the convexity of the negative conditional log-likelihood function with respect to the shape parameter.
- We show that the proposed penalty method naturally extends to M-estimation of GARCH, estimation of GARCH model variants, and joint estimation of GARCH models with conditional mean models.
- Numerical experiments on synthetic data show that the proposed algorithm outperforms state-of-the-art methods in GARCH estimation. Its effectiveness is further confirmed through evaluations on real-world datasets.

The remainder of this paper is organized as follows: Section II introduces the GARCH model. In Section III, we provide a brief overview of the BMM algorithmic framework. Section IV discusses the Gaussian ML estimation problem for GARCH and the corresponding development of the BMM algorithm. Section V focuses on the ML estimation of Student's t GARCH. The convergence and complexity of the proposed BMM algorithms are analyzed in Section VI. Section VII discusses how the penalty method can be extended to other GARCH-related estimation problems. The BMM algorithms for GARCH estimation are compared numerically with existing methods on both synthetic and real datasets in Section VIII. Finally, we conclude in Section IX.

Notation. The following notation is adopted. Standard lowercase letters denote scalars. Boldface lowercase and uppercase letters represent vectors and matrices, respectively. Uppercase calligraphic letters stand for sets. $\mathbf{0}$ and $\mathbf{1}$ denote column vectors where all entries are zero and one, respectively. $\boldsymbol{I}$ stands for the identity matrix. $\mathrm{E}(x)$ and $\mathrm{Var}(x)$ denote the mean and variance of the random variable x. $|x|$ is the modulus of the scalar x. $[\boldsymbol{x}]_i$ denotes the i-th entry of the vector $\boldsymbol{x}$. $\|\boldsymbol{x}\|$ denotes the 2-norm of the vector $\boldsymbol{x}$. $\boldsymbol{X}^\top$ denotes the transpose of $\boldsymbol{X}$. $\lambda_{\max}(\boldsymbol{X})$ is the maximum eigenvalue of $\boldsymbol{X}$. For symmetric matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$, $\boldsymbol{X} \succeq \boldsymbol{Y}$ denotes that $\boldsymbol{X} - \boldsymbol{Y}$ is symmetric positive semidefinite. $\mathrm{Card}(\mathcal{I})$ denotes the size (cardinality) of the set $\mathcal{I}$. The notation $O(\cdot)$ describes the asymptotic upper bound of the time complexity with respect to the input size. Given a function f, we use f' and f'' to represent its first- and second-order derivatives, respectively.

II. The GARCH Model

Consider a random process $\{y_t\}$. The residual process $\{\epsilon_t\}$ is defined as

$$\epsilon_t = y_t - \mu_t,$$

where $\mu_t = \mathrm{E}(y_t \mid \mathcal{F}_{t-1})$ is the conditional mean of y_t with $\mathcal{F}_{t-1}$ denoting the information set available up to time $t-1$. The residual ϵ_t is said to follow a GARCH process [5] if its conditional variance, defined as $h_t = \mathrm{Var}(\epsilon_t \mid \mathcal{F}_{t-1})$, satisfies the following autoregressive moving average equation:

$$h_t = \omega + \sum_{i=1}^{q} \alpha_i \epsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j h_{t-j}, \tag{1}$$

where $\omega, \alpha_1, \ldots, \alpha_q$, and $\beta_1, \ldots, \beta_p$ are the model parameters. Based on (1), the residual series $\{\epsilon_t\}$ can be represented as

$$\epsilon_t = \sqrt{h_t} z_t, \tag{2}$$

where $\{z_t\}$ is a sequence of independent and identically distributed (i.i.d.) random variables with zero mean and unit variance. The model in (1) is referred to as GARCH(q, p) with q and p specifying the order of the model (often both set to one in practice). Denote $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_q]^\top$ and $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_p]^\top$. When $\boldsymbol{\beta} = \mathbf{0}$, the GARCH model reduces to the ARCH model [4]; when $\boldsymbol{\alpha} = \boldsymbol{\beta} = \mathbf{0}$, the GARCH model simplifies to the constant variance model. Define

$$\boldsymbol{\gamma} = \left[\boldsymbol{\alpha}^\top, \boldsymbol{\beta}^\top\right]^\top,$$

and

$$\boldsymbol{c}_t = \left[\epsilon_{t-1}^2, \ldots, \epsilon_{t-q}^2, h_{t-1}, \ldots, h_{t-p}\right]^\top. \tag{3}$$

The equation (1) can be rewritten as

$$h_t = \omega + \boldsymbol{\gamma}^\top \boldsymbol{c}_t. \tag{4}$$

To ensure the validity of the model, we require the following regularity conditions:

$$\omega > 0, \quad \boldsymbol{\gamma} \geq \mathbf{0}, \tag{5}$$

which are sufficient to guarantee the positivity of $\{h_t\}$.

In GARCH modeling, the stationarity of the process $\{\epsilon_t\}$ is a central concern for both theory and practice. In the seminal work on GARCH [5], the following stationarity constraint is considered

$$\boldsymbol{\gamma}^\top \mathbf{1} < 1.$$

This condition guarantees that the process $\{\epsilon_t\}$ is strictly stationary and has a finite unconditional variance, which corresponds to wide-sense stationarity. In certain scenarios, such as financial time series, empirical evidence suggests that volatility shocks tend to exhibit long-lasting effects [56], [57]. To model such persistent behavior and generate long memory in the variance process, the following constraint is imposed

$$\boldsymbol{\gamma}^\top \mathbf{1} = 1.$$

In this case, although the process $\{\epsilon_t\}$ remains strictly stationary, its unconditional variance becomes infinite. Due to this

unit-root-like behavior in the variance dynamics, the resulting model is referred to as the integrated GARCH model [6]. The stationarity constraint is usually imposed during estimation to ensure a stationary GARCH model, though some studies [54], [58], [59] estimate parameters without it and check stationarity afterward. In this paper, we consider both settings and denote the constraint set on $\boldsymbol{\gamma}$ uniformly by $\mathcal{S}$.

III. Block Majorization-Minimization Algorithm

In this section, we introduce a generic iterative optimization framework known as block majorization-minimization (BMM) [55]. Consider the following optimization problem:

$$\min_{\boldsymbol{x}\in\mathcal{X}} \quad f(\boldsymbol{x}),$$

where the optimization variable $\boldsymbol{x}$ is partitioned into n blocks, denoted by $\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\}$, with each $\boldsymbol{x}_i\in\mathcal{X}_i$. We assume that the feasible set is a Cartesian product $\mathcal{X}=\prod_{i=1}^n\mathcal{X}_i$, where each $\mathcal{X}_i$ is closed and convex, and that the objective function $f:\mathcal{X}\to\mathbb{R}$ is continuous. In BMM, each variable block $\boldsymbol{x}_i$ is updated in a cyclic order by solving the following subproblem[1]:

$$\min_{\boldsymbol{x}_i\in\mathcal{X}_i} \quad \bar{f}_i(\boldsymbol{x}_i;\underline{\boldsymbol{x}}),$$

where $\bar{f}_i$ is an upper-bound surrogate function of f with respect to $\boldsymbol{x}_i$, and satisfies the following conditions:

$$\begin{aligned}&\bar{f}_i(\underline{\boldsymbol{x}}_i;\underline{\boldsymbol{x}})=f(\underline{\boldsymbol{x}}), \quad \forall\underline{\boldsymbol{x}}\in\mathcal{X},\ i=1,\ldots,n,\\ &\bar{f}_i(\boldsymbol{x}_i;\underline{\boldsymbol{x}})\geq f\left(\underline{\boldsymbol{x}}_1,\ldots,\underline{\boldsymbol{x}}_{i-1},\boldsymbol{x}_i,\underline{\boldsymbol{x}}_{i+1},\ldots,\underline{\boldsymbol{x}}_n\right),\\ &\qquad\qquad \forall\boldsymbol{x}_i\in\mathcal{X}_i,\ \forall\underline{\boldsymbol{x}}\in\mathcal{X},\ i=1,\ldots,n.\end{aligned}$$

In practice, the surrogate function $\bar{f}_i$ should approximate f closely to promote fast convergence, while remaining simple enough to ensure low per-iteration computational complexity for updating $\boldsymbol{x}_i$. The BMM algorithm proceeds iteratively until a convergence criterion is satisfied.

IV. ML Estimation of Gaussian GARCH

A. The Gaussian ML Estimation Problem

This section considers the Gaussian GARCH model, where the conditional distribution of z_t in (2) is standard Gaussian. The standard Gaussian distribution is defined as follows:

$$d_{\text{Gau}}(z_t)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{z_t^2}{2}\right).$$

Using the transformation $z_t=\frac{\epsilon_t}{\sqrt{h_t}}\,|\,\mathcal{F}_{t-1}$, where z_t is a conditional random variable, h_t is defined in (4), and $\mathcal{F}_{t-1}=\{\epsilon_{1-q},\ldots,\epsilon_{t-1},h_{1-p},\ldots,h_0\}$, we obtain the negative conditional log-likelihood function of ϵ_t, for $t=1,\ldots,n$, as follows:

$$\begin{aligned}&-\log L_{\text{Gau}}(\omega,\boldsymbol{\gamma};\epsilon_1,\ldots,\epsilon_n\,|\,\mathcal{F}_0)\\ &=-\log\prod_{t=1}^n\frac{1}{\sqrt{h_t}}d_{\text{Gau}}\left(\frac{\epsilon_t}{\sqrt{h_t}}\,|\,\mathcal{F}_{t-1}\right)\\ &=\sum_{t=1}^n\left[\frac{1}{2}\log h_t-\log d_{\text{Gau}}\left(\frac{\epsilon_t}{\sqrt{h_t}}\,|\,\mathcal{F}_{t-1}\right)\right].\end{aligned}$$

Then, we derive the ML estimation problem for the Gaussian GARCH model as follows:

$$\begin{aligned}\min_{\omega,\boldsymbol{\gamma},\boldsymbol{h}}\quad &\sum_{t=1}^n\left[\frac{1}{2}\log(2\pi)+\frac{1}{2}\log h_t+\frac{1}{2}\frac{\epsilon_t^2}{h_t}\right]\\ \text{s.t.}\quad &h_t=\omega+\boldsymbol{\gamma}^\top\boldsymbol{c}_t,\ t=1,\ldots,n\\ &\omega\geq 0_\varepsilon,\ \boldsymbol{\gamma}\geq\mathbf{0},\ \boldsymbol{\gamma}\in\mathcal{S},\end{aligned}\tag{6}$$

where

$$\boldsymbol{h}=[h_1,\ldots,h_n]^\top,$$

and $\boldsymbol{c}_t$ is defined in (3). All values of ϵ_t for $t=1-q,\ldots,0,\ldots,n$ and h_t for $t=1-p,\ldots,0$ are assumed to be known. To handle the strict inequality constraint in (5), namely $\omega>0$, we relax it to a closed constraint of the form $\omega\geq 0_\varepsilon$, where $0_\varepsilon=0+\varepsilon$ with ε being a small positive constant (e.g., 10^{-6}). This relaxation is commonly adopted by existing GARCH solvers, including the R packages tseries [47], fGarch [48], and rugarch [49], the Python package arch [50], as well as the Econometrics Toolbox [51] in MATLAB, to facilitate the design and implementation of numerical algorithms. For the stationarity constraint, when $\boldsymbol{\gamma}^\top\mathbf{1}<1$ is considered, it is similarly relaxed to $\boldsymbol{\gamma}^\top\mathbf{1}\leq 1_\varepsilon$, where $1_\varepsilon=1-\varepsilon$.

Problem (6) is a non-convex constrained optimization problem. A significant challenge of the problem arises from the fact that the optimization variables are recursively coupled in the equality constraints. A common approach in existing numerical methods is to eliminate these constraints by recursively substituting them into the objective function. However, this substitution significantly increases computational complexity, as it requires computing gradients via the chain rule, which not only incurs a high computational cost but also amplifies numerical errors and may ultimately lead to convergence failure. To overcome this limitation, we reformulate the problem by incorporating the equality constraints using a penalty function approach. Based on this reformulation, we develop an efficient algorithm under the BMM framework.

B. Problem Reformulation via Penalization

Applying the quadratic penalty function to the equality constraints, we obtain the following estimation problem

$$\begin{aligned}\min_{\omega,\boldsymbol{\gamma},\boldsymbol{h}}\quad &\sum_{t=1}^n\left[\log h_t+\frac{\epsilon_t^2}{h_t}+\frac{\eta}{2}\left(h_t-\omega-\boldsymbol{\gamma}^\top\boldsymbol{c}_t\right)^2\right]\\ \text{s.t.}\quad &\omega\geq 0_\varepsilon,\ \boldsymbol{\gamma}\geq\mathbf{0},\ \boldsymbol{\gamma}\in\mathcal{S},\end{aligned}\tag{7}$$

where $\eta>0$ is a predefined penalty parameter. In contrast to existing approaches that handle equality constraints by recursive substitution into the objective, the proposed penalized formulation enables a more efficient algorithm by leveraging the specific structure of each variable block, an advantage naturally supported by the BMM framework.

[1]In this paper, underlined letters indicate variables that are held fixed during the update.

C. Problem Solving

Based on the idea of BMM, we divide the optimization variables in (7) into three blocks, namely ω, $\boldsymbol{\gamma}$, and $\boldsymbol{h}$. By carefully constructing surrogate functions, we derive closed-form update rules for each variable block.

1) Solving the ω-Block: Only the last term in the objective of (7) involves variable ω. Thus, the subproblem for ω is

$$\begin{aligned} \min_{\omega} \quad & \sum_{t=1}^{n} \left(\underline{h}_t - \omega - \underline{\boldsymbol{\gamma}}^\top \underline{\boldsymbol{c}}_t\right)^2 \\ \text{s.t.} \quad & \omega \geq 0_\varepsilon. \end{aligned} \tag{8}$$

Problem (8) is a convex quadratic program. It has a closed-form solution given by

$$\omega^+ = \max\left\{\frac{1}{n}\sum_{t=1}^{n}\left(\underline{h}_t - \underline{\boldsymbol{\gamma}}^\top \underline{\boldsymbol{c}}_t\right), 0_\varepsilon\right\}. \tag{9}$$

2) Solving the $\boldsymbol{\gamma}$-Block: The subproblem regarding $\boldsymbol{\gamma}$ is

$$\begin{aligned} \min_{\boldsymbol{\gamma}} \quad & \sum_{t=1}^{n} \left(\underline{h}_t - \underline{\omega} - \boldsymbol{\gamma}^\top \underline{\boldsymbol{c}}_t\right)^2 \\ \text{s.t.} \quad & \boldsymbol{\gamma} \geq \mathbf{0},\ \boldsymbol{\gamma} \in \mathcal{S}. \end{aligned} \tag{10}$$

Problem (10) is a convex quadratic program; however, it generally lacks a closed-form solution and must be solved numerically. Although standard solvers are applicable, we aim to derive a more efficient update by constructing appropriate surrogate functions, guided by the BMM framework. To this end, we first present a useful result.

Lemma 1 ([55]): Given $u\boldsymbol{I} \succeq \boldsymbol{D}$, it follows that

$$\boldsymbol{x}^\top \boldsymbol{D}\boldsymbol{x} \leq u\boldsymbol{x}^\top \boldsymbol{x} + 2\boldsymbol{x}^\top (\boldsymbol{D} - u\boldsymbol{I})\,\underline{\boldsymbol{x}} + \underline{\boldsymbol{x}}^\top (u\boldsymbol{I} - \boldsymbol{D})\,\underline{\boldsymbol{x}},$$

where the equality is attained when $\boldsymbol{x} = \underline{\boldsymbol{x}}$.

Based on Lemma 1, for the objective of (10), we have

$$\begin{aligned} & \sum_{t=1}^{n} \left(\underline{h}_t - \underline{\omega} - \boldsymbol{\gamma}^\top \underline{\boldsymbol{c}}_t\right)^2 \\ &= \boldsymbol{\gamma}^\top \left(\sum_{t=1}^{n} \underline{\boldsymbol{c}}_t \underline{\boldsymbol{c}}_t^\top\right)\boldsymbol{\gamma} - 2\sum_{t=1}^{n}\left(\underline{h}_t - \underline{\omega}\right)\boldsymbol{\gamma}^\top \underline{\boldsymbol{c}}_t + \text{const.} \\ &\leq u\boldsymbol{\gamma}^\top \boldsymbol{\gamma} - 2\boldsymbol{\gamma}^\top \boldsymbol{v} + \text{const.}, \end{aligned} \tag{11}$$

where

$$\begin{aligned} u &\geq \lambda_{\max}\left(\sum_{t=1}^{n} \underline{\boldsymbol{c}}_t \underline{\boldsymbol{c}}_t^\top\right), \\ \boldsymbol{v} &= u\underline{\boldsymbol{\gamma}} - \sum_{t=1}^{n}\left(\underline{\boldsymbol{\gamma}}^\top \underline{\boldsymbol{c}}_t - \underline{h}_t + \underline{\omega}\right)\underline{\boldsymbol{c}}_t, \end{aligned}$$

and const. represents a constant term that is irrelevant to $\boldsymbol{\gamma}$. Taking the last line in (11) as the surrogate function, the subproblem for $\boldsymbol{\gamma}$ is given by

$$\begin{aligned} \min_{\boldsymbol{\gamma}} \quad & u\boldsymbol{\gamma}^\top \boldsymbol{\gamma} - 2\boldsymbol{\gamma}^\top \boldsymbol{v} \\ \text{s.t.} \quad & \boldsymbol{\gamma} \geq \mathbf{0},\ \boldsymbol{\gamma} \in \mathcal{S}. \end{aligned} \tag{12}$$

We solve this problem case by case.

a) $\mathcal{S}$ is the $(q+p)$-dimensional Euclidean space

In this case, the solution for $\boldsymbol{\gamma}$ is given in closed form as

$$\gamma_i^+ = \max\left\{0, \frac{v_i}{u}\right\},\ \text{for } i = 1, \ldots, q+p. \tag{13}$$

b) $\mathcal{S} = \{\boldsymbol{\gamma} \mid \boldsymbol{\gamma}^\top \mathbf{1} \leq 1_\varepsilon\}$ and $\mathcal{S} = \{\boldsymbol{\gamma} \mid \boldsymbol{\gamma}^\top \mathbf{1} = 1\}$

In these cases, the solutions to problem (12) can still be computed in a finite number of steps, since the constraint set defines a simplex-type projection.

Proposition 2: Define an index set as follows:

$$\mathcal{V} = \{i \mid v_i > 0,\ \ i = 1, \ldots, q+p\}.$$

The solution to problem (12) with $\mathcal{S} = \{\boldsymbol{\gamma} \mid \boldsymbol{\gamma}^\top \mathbf{1} \leq 1_\varepsilon\}$ is given by

$$\gamma_i^+ = \begin{cases} \frac{v_i}{u}, & \sum\limits_{j\in\mathcal{V}} \frac{v_j}{u} < 1_\varepsilon, i \in \mathcal{V} \\ 0, & \sum\limits_{j\in\mathcal{V}} \frac{v_j}{u} < 1_\varepsilon, i \notin \mathcal{V} \\ 0, & \sum\limits_{j\in\mathcal{V}} \frac{v_j}{u} \geq 1_\varepsilon, i \in \mathcal{I}_\varepsilon \\ \frac{v_i}{u} - \frac{\sum_{j\notin\mathcal{I}} \frac{v_j}{u} - 1_\varepsilon}{q+p-\text{Card}(\mathcal{I}_\varepsilon)}, & \sum\limits_{j\in\mathcal{V}} \frac{v_j}{u} \geq 1_\varepsilon, i \notin \mathcal{I}_\varepsilon, \end{cases}$$

for $i = 1, \ldots, q+p$, where

$$\mathcal{I}_\varepsilon = \left\{ i \,\middle|\, \frac{v_i}{u} + \frac{\sum_{j\notin\mathcal{I}} \frac{v_j}{u} - 1_\varepsilon}{q+p-\text{Card}\left(\mathcal{I}_\varepsilon\right)} \geq 0,\ \ i = 1, \ldots, q+p \right\}.$$

An efficient computational procedure is given in Algorithm 1.

The solution to problem (12) with $\mathcal{S} = \{\boldsymbol{\gamma} \mid \boldsymbol{\gamma}^\top \mathbf{1} = 1\}$ is given by

$$\gamma_i^+ = \begin{cases} 0, & i \in \mathcal{I} \\ \frac{v_i}{u} - \frac{\sum_{j\notin\mathcal{I}} \frac{v_j}{u} - 1}{q+p-\text{Card}(\mathcal{I})}, & i \notin \mathcal{I}, \end{cases}$$

for $i = 1, \ldots, q+p$, where

$$\mathcal{I} = \left\{ i \,\middle|\, \frac{v_i}{u} + \frac{\sum_{j\notin\mathcal{I}} \frac{v_j}{u} - 1}{q+p-\text{Card}\left(\mathcal{I}\right)} \geq 0,\ \ i = 1, \ldots, q+p \right\}.$$

An efficient computational procedure is given in Algorithm 2.

Proof: It can be proved similarly to [[60], Lemma 1]. ■

3) Solving the $\boldsymbol{h}$-Block: Let

$$o_t = \underline{\omega} + \sum_{i=1}^{q} \underline{\alpha}_i \epsilon_{t-i}^2, \quad t = 1, \ldots, n.$$

The subproblem for $\boldsymbol{h}$ is

$$\min_{\boldsymbol{h}} \sum_{t=1}^{n} \left[\log h_t + \frac{\epsilon_t^2}{h_t} + \frac{\eta}{2}\left(h_t - \sum_{j=1}^{p} \underline{\beta}_j h_{t-j} - o_t\right)^2\right], \tag{14}$$

which is an unconstrained non-convex optimization problem.

Let the objective in (14) be denoted by $\varphi_{\text{Gau}}(\boldsymbol{h})$. The optimal solution $\boldsymbol{h}^+$ is the first-order stationary point of φ_{Gau}. Therefore, we solve for the zeros of the partial derivatives of $\varphi_{\text{Gau}}(\boldsymbol{h})$ with respect to h_t for $t = 1, \ldots, n$, respectively, which leads to a system of cubic equations. In this system, the optimization variables h_t, for $t = 1, \ldots, n$, are interdependent across the n

Algorithm 1 Solving problem (12) with $\mathcal{S}=\{\boldsymbol{\gamma}\,|\,\boldsymbol{\gamma}^\top\mathbf{1}\le 1_\varepsilon\}$

if $\sum_{i\in\mathcal{V}}\frac{v_i}{u}<1_\varepsilon$ **then**
  $\gamma_i=\frac{v_i}{u}$ with $i\in\mathcal{V}$ and $\gamma_i=0$ with $i\notin\mathcal{V}$
else
  build a set $\mathcal{I}_\varepsilon$, let $i\notin\mathcal{V}$ be in the set $\mathcal{I}_\varepsilon$
  start a loop
  find the max element v_i with $i\notin\mathcal{I}_\varepsilon$
  put i into the set $\mathcal{I}_\varepsilon$
  if $\frac{v_i}{u}<\frac{\sum_{j\notin\mathcal{I}}\frac{v_j}{u}-1_\varepsilon}{q+p-\mathrm{Card}(\mathcal{I}_\varepsilon)}$ **then**
    remove i from the set $\mathcal{I}_\varepsilon$
  else
    $\gamma_i=0$ with $i\in\mathcal{I}_\varepsilon$
    $\gamma_i=\frac{v_i}{u}-\frac{\sum_{j\notin\mathcal{I}}\frac{v_j}{u}-1_\varepsilon}{q+p-\mathrm{Card}(\mathcal{I}_\varepsilon)}$ with $i\notin\mathcal{I}_\varepsilon$
    terminate the loop
  end if
end if

Algorithm 2 Solving problem (12) with $\mathcal{S}=\{\boldsymbol{\gamma}\,|\,\boldsymbol{\gamma}^\top\mathbf{1}=1\}$

build a set $\mathcal{I}$, let $i\notin\mathcal{V}$ be in the set $\mathcal{I}$
start a loop
find the max element v_i with $i\notin\mathcal{I}$
put i into the set $\mathcal{I}$
if $\frac{v_i}{u}<\frac{\sum_{j\notin\mathcal{I}}\frac{v_j}{u}-1}{q+p-\mathrm{Card}(\mathcal{I})}$ **then**
  remove i from the set $\mathcal{I}$
else
  $\gamma_i=0$ with $i\in\mathcal{I}$
  $\gamma_i=\frac{v_i}{u}-\frac{\sum_{j\notin\mathcal{I}}\frac{v_j}{u}-1}{q+p-\mathrm{Card}(\mathcal{I})}$ with $i\notin\mathcal{I}$
  terminate the loop
end if

equations, which complicates the solution process. To address this, we consider an alternative cubic system derived from a suitable surrogate function for φ_{Gau}, which enables a more efficient solution.

Define $\tilde{\underline{\boldsymbol{\beta}}}=\left[1,-\underline{\boldsymbol{\beta}}^\top\right]^\top$, and $\boldsymbol{h}_t=[h_t,\ldots,h_{t-p}]^\top$, for $t=1,\ldots,n$. Applying Lemma 1 to the third term in the summation of φ_{Gau}, we have

$$\begin{aligned}\left(\boldsymbol{h}_t^\top\tilde{\underline{\boldsymbol{\beta}}}-o_t\right)^2&=\boldsymbol{h}_t^\top\tilde{\underline{\boldsymbol{\beta}}}\tilde{\underline{\boldsymbol{\beta}}}^\top\boldsymbol{h}_t-2o_t\boldsymbol{h}_t^\top\tilde{\underline{\boldsymbol{\beta}}}+\mathrm{const.}\\&\le\left\|\tilde{\underline{\boldsymbol{\beta}}}\right\|^2\boldsymbol{h}_t^\top\boldsymbol{h}_t+2\boldsymbol{h}_t^\top\boldsymbol{r}_t+\mathrm{const.},\end{aligned}\tag{15}$$

where const. represents a constant term and

$$\boldsymbol{r}_t=\left(\underline{\boldsymbol{h}}_t^\top\tilde{\underline{\boldsymbol{\beta}}}-o_t\right)\tilde{\underline{\boldsymbol{\beta}}}-\left\|\tilde{\underline{\boldsymbol{\beta}}}\right\|^2\underline{\boldsymbol{h}}_t.$$

By integrating the other two terms in φ_{Gau}, we derive the surrogate function $\bar{\varphi}_{\mathrm{Gau}}$ as follows:

$$\bar{\varphi}_{\mathrm{Gau}}(\boldsymbol{h})=\sum_{t=1}^{n}\left[\log h_t+\frac{\epsilon_t^2}{h_t}+\frac{\eta}{2}\left\|\tilde{\underline{\boldsymbol{\beta}}}\right\|^2\boldsymbol{h}_t^\top\boldsymbol{h}_t+\eta\boldsymbol{h}_t^\top\boldsymbol{r}_t\right],$$

where the variables h_t, for $t=1,\ldots,n$, become separable. Now, we are prepared to solve the subproblems for $\boldsymbol{h}$ based on the surrogate function $\bar{\varphi}_{\mathrm{Gau}}$. By computing the partial derivative of $\bar{\varphi}_{\mathrm{Gau}}$ with respect to each h_t, setting it to zero, and multiplying both sides by h_t^2, we obtain

$$\eta\kappa_t\left\|\tilde{\underline{\boldsymbol{\beta}}}\right\|^2h_t^3+\eta\sum_{j=0}^{p}\left[\boldsymbol{r}_{t+j}\right]_{j+1}h_t^2+h_t-\epsilon_t^2=0,\tag{16}$$

where

$$\kappa_t=\begin{cases}p+1, & 1\le t\le n-p,\\ n-t+1, & n-p<t\le n.\end{cases}$$

The solution h_t^+ is the positive root of (16), required by the domain of the logarithmic function, and can be computed using the cubic formula. The result is stated below.

Proposition 3: For the cubic equation in (16), it has either one or three positive roots. Besides,

1) If there is only one root h_t^+, it attains the minimum of $\bar{\varphi}_{\mathrm{Gau}}$;
2) If there are three roots, either the least one $h_t^\vee$ or the greatest one $h_t^\wedge$ attains the minimum of $\bar{\varphi}_{\mathrm{Gau}}$, and hence h_t^+ is determined by $h_t^+=\arg\min_{h_t\in\{h_t^\vee,h_t^\wedge\}}\bar{\varphi}_{\mathrm{Gau}}\left(\underline{h}_1,\ldots,\underline{h}_{t-1},h_t,\underline{h}_{t+1},\ldots,\underline{h}_n\right)$.

Proof: See Appendix A. ■

To summarize, the ML estimation of Gaussian GARCH using BMM involves cyclically updating the three variables: ω, $\boldsymbol{\gamma}$, and $\boldsymbol{h}$, as detailed in Sections IV-C1, IV-C2, and IV-C3. The complete procedure is outlined in Algorithm 3.

V. ML Estimation of Student's t GARCH

A. The Student's t ML Estimation Problem

In many applications, the residual series $\{\epsilon_t\}$ is often observed to deviate from conditional normality, which calls for considering the non-Gaussian noise [45]. In this paper, we also consider the Student's t GARCH model [14], which captures the heavy-tailed property in volatility modeling, making it more robust to outliers and deviations from normality in data [46]. Compared to the Gaussian distribution, a Student's t distribution introduces an additional parameter, referred to as the shape parameter (also known as the degrees of freedom), which governs the tail heaviness of the distribution.

The probability density function of the Student's t distribution with shape parameter ν is given by

$$s(x)=\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}},$$

where Γ is the usual gamma function (i.e., $\Gamma(x)=\int_0^\infty y^{x-1}\exp(-y)\mathrm{d}y$). When $\nu\to\infty$, the Student's t distribution will degenerate to the Gaussian distribution. For the Student's t distribution, we have $\mathrm{Var}(x)=\frac{\nu}{\nu-2}$ for $\nu>2$. We use $x=\sqrt{\frac{\nu}{\nu-2}}z_t$ to obtain the following probability density function of z_t:

$$d_{\mathrm{Stu}}(z_t)=\frac{(\nu-2)^{\frac{\nu}{2}}\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{\nu}{2}\right)}\left(\nu-2+z_t^2\right)^{-\frac{\nu+1}{2}}.$$

Let $z_t = \frac{\epsilon_t}{\sqrt{h_t}} \,|\, \mathcal{F}_{t-1}$, where h_t is defined in (4) and $\mathcal{F}_{t-1} = \{\epsilon_{1-q}, \ldots, \epsilon_{t-1}, h_{1-p}, \ldots, h_0\}$. The negative conditional log-likelihood function of ϵ_t, for $t = 1, \ldots, n$, is given by

$$\begin{aligned} &-\log L_{\mathrm{Stu}}\left(\omega, \boldsymbol{\gamma}, \nu; \epsilon_1, \ldots, \epsilon_n \,|\, \mathcal{F}_0\right) \\ =& -\log \prod_{t=1}^{n} \frac{1}{\sqrt{h_t}} d_{\mathrm{Stu}}\left(\frac{\epsilon_t}{\sqrt{h_t}} \,\middle|\, \mathcal{F}_{t-1}\right) \\ =& \sum_{t=1}^{n}\left[\frac{1}{2}\log h_t - \log d_{\mathrm{Stu}}\left(\frac{\epsilon_t}{\sqrt{h_t}} \,\middle|\, \mathcal{F}_{t-1}\right)\right]. \end{aligned}$$

The ML estimation problem for Student's t GARCH is given in (17), where $2_\varepsilon = 2 - \varepsilon$. In problem (17), the shape parameter ν is treated as an unknown to be jointly estimated along with the GARCH parameters from the data. We impose an upper limit l for ν, a common practice in existing numerical solvers for Student's t GARCH estimation. Since a Student's t distribution is much like a Gaussian distribution when ν is relatively large (for instance, $\nu > 30$), it is customary to set the parameter l to 100. As in Section IV-B, we further transform problem (17) into a penalty form in (18).

B. Problem Solving

In this section, we derive an algorithm to solve problem (18) via BMM, where the optimization variables are divided into four blocks: ω, $\boldsymbol{\gamma}$, $\boldsymbol{h}$, and ν.

1) Solving the ω-Block and $\boldsymbol{\gamma}$-Block: The subproblems with respect to ω and $\boldsymbol{\gamma}$ are identical to problems (8) and (10), respectively, and can thus be solved using the methods presented in Sections IV-C1 and IV-C2.

2) Solving the $\boldsymbol{h}$-Block: The subproblem with respect to $\boldsymbol{h}$ is given by

$$\begin{aligned} \min_{\boldsymbol{h}} \quad & \sum_{t=1}^{n}\left[\log h_t + (\underline{\nu}+1)\log\left(\underline{\nu} - 2 + \frac{\epsilon_t^2}{h_t}\right)\right. \\ & \left. + \frac{\eta}{2}\left(\boldsymbol{h}_t^\top \tilde{\boldsymbol{\beta}}_t - o_t\right)^2\right], \end{aligned} \tag{19}$$

where definitions of $\boldsymbol{h}_t$, $\tilde{\boldsymbol{\beta}}$, and o_t are the same as those defined in Section IV-C3. Problem (19) is an unconstrained non-convex program. Denote the objective function as φ_{Stu}. As in the Gaussian case, we majorize φ_{Stu} by the same tricks in (15) to obtain a surrogate function $\bar{\varphi}_{\mathrm{Stu}}$ as follows:

$$\begin{aligned} \bar{\varphi}_{\mathrm{Stu}}(\boldsymbol{h}) = \sum_{t=1}^{n}\Bigg[&\log h_t + (\underline{\nu}+1)\log\left(\underline{\nu} - 2 + \frac{\epsilon_t^2}{h_t}\right) \\ &+ \frac{\eta}{2}\left\|\tilde{\underline{\boldsymbol{\beta}}}\right\|^2 \boldsymbol{h}_t^\top \boldsymbol{h}_t + \eta \boldsymbol{h}_t^\top \boldsymbol{r}_t\Bigg]. \end{aligned}$$

where $\boldsymbol{r}_t$ is defined as in Section IV-C. Then, we compute the first derivative of $\bar{\varphi}_{\mathrm{Stu}}$ with respect to each h_t for $t = 1, \ldots, n$, and set it equal to 0, leading to the following cubic equation

$$\begin{aligned} &\eta\kappa_t\left\|\tilde{\underline{\boldsymbol{\beta}}}\right\|^2 h_t^3 + \eta\left(\sum_{i=0}^{p}[\boldsymbol{r}_{t+i}]_{i+1} + \kappa_t\left\|\tilde{\underline{\boldsymbol{\beta}}}\right\|^2 \frac{\epsilon_t^2}{\underline{\nu}-2}\right) h_t^2 \\ &+ \left(\eta\frac{\epsilon_t^2}{\underline{\nu}-2}\sum_{i=0}^{p}[\boldsymbol{r}_{t+i}]_{i+1} + 1\right) h_t - \frac{\underline{\nu}}{\underline{\nu}-2}\epsilon_t^2 = 0. \end{aligned} \tag{20}$$

It can be verified that the positivity of the coefficients in (20) is the same as that in (16). By Proposition 3, this cubic equation can be resolved in the same manner.

3) Solving the ν-Block: The ν-block subproblem is

$$\begin{aligned} \min_{\nu} \quad & -n\nu\log(\nu-2) + 2n\log\Gamma\left(\frac{\nu}{2}\right) \\ & - 2n\log\Gamma\left(\frac{\nu+1}{2}\right) + \sum_{t=1}^{n}(\nu+1)\log(\nu - 2 + w_t) \\ \text{s.t.} \quad & 2_\varepsilon \le \nu \le l, \end{aligned} \tag{21}$$

where $w_t = \frac{\epsilon_t^2}{h_t}$. Denote the objective function as Υ. The solution to the above problem is related to the zeros of Υ', for which we present the following result.

Theorem 4: For the function Υ', the following relations hold true:

1) If $w_t \in [3 - \sqrt{6}, 3 + \sqrt{6}]$ for $t = 1, \ldots, n$, Υ' has no zeros on $(2, \infty)$;
2) If $w_t \in \left[0, 3 - \sqrt{6}\right) \cup \left(3 + \sqrt{6}, \infty\right)$ for $t = 1, \ldots, n$, Υ' has at least one zero on $(2, \infty)$.

Proof: See Appendix B of the Supplementary Material. ■

In Theorem 4, we have provided two sufficient conditions on the zeros of Υ'. However, these conditions on w_t may be too strict to satisfy in practice. In the following, based on the BMM idea, we propose to solve ν via optimizing a proper surrogate function for Υ.

$$\begin{aligned} \min_{\omega, \boldsymbol{\gamma}, \boldsymbol{h}, \nu} \quad & \sum_{t=1}^{n}\left[\frac{1}{2}\log\pi - \frac{\nu}{2}\log(\nu-2) + \log\Gamma\left(\frac{\nu}{2}\right) - \log\Gamma\left(\frac{\nu+1}{2}\right) + \frac{1}{2}\log h_t + \frac{\nu+1}{2}\log\left(\nu - 2 + \frac{\epsilon_t^2}{h_t}\right)\right] \\ \text{s.t.} \quad & h_t = \omega + \boldsymbol{\gamma}^\top \boldsymbol{c}_t,\ t = 1, \ldots, n \\ & \omega \ge 0_\varepsilon,\ \boldsymbol{\gamma} \ge \mathbf{0},\ \boldsymbol{\gamma} \in \mathcal{S},\ 2_\varepsilon \le \nu \le l \end{aligned} \tag{17}$$

$$\begin{aligned} \min_{\omega, \boldsymbol{\gamma}, \boldsymbol{h}, \nu} \quad & \sum_{t=1}^{n}\left[-\nu\log(\nu-2) + 2\log\Gamma\left(\frac{\nu}{2}\right) - 2\log\Gamma\left(\frac{\nu+1}{2}\right) + \log h_t + (\nu+1)\log\left(\nu - 2 + \frac{\epsilon_t^2}{h_t}\right) + \frac{\eta}{2}\left(h_t - \omega - \boldsymbol{\gamma}^\top \boldsymbol{c}_t\right)^2\right] \\ \text{s.t.} \quad & \omega \ge 0_\varepsilon,\ \boldsymbol{\gamma} \ge \mathbf{0},\ \boldsymbol{\gamma} \in \mathcal{S},\ 2_\varepsilon \le \nu \le l \end{aligned} \tag{18}$$

Algorithm 3 ML Estimation of Gaussian GARCH

Input: ϵ_t $(t=1-q,\ldots,n)$, h_t $(t=1-p,\ldots,0)$, η
Initialization: ω, $\boldsymbol{\gamma}$, $\boldsymbol{h}$
while stopping criteria not met **do**
    update ω with (9);
    update $\boldsymbol{\gamma}$ with (13) or Proposition 2;
    update $\boldsymbol{h}$ with Proposition 3 applied to (16);
end while
Output: ω, $\boldsymbol{\gamma}$

Algorithm 4 ML Estimation of Student's t GARCH

Input: ϵ_t $(t=1-q,\ldots,n)$, h_t $(t=1-p,\ldots,0)$, η
Initialization: ω, $\boldsymbol{\gamma}$, $\boldsymbol{h}$, ν
while stopping criteria not met **do**
    update ω with (9);
    update $\boldsymbol{\gamma}$ with (13) or Proposition 2;
    update $\boldsymbol{h}$ with Proposition 3 applied to (20);
    update ν with the zero of (24);
end while
Output: ω, $\boldsymbol{\gamma}$, ν

It can be verified that the combination of the first two terms in Υ, i.e., $-n\nu\log(\nu-2)+2n\log\Gamma\left(\frac{\nu}{2}\right)$, and the last two terms, i.e, $-2n\log\Gamma\left(\frac{\nu+1}{2}\right)+(\nu+1)\sum_{t=1}^{n}\log(\nu-2+w_t)$, are convex and concave in ν, respectively. Therefore, one way to obtain a convex surrogate function is to retain the first two terms and linearize the last two at a given point $\underline{\nu}$. This is precisely the surrogate function employed in the classical expectation-maximization (EM) algorithm for Student's t ML estimation [61]. In this paper, we develop a more refined surrogate function for Υ, which offers a tighter approximation than that used in EM. To achieve this, we decompose the last term in Υ into two parts as follows:

$$\sum_{t=1}^{n}(\nu+1)\log(\nu-2+w_t)$$
$$=n(\nu+1)\log(\nu+1)+\sum_{t=1}^{n}(\nu+1)\log\left(\frac{\nu-2+w_t}{\nu+1}\right), \tag{22}$$

where the first term is combined with the remaining terms in Υ to form

$$\Upsilon_1(\nu)=-n\nu\log(\nu-2)+2n\log\Gamma\left(\frac{\nu}{2}\right)$$
$$-2n\log\Gamma\left(\frac{\nu+1}{2}\right)+n(\nu+1)\log(\nu+1), \tag{23}$$

and the second term is defined as

$$\Upsilon_2(\nu)=\sum_{t=1}^{n}(\nu+1)\log\left(\frac{\nu-2+w_t}{\nu+1}\right).$$

Lemma 5: The function Υ_1 in (22) is strictly convex on $(2,\infty)$; the function Υ_2 in (23) is concave on $(2,\infty)$.

Proof: See Appendix C of the Supplementary Material. ■

Based on Lemma 5, we construct a linear surrogate function for Υ_2 as follows:

$$\bar{\Upsilon}_2(\nu)=\sum_{t=1}^{n}\left[\left(\log\left(\frac{\underline{\nu}-2+w_t}{\underline{\nu}+1}\right)+\frac{3-w_t}{\underline{\nu}-2+w_t}\right)(\nu-\underline{\nu})\right.$$
$$\left.+(\underline{\nu}+1)\log\left(\frac{\underline{\nu}-2+w_t}{\underline{\nu}+1}\right)\right].$$

Therefore, a strictly convex surrogate for Υ is obtained as $\bar{\Upsilon}=\Upsilon_1+\bar{\Upsilon}_2$. One can verify that $\bar{\Upsilon}$ is a tighter surrogate to Υ than the surrogate function derived from EM since it preserves more convexity of Υ.

Then, we update ν on $(2,\infty)$ by finding the unique zero of $\bar{\Upsilon}'(\nu)$, which is computed as

$$\bar{\Upsilon}'(\nu)=-n\log(\nu-2)-n\frac{\nu}{\nu-2}+n\psi\left(\frac{\nu}{2}\right)$$
$$-n\psi\left(\frac{\nu+1}{2}\right)+n\log(\nu+1)+n$$
$$+\sum_{t=1}^{n}\left[\log\left(\frac{\underline{\nu}-2+w_t}{\underline{\nu}+1}\right)+\frac{3-w_t}{\underline{\nu}-2+w_t}\right], \tag{24}$$

where $\psi(x)=\frac{\Gamma'(x)}{\Gamma(x)}=\int_0^\infty\frac{\exp(-y)}{y}-\frac{\exp(-xy)}{1-\exp(-y)}\mathrm{d}y$ is the digamma function. Regarding $\bar{\Upsilon}'$, we have the following result.

Proposition 6: For the function $\bar{\Upsilon}'$, the following relations hold true:

1) If $w_t=3$ for all $t=1,\ldots,n$, $\bar{\Upsilon}'$ has no zeros on $(2,\infty)$;
2) If there exists $w_t\neq 3$ for $t=1,\ldots,n$, $\bar{\Upsilon}'$ has a unique zero on $(2,\infty)$.

Proof: See Appendix D of the Supplementary Material. ■

The first case in Proposition 6 can hardly arise in practice. Therefore, we focus only on the second case, i.e., the update ν^+ can be chosen as the only stationary point of $\bar{\Upsilon}(\nu)$. Finally, considering $2_\varepsilon\leq\nu\leq l$, one can obtain ν^+ by applying a one-dimensional search method, e.g., Brent's method [62, Chapter 4] on the surrogate function $\bar{\Upsilon}(\nu)$.

We summarize the overall BMM algorithm for solving the Student's t ML estimation problem in Algorithm 4.

VI. Convergence and Complexity Analysis

A. Convergence Properties

The convergence of Algorithms 3 and 4 can be established based on the generic proof in [63]. Since all the variable updates in Algorithms 3 and 4 have unique solutions, it can be shown that the sequence $(\omega,\boldsymbol{\gamma},\boldsymbol{h})$ generated by Algorithm 3 converges to the set of Karush-Kuhn-Tucker points of problem (7), and the sequence $(\omega,\boldsymbol{\gamma},\boldsymbol{h},\nu)$ generated by Algorithm 4 converges to the set of Karush-Kuhn-Tucker points of problem (18). Since Algorithms 3 and 4 both utilize the BMM framework, they can benefit from some "off-the-shelf" accelerators to enhance algorithm convergence, such as the squared iterative method (SQUAREM) [64]. SQUAREM treats each iteration of an EM algorithm as a mapping function. The SQUAREM then employs

an approximated Newton's method to find the fixed point of this mapping, which allows it to attain superlinear convergence. SQUAREM only requires the EM updating map, so it can be readily extended to any EM-type algorithm. That is to say, SQUAREM can be applied to any BMM algorithm, including Algorithm 3 and Algorithm 4. Interested readers may refer to [64] to see why it works.

B. Computational Complexity

We analyze the per-iteration computational complexity of Algorithm 3 and Algorithm 4. For Algorithm 3, we update ω by (9) with complexity $O\left(n(q+p)\right)$. In the update of $\boldsymbol{\gamma}$, since $\sum_{t=1}^{n}\left\|\underline{\boldsymbol{c}}_t\right\|^2 \geq \lambda_{\max}\left(\sum_{t=1}^{n}\underline{\boldsymbol{c}}_t\underline{\boldsymbol{c}}_t^\top\right)$, we approximate u as $\sum_{t=1}^{n}\left\|\underline{\boldsymbol{c}}_t\right\|^2$ with the computational complexity being $O\left(n(q+p)\right)$; the time complexity for computing $\boldsymbol{v}$ is $O\left(q+p\right)$. If the stationary constraint is not considered, $\boldsymbol{\gamma}$ can be updated with (13) in time complexity $O\left(n(q+p)\right)$. Hence, the total time complexity is $O\left(n(q+p)\right)$. If the constraint set $\mathcal{S}$ is given by either $\left\{\boldsymbol{\gamma}\,|\,\boldsymbol{\gamma}^\top\mathbf{1}\leq 1_\varepsilon\right\}$ or $\left\{\boldsymbol{\gamma}\,|\,\boldsymbol{\gamma}^\top\mathbf{1}=1\right\}$, then in the worst case we need to check $q+p$ elements to determine whether they belong to the set $\mathcal{I}_\varepsilon$ or $\mathcal{I}$, respectively. The complexity for updating $\boldsymbol{\gamma}$ is $O\left((q+p)^2\right)$. Hence, the total time complexity is $O\left(n(q+p)+(q+p)^2\right)$. For updating $\boldsymbol{h}$, the complexity of computing the coefficients of each cubic equation in (16) is $O(q+p)$, and solving each cubic equation by the cubic formula costs $O(1)$. Since there are n equations, the total complexity of solving the cubic system of equations is $O\left(n(q+p)\right)$. For Algorithm 4, the time complexities for computing the ω, $\boldsymbol{\gamma}$, and $\boldsymbol{h}$ blocks are identical to those in Algorithm 3. For updating ν, we need to take $O\left(n\right)$ to obtain $\bar{Y}'$ in (24). Then, if we choose Brent's method to solve the ν-block, it takes $O\left(\log_2\frac{l-2_\varepsilon}{\delta}\right)$ steps to find the zero, where δ is the error tolerance. Hence, the total time complexity is $O\left(n+\log_2\frac{l-2_\varepsilon}{\delta}\right)$.

VII. Extensions

A. M-Estimation

In the previous sections, we have demonstrated how the proposed unified penalty method can be applied to the ML estimation of GARCH models, using Gaussian and Student's t distributions as illustrative examples. It is worth noting that our approach is also applicable to other conditional distributions, such as the generalized error distribution [7] and the generalized hyperbolic distribution [65]. All the aforementioned ML estimation procedures aim to maximize the conditional log-likelihood function. In more general settings, maximum likelihood-type estimation, commonly referred to as M-estimation, is widely adopted for GARCH models [42], [43], [66], [67]. This framework is particularly appealing because it allows the use of robust loss functions such as the Huber loss [66], which reduce the influence of outliers and lead to robust parameter estimates. The general M-estimation objective takes the form

$$\sum_{t=1}^{n}\left[\frac{1}{2}\log h_t+\rho\left(\frac{\epsilon_t}{\sqrt{h_t}}\,\middle|\,\mathcal{F}_{t-1}\right)\right],$$

where ρ denotes a user-defined loss function. Here, $\sqrt{h_t}$ represents a time-varying dispersion parameter, which can be the conditional volatility when second-order moments exist, and more generally serves as a conditional scale parameter in heavy-tailed settings [68], [69], [70], [71]. The M-estimation framework encompasses a wide range of GARCH estimation models. For example, when $\rho(x)=\frac{1}{2}x^2$, the objective corresponds to the Gaussian ML estimation loss in (6); when $\rho(x)=\frac{1}{2}|x|$, it corresponds to the Laplace ML estimation loss; when $\rho(x)=\frac{1}{2}x^2I(|x|\leq k)+\left(k|x|-\frac{1}{2}k^2\right)I(|x|>k)$ with I the indicator function and k a constant, it becomes the Huber loss. The proposed penalty method is applicable to solving this generalized formulation.

Remark 7 (On quasi-ML estimation of GARCH): Quasi-ML estimation frequently appears in the GARCH literature [72], [73], [74]. It involves estimating model parameters under an assumed conditional distribution that differs from the true distribution of the innovations, and the resulting objective is referred to as a quasi-likelihood function. When the assumed distribution matches the true one, the quasi-ML estimation reduces to the ML estimation. Both ML and quasi-ML estimation fall within the broader class of M-estimation. Given that the assumed distribution is generally misspecified, it is crucial to study the statistical properties of the resulting estimators under model misspecification. Accordingly, much of the existing work has focused on theoretical analysis, establishing conditions under which quasi-ML estimators are consistent and asymptotically normal [73]. Although this paper primarily focuses on ML estimation, the proposed method is also directly applicable to estimation procedures based on quasi-likelihood functions.

B. Estimation of GARCH Variants

Although this paper focuses on the ML estimation of standard GARCH and integrated GARCH models, the proposed penalty method for decoupling the volatility equations can also be extended to other GARCH variants. In particular, it is applicable to asymmetric GARCH models [75], which are designed to capture the leverage effect in financial time series by allowing positive and negative shocks to have different impacts on volatility. Examples include the exponential GARCH model [7], the threshold GARCH model [8] (or GJR-GARCH [9]), and the asymmetric power ARCH model [10]. Besides, the method is also applicable to GARCH models with exogenous variables in the generalized GARCH framework of Bollerslev [5]. In Appendix E of the Supplementary Material, we illustrate how the proposed method extends to the estimation of GARCH variants.

C. Joint Estimation With a Conditional Mean Model

In many time series applications, it is common to jointly estimate a conditional mean model alongside a GARCH variance model. The mean component may take various forms, such as a constant, a linear regression with exogenous variables, or an autoregressive structure. Joint estimation enables the model to capture both the mean dynamics and the time-varying volatility of the data. In Appendix F of the Supplementary Material, we

demonstrate the application of the BMM algorithm to joint estimation with an autoregressive model.

VIII. Numerical Experiments

In this section, we evaluate the performance of the proposed BMM method through numerical experiments on both synthetic and real-world datasets. Owing to space limitations, we focus on the estimation performance of the vanilla GARCH model [5]. The benchmark algorithms include BHHH [5], [14], BFGS [47], L-BFGS [48], [49], SQP [49], [50], [51], and IPM [51]. All these iterative algorithms are implemented in MATLAB with random initialization to ensure a consistent and fair comparison. Simulations are conducted on a machine equipped with a 3.3 GHz Intel Xeon W CPU.

A. Synthetic Data

1) Specification of GARCH Models: We consider two GARCH models as follows:

- Gaussian GARCH$(2,3)$ with parameters $\omega^\star = 0.01$, $\boldsymbol{\alpha}^\star = [0.1, 0.3]^\top$, $\boldsymbol{\beta}^\star = [0.2, 0.29, 0.1]^\top$;
- Student's t GARCH$(1,1)$ with parameters $\omega^\star = 0.01$, $\boldsymbol{\alpha}^\star = 0.4$, $\boldsymbol{\beta}^\star = 0.59$, $\nu^\star = 4$.

2) Performance Metrics: Let $\boldsymbol{x}^\star$ denote the ground truth. Consider m Monte Carlo simulations, where the estimate from the i-th experiment, $i = 1, \ldots, m$, is denoted by $\hat{\boldsymbol{x}}_i$. The estimation accuracy of parameters is measured based on the root mean squared error (RMSE), defined as

$$\mathrm{RMSE}(\boldsymbol{x}) = \sqrt{\frac{1}{m}\sum_{i=1}^{m} \|\hat{\boldsymbol{x}}_i - \boldsymbol{x}^\star\|^2}.$$

The variance of the estimator is measured by the standard error (SE), which is given by

$$\mathrm{SE}(\boldsymbol{x}) = \sqrt{\frac{1}{m}\sum_{i=1}^{m} \left\| \hat{\boldsymbol{x}}_i - \frac{1}{m}\sum_{j=1}^{m} \hat{\boldsymbol{x}}_j \right\|^2}.$$

We report the number of cases in which the estimates violate the stationarity constraint. Additionally, we also report the average value of the negative conditional log-likelihood function (the objective function in (6) for Gaussian GARCH and the objective function in (17) for Student's t GARCH) and the average CPU time in seconds.

3) Estimation Performance: We consider $m = 500$ simulation experiments, where each experiment contains $n = 50$ samples and p pre-sample observations needed for GARCH initialization. Since the quasi-Newton algorithms, including BHHH, BFGS, and L-BFGS, frequently report estimates that violate the constraints, we only include results from experiments where all six algorithms observe the constraints. We summarize the results for Gaussian GARCH$(2,3)$ in Table I. From Table I, we find that the proposed BMM algorithm can attain the lowest RMSE for all the GARCH parameters and also for $\boldsymbol{h}$[2]. Moreover, the BMM algorithm achieves the lowest

[2] $\boldsymbol{h}^\star$ is computed from the GARCH model (4) based on $\omega^\star$, $\boldsymbol{\alpha}^\star$, and $\boldsymbol{\beta}^\star$. Since $\boldsymbol{h}^\star$ is different for different Monte Carlo iterations, we did not report the SE for $\boldsymbol{h}$.

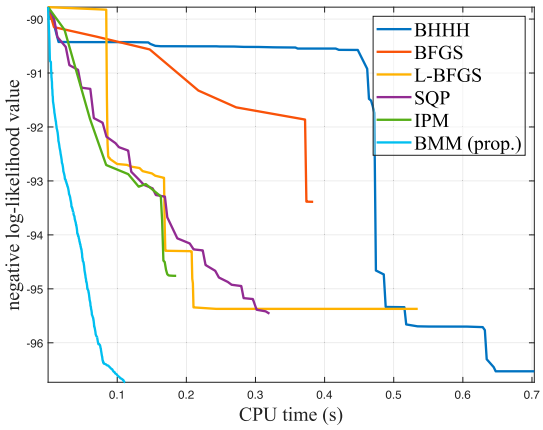


Fig. 2. Convergence of the objective value from six different methods.

average log-likelihood value and average CPU time among all methods. In Fig. 2, we depict the average log-likelihood value versus CPU time for all six algorithms. It can be observed that the BMM method achieves a faster convergence speed with monotonic convergence behavior. The comparison results for the Student's t GARCH$(1,1)$ model are reported in Table II. Similar to the Gaussian case, the proposed BMM algorithm achieves the lowest RMSE, objective values, and CPU time compared to existing methods.

B. Real Data

1) Log-Return Modeling in Finance: In this section, we examine the estimation performance of the BMM algorithm for GARCH estimation based on real stock market data. We select the daily log-returns of HSBC Holdings plc (HSBC) from the New York Stock Exchange, spanning from January 1st, 2022, to December 31st, 2022, which comprise a total of 251 observations. We remove the mean of this time series using its sample mean estimate to obtain the residual series $\{\epsilon_t\}$. We split the first 201 observations to train both the Gaussian and Student's t GARCH models, and the last 50 observations are used to test the forecasting performance. For the model order selection, since low-order GARCH models are typically employed in practice, we choose from GARCH$(0,0)$, GARCH$(1,0)$, GARCH$(1,1)$, GARCH$(1,2)$, GARCH$(2,0)$, GARCH$(2,1)$, and GARCH$(2,2)$. We then utilize the Bayesian information criterion (BIC) to determine the specific orders q and p. As reported in Table III, we observe that within the Gaussian GARCH models, GARCH$(1,1)$ and GARCH$(2,1)$ exhibit relatively lower BIC values. A similar trend is observed for the Student's t GARCH models, where GARCH$(1,1)$ and GARCH$(2,1)$ provide superior performance. Furthermore, the Student's t GARCH models generally exhibit smaller BIC values than the Gaussian GARCH models, suggesting that Student's t GARCH provides a better fit for stock market data. Accordingly, we select the Student's t GARCH$(2,1)$ for volatility prediction, and the results are shown in Fig. 1.

The goodness of the models is tested based on their forecasting performance. However, the true volatility is unobservable, which means that the accuracy of volatility forecasting cannot be directly measured in practice. Hence, it is essential to establish a metric to evaluate the volatility forecasting performance.

TABLE I
PERFORMANCE COMPARISONS OF SIX METHODS FOR GAUSSIAN GARCH$(2,3)$

Method	RMSE(ω) (SE(ω))	RMSE($\boldsymbol{\alpha}$) (SE($\boldsymbol{\alpha}$))	RMSE($\boldsymbol{\beta}$) (SE($\boldsymbol{\beta}$))	RMSE($\boldsymbol{h}$)	Constraint Violation	Log-likelihood	CPU Time (s)
BHHH	0.0424 (0.0094)	0.1300 (0.0045)	0.1389 (0.0047)	0.9823	282/500	10.3007	0.1849
BFGS	0.3391 (0.0109)	0.0959 (0.0056)	0.1091 (0.0049)	0.7087	277/500	10.8624	0.2192
L-BFGS	0.0748 (0.0029)	0.0964 (0.0067)	0.1100 (0.0056)	0.8774	120/500	10.3158	0.0894
SQP	0.0831 (0.0027)	0.1118 (0.0083)	0.1315 (0.0059)	0.7266	**0/500**	9.4549	0.2853
IPM	0.0424 (0.0011)	0.0872 (0.0038)	0.1131 (0.0027)	0.5799	**0/500**	9.3266	0.1308
BMM (prop.)	0.0141 (0.0059)	0.0954 (0.0042)	0.1077 (0.0057)	**0.5585**	**0/500**	**8.8040**	**0.0850**

TABLE II
PERFORMANCE COMPARISONS OF SIX METHODS FOR STUDENT'S t GARCH$(1,1)$

Method	RMSE(ω) (SE(ω))	RMSE($\boldsymbol{\alpha}$) (SE($\boldsymbol{\alpha}$))	RMSE($\boldsymbol{\beta}$) (SE($\boldsymbol{\beta}$))	RMSE(ν) (SE(ν))	RMSE($\boldsymbol{h}$)	Constraint Violation	Log-likelihood	CPU Time (s)
BHHH	0.5213 (0.0667)	0.2105 (0.0447)	0.2653 (0.0526)	0.2381 (0.2202)	3.5912	419/500	49.2463	0.2085
BFGS	0.4593 (0.0817)	0.2579 (0.0522)	0.2884 (0.0837)	0.2054 (1.4983)	2.4222	362/500	47.9485	0.2111
L-BFGS	0.3571 (0.0566)	0.2441 (0.0572)	0.2612 (0.0694)	0.1120 (0.2636)	2.5306	346/500	48.0595	0.1935
SQP	0.4445 (0.0561)	0.2385 (0.0363)	0.2811 (0.0409)	0.2189 (1.8173)	2.7815	**0/500**	49.9377	0.1708
IPM	0.3197 (0.0451)	0.2163 (0.0141)	0.2315 (0.0260)	0.6303 (4.2433)	1.9352	**0/500**	49.0684	0.1664
BMM (prop.)	0.2445 (0.0447)	0.2126 (0.0459)	0.2110 (0.0437)	0.1102 (0.6340)	**1.7234**	**0/500**	**47.1058**	**0.1625**

TABLE III
THE BIC OF GAUSSIAN AND STUDENT'S t GARCH MODELS FOR HSBC

Gaussian GARCH				Student's t GARCH			
(q,p)	BIC	(q,p)	BIC	(q,p)	BIC	(q,p)	BIC
$(0,0)$	-941.4370			$(0,0)$	-965.8462		
$(1,0)$	-943.0760	$(2,0)$	-939.0760	$(1,0)$	-964.4285	$(2,0)$	-956.0430
$(1,1)$	$\mathbf{-987.1239}$	$(2,1)$	$\mathbf{-948.9591}$	$(1,1)$	$\mathbf{-1015.4725}$	$(2,1)$	$\mathbf{-1027.3791}$
$(1,2)$	-934.4087	$(2,2)$	-943.6632	$(1,2)$	-948.9946	$(2,2)$	-989.5558

Given the observed residuals $\{\epsilon_t\}$ in the test data and the corresponding predicted conditional standard deviation $\left\{\hat{h}_t\right\}$, we can obtain the standardized residual series $\left\{\hat{z}_t \,\middle|\, \hat{z}_t = \frac{\epsilon_t}{\sqrt{\hat{h}_t}}\right\}$. Since z_t is defined as an i.i.d. random variable, we can employ the Ljung-Box Q-test as a benchmark to measure the autocorrelation of the standardized residuals $\{\hat{z}_t\}$ [2], [76]. The Ljung-Box Q-test statistic for lag ℓ is defined as follows:

$$Q(\ell) = n(n+2)\sum_{k=1}^{\ell}\frac{\hat{\varrho}^2(k)}{n-k},$$

where n denotes the sample size, and

$$\hat{\varrho}(k) = \frac{\sum_{t=k+1}^{n}\left(\hat{z}_t - \frac{1}{n}\sum_{i=1}^{n}\hat{z}_i\right)\left(\hat{z}_{t-k} - \frac{1}{n}\sum_{i=1}^{n}\hat{z}_i\right)}{\sum_{t=1}^{n}\left(\hat{z}_t - \frac{1}{n}\sum_{i=1}^{n}\hat{z}_i\right)^2}$$

is the sample autocorrelation coefficient of series $\{\hat{z}_t\}$ at lags k. The volatility we forecast is more accurate if the Q-test statistic has a lower value.

Besides, the accuracy of the forecasted volatility can also be evaluated by the skewness and kurtosis of the standardized residual series $\{\hat{z}_t\}$. Specifically, if the measured skewness and kurtosis are close to the theoretical values under the specific conditional distribution assumptions on the standardized residual series $\{z_t\}$, the forecasted volatility is considered accurate. Since the skewness and kurtosis of standardized residuals differ between Gaussian and Student's t distributions, to directly compare the relative performance of Gaussian GARCH and Student's t GARCH models, we introduce the Rosenblatt transformation [77], [78], which transforms any time series to a Gaussian series with zero mean and unit variance. We apply the Rosenblatt transformation to transform the obtained

TABLE IV
PERFORMANCE COMPARISONS OF SIX METHODS USING GAUSSIAN GARCH ON HSBC

Method	GARCH$(1,1)$				GARCH$(2,1)$			
	Log-likelihood	$Q(20)$	Skewness	Kurtosis	Log-likelihood	$Q(20)$	Skewness	Kurtosis
BHHH	-456.1237	6.3281	0.3851	6.1643	-461.0942	5.7586	0.4987	6.5563
BFGS	-451.2235	6.4127	0.3506	6.1831	-439.5896	6.4678	0.3458	6.3537
L-BFGS	-445.8975	6.4451	0.3473	6.3025	-440.9587	6.4554	0.3493	6.3492
SQP	-445.4572	6.4485	0.3468	6.3092	-437.0015	6.4830	0.3487	6.4003
IPM	-437.4983	6.5006	0.3621	6.5320	-442.2206	6.6693	0.3419	6.5413
BMM (prop.)	$\mathbf{-501.5057}$	**6.0364**	**0.1629**	**6.1011**	$\mathbf{-485.0711}$	**4.3559**	**0.0480**	**6.2379**

TABLE V
PERFORMANCE COMPARISONS OF SIX METHODS USING STUDENT'S t GARCH ON HSBC

Method	GARCH$(1,1)$				GARCH$(2,1)$			
	Log-likelihood	$Q(20)$	Skewness	Kurtosis	Log-likelihood	$Q(20)$	Skewness	Kurtosis
BHHH	-461.8468	5.9545	0.1847	3.7984	-479.6861	6.4385	0.1107	2.9463
BFGS	-444.9232	6.4245	0.1005	3.7421	-474.4392	6.1600	0.1436	3.4109
L-BFGS	-444.9228	6.4245	0.1005	3.7420	-431.8533	6.5713	0.0924	3.6846
SQP	-435.9413	6.4419	0.1680	4.5358	-424.9679	6.4872	0.2028	4.9519
IPM	-434.9634	6.4306	0.1545	4.4693	-442.0838	6.6285	0.1369	4.3235
BMM (prop.)	$\mathbf{-520.9695}$	**5.6841**	**0.0039**	**3.1216**	$\mathbf{-526.9228}$	**5.3784**	**0.0115**	**3.1179**

standardized residual series in Student's t GARCH into a Gaussian i.i.d. series, which is analogous to the standardized residuals from Gaussian GARCH models. Since Gaussian has a skewness of 0 and a kurtosis of 3, models producing series that closely match these skewness and kurtosis values tend to perform better.

The volatility forecasting performance from different estimation methods on the Gaussian GARCH models is shown in Table IV, and the performance for the Student's t GARCH models is presented in Table V (we chose $\ell = 20$ in the Ljung-Box test statistics as suggested by [79]). The proposed BMM method achieves lower objective function values than existing methods for both GARCH models, indicating a better fit to the data. Additionally, the BMM method demonstrates lower Ljung-Box Q-test statistics and better skewness and kurtosis values among all the models.

2) Target Detection in Radar: In radar target detection, the goal is to identify the desired targets from the received signals, which are often affected by unwanted clutter that obscures detection. Hence, accurately modeling the clutter is crucial for increasing the detection accuracy. Since the clutter observed in real-world data typically exhibits heavy-tailedness and volatility clustering, we can model it as a time series using a complex GARCH model [80]. In [22], [23], the Gaussian GARCH model was adopted for clutter modeling, while in this paper, we also consider the Student's t GARCH. A complex GARCH process ϵ_t in (1) is specified as [81]

$$\epsilon_t = \sqrt{h_t} z_t,$$

where the conditional variance h_t is defined as

$$h_t = \omega + \sum_{i=1}^{q} \alpha_i \left|\epsilon_{t-i}\right|^2 + \sum_{j=1}^{p} \beta_j h_{t-j},$$

and $\{z_t\}$ is a sequence of standardized circularly symmetric i.i.d. random variables. In the experiment, the clutter ϵ_t is modeled using the McMaster University IPIX radar dataset [82], where it has a conditional mean of zero. Specifically, we use the first 2048 samples from the file "19931107_141630_starea.cdf" in the IPIX radar dataset. The IPIX radar provides polarimetric information; the data shown corresponds to vertical polarization. We split the 2048 samples into two parts. The first 1024 samples are used as the training data. The last 1024 samples are used to construct the test data, where a synthetic target signal, denoted as e_t is added onto the clutter ϵ_t at specific time points. We estimate the GARCH models based on the training data. Then, the estimated model is used to detect whether the test data at each time point includes an echo signal from the target. We model e_t according to the Swerling I model, which assumes it follows a circular Gaussian distribution with zero mean and constant variance h_e [83].

The detection procedure at time t involves making decisions on the received signal between these two hypotheses:

$$H_0: \quad \text{target is absent}$$
$$H_1: \quad \text{target is present}$$

where in the null hypothesis H_0, the received signal is assumed to contain only the clutter ϵ_t; in the alternative hypothesis H_1, the received signal additionally includes the target e_t that we aim to detect. Let y_t denote the received signal. The decision between H_0 and H_1 is made using the Neyman-Pearson criterion [84]:

$$\Lambda(y_t) = \frac{P(y_t \mid H_1)}{P(y_t \mid H_0)} = \frac{\int P(y_t \mid e_t, H_1) P(e_t) \mathrm{d}e_t}{P(y_t \mid H_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \zeta_t,$$

where $\Lambda(y_t)$ is called the likelihood ratio, $P(\cdot \mid \cdot)$ denotes the probability of the corresponding event, and ζ_t is a nonnegative threshold. We choose H_1 if the likelihood ratio $\Lambda(y_t)$ is larger

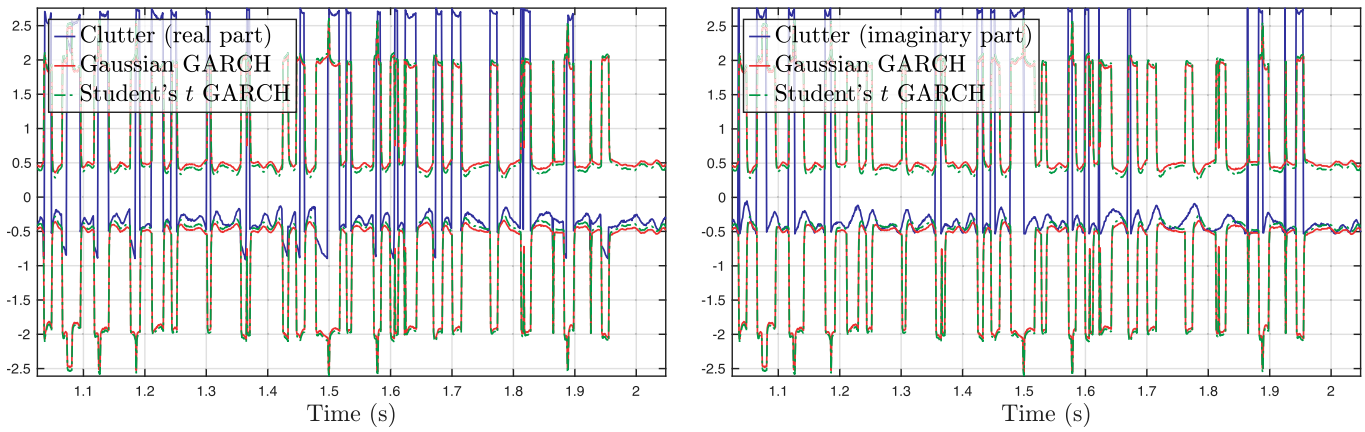


Fig. 3. The real and imaginary parts of the radar clutter signal, along with the estimated gaussian and student's t GARCH models.

than the threshold ζ_t, and we fail to reject H_0 if the likelihood ratio is smaller than ζ_t.

In practice, we are often given a probability of false alarm P_{fa}, which is defined as the probability of declaring a received signal under H_1 when it is actually under H_0. Then, we have the following relation:

$$P_{\text{fa}} = P\left(\Lambda(y_t) > \zeta_t \mid H_0\right). \tag{25}$$

The value of ζ_t can be calculated based on (25). When ϵ_t follows a complex Gaussian GARCH model, $\Lambda(y_t)$ is calculated by

$$\Lambda(y_t) = \sqrt{\frac{h_t}{h_t + h_e}} \exp\left(\frac{|y_t|^2}{h_t} - \frac{|y_t|^2}{h_t + h_e}\right).$$

Since $P\left(\Lambda(y_t) > \zeta_t \mid H_0\right) = P\left(\log \Lambda(y_t) > \log \zeta_t \mid H_0\right)$, we have

$$\begin{aligned} P_{\text{fa}} &= P\left(\frac{1}{2}\log\left(\frac{h_t}{h_t + h_e}\right) + \frac{|y_t|^2}{h_t} - \frac{|y_t|^2}{h_t + h_e} > \log \zeta_t \,\middle|\, H_0\right) \\ &= P\left(|y_t| > \sqrt{\frac{1}{2}\frac{h_t(h_t + h_e)}{h_e}\log\left(\frac{\zeta_t^{-2} h_t}{h_t + h_e}\right)} \,\middle|\, H_0\right) \\ &= P\left(|\epsilon_t| > \sqrt{\frac{1}{2}\frac{h_t(h_t + h_e)}{h_e}\log\left(\frac{\zeta_t^{-2} h_t}{h_t + h_e}\right)}\right). \end{aligned}$$

Since the modulus of a circular Gaussian random variable $|\epsilon_t|$ is a Rayleigh random variable, we can compute ζ_t in a closed form. However, when ϵ_t follows a complex Student's t GARCH model, a closed-form solution for ζ_t does not exist. In practice, for a given threshold ζ_t, we can use the Monte Carlo method (10^6 Monte Carlo iterations are used in the following calculation) to approximate the relation in (25) [85]:

$$\hat{P}_{\text{fa}} = \frac{1}{10^6}\sum_{i=1}^{10^6} I_{\{\Lambda(y_t) > \zeta_t | H_0\}}(\epsilon_t^i), \tag{26}$$

where $I_{\mathcal{A}}(\cdot)$ is the indicator function of set $\mathcal{A}$, and ϵ_t^i denotes the i-th Monte Carlo sample from ϵ_t. The approximation in (26) can be run for multiple different ζ_t. Therefore, for a given P_{fa}, the value of ζ_t can be approximately computed.

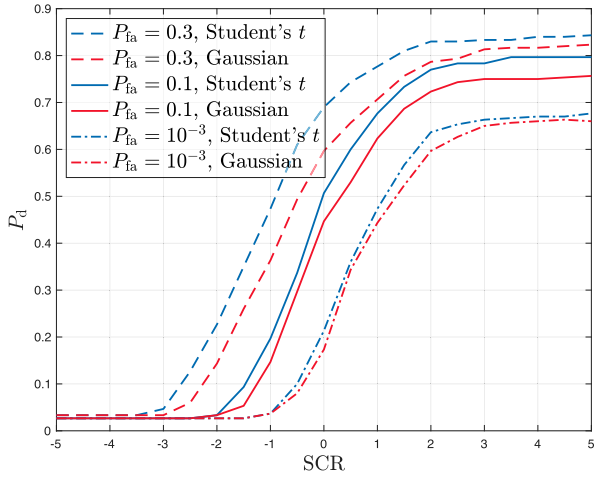


Fig. 4. P_{d} versus SCR for the Gaussian and Student's t GARCH models.

In the following, we compare the performance of Gaussian and Student's t GARCH models on clutter modeling. In Fig. 3, we demonstrate the clutter from the test data with the volatility envelopes that are forecasted by Gaussian GARCH$(1,1)$ and Student's t GARCH$(1,1)$ (orders of these models are determined based on BIC). We report the probability of detection P_{d}, defined as the probability of declaring a received signal under H_1 when it is indeed under H_1, under different values of signal-to-clutter ratio (SCR)[3] and P_{fa}. Fig. 4 illustrates how P_{d} changes with varying SCRs, while Fig. 5 presents the relationship between P_{d} and P_{fa}. From both figures, we can observe that P_{d} increases with the increase in SCR or P_{fa} under both GARCH models. Moreover, under the same SCR and P_{fa}, the Student's t GARCH model consistently demonstrates better performance than the Gaussian GARCH model.

[3]SCR is defined as $\text{SCR} = \log \frac{h_e}{\hat{h}}$, where $\hat{h}$ is the sample variance of ϵ_t in the test data.

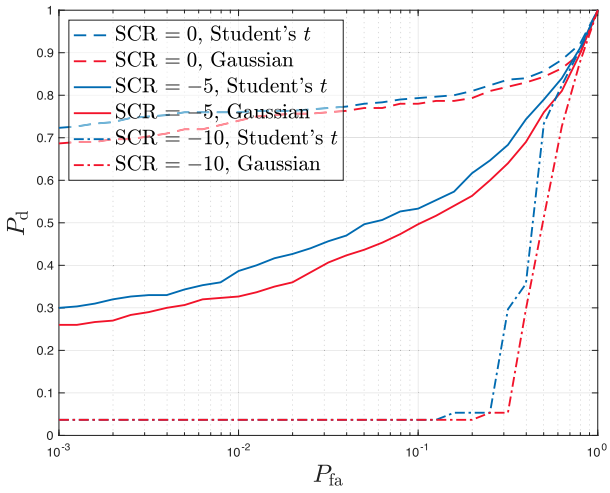


Fig. 5. $P_{\rm d}$ versus $P_{\rm fa}$ for the Gaussian and Student's t GARCH models.

IX. Conclusion

In this paper, we have studied the parameter estimation of GARCH models under the Gaussian and Student's t assumptions. We have proposed a novel penalty method for the ML estimation problems and have developed structure-aware block majorization-minimization algorithms for problem resolution. Compared with existing methods, our approach effectively exploits the structure of the ML estimation problems and avoids the heavy computational overhead of volatility recursion constraints, which often increase the risk of round-off errors and may cause convergence failures. Experiments on both synthetic and real data have demonstrated that the proposed algorithm achieves more accurate and efficient parameter estimation. Furthermore, we have discussed how the developed methods can be adapted to M-estimation of GARCH, estimation of GARCH variants, and joint estimation of GARCH and conditional mean models. Beyond GARCH, the proposed penalty approach for enforcing recursive coupling constraints can also be extended to ML estimation of stochastic volatility models, which we leave for future research.

Appendix A
Proof for Proposition 3

Proof: We represent each cubic equation in (16) as

$$g(x) = a_1x^3 + a_2x^2 + a_3x + a_4 = 0,$$

where $a_1 > 0$, $a_3 > 0$, $a_4 < 0$, and the sign of a_2 is undetermined. By Descartes' rule of signs, the number of positive solutions is related to the number of sign changes in the coefficients. Hence,

- If $a_2 \geq 0$, the signs change once, indicating the cubic equation in (16) has a unique positive solution;
- If $a_2 < 0$, there are three sign changes, indicating the cubic equation in (16) has one or three positive solutions.

In summary, each cubic equation in (16) has either one or three positive roots. Then, we prove that minimum points exist among the positive solutions by examining the monotonicity of $g(x)$. We have

$$g'(x) = 3a_1x^2 + 2a_2x + a_3.$$

Since both a_1 and a_3 are positive, $g'(x)$ cannot have only one positive zero. Hence, we consider the following two possible cases:

- If $g'(x)$ has no positive zero, then since the coefficient of order 2 is positive, $g'(x) \geq 0$, indicating that $g(x)$ is a monotonically increasing function for $x > 0$. Considering $\lim_{x\to 0} g(x) = a_4 < 0$, the unique zero for $g(x)$ is the minimum point;
- If $g'(x)$ has two positive zeros, then since the coefficient of order 2 is positive, $g(x)$ is increasing, decreasing, and then increasing for $x > 0$. Considering $\lim_{x\to 0} g(x) = a_4 < 0$, if $g(x)$ has a unique zero, it is the minimum point; if $g(x)$ has three zeros, the minimum is either the smallest or the largest zero.

■

References

[1] C. Gao, Z. Zhao, and D. P. Palomar, "A novel algorithm for GARCH model estimation," in *Proc. IEEE Statist. Signal Process. Workshop (SSP)*, 2023, pp. 210–214.
[2] R. S. Tsay, *Analysis of Financial Time Series*. Hoboken, NJ, USA: Wiley, 2005.
[3] D. S. Matteson and D. Ruppert, "Time-series models of dynamic volatility and correlation," *IEEE Signal Process. Mag.*, vol. 28, no. 5, pp. 72–82, Sep. 2011.
[4] R. F. Engle, "Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation," *Econometrica*, vol. 50, no. 4 pp. 987–1007, 1982.
[5] T. Bollerslev, "Generalized autoregressive conditional heteroskedasticity," *J. Econometrics*, vol. 31, no. 3, pp. 307–327, 1986.
[6] R. F. Engle and T. Bollerslev, "Modelling the persistence of conditional variances," *Econometrics Rev.*, vol. 5, no. 1, pp. 1–50, 1986.
[7] D. B. Nelson, "Conditional heteroskedasticity in asset returns: A new approach," *Econometrica*, vol. 59, no. 2 pp. 347–370, 1991.
[8] J.-M. Zakoïan, "Threshold heteroskedastic models," *J. Econ. Dyn. Control*, vol. 18, no. 5, pp. 931–955, 1994.
[9] L. R. Glosten, R. Jagannathan, and D. E. Runkle, "On the relation between the expected value and the volatility of the nominal excess return on stocks," *J. Finance*, vol. 48, no. 5, pp. 1779–1801, 1993.
[10] Z. Ding, C. W. Granger, and R. F. Engle, "A long memory property of stock market returns and a new model," *J. Empirical Finance*, vol. 1, no. 1, pp. 83–106, 1993.
[11] R. F. Engle, "GARCH 101: The use of ARCH/GARCH models in applied econometrics," *J. Econ. Perspectives*, vol. 15, no. 4, pp. 157–168, 2001.
[12] T. Bollerslev, "Glossary to ARCH (GARCH)," *CREATES Res. Paper*, vol. 49, p. 44, Sep. 2008.
[13] K. Johansson et al., "A simple method for predicting covariance matrices of financial returns," *Found. Trends Econometrics*, vol. 12, no. 4, pp. 324–407, 2023.
[14] T. Bollerslev, "A conditionally heteroskedastic time series model for speculative prices and rates of return," *Rev. Econ. Statist.*, vol. 69, no. 3, p. 542, 1987.
[15] R. Cont, "Empirical properties of asset returns: Stylized facts and statistical issues," *Quantitative Finance*, vol. 1, no. 2, p. 223, 2001.
[16] I. Cohen, "Modeling speech signals in the time–Frequency domain using GARCH," *Signal Process.*, vol. 84, no. 12, pp. 2453–2459, 2004.
[17] A. Abramson and I. Cohen, "Recursive supervised estimation of a Markov-switching GARCH process in the short-time Fourier transform domain," *IEEE Trans. Signal Process.*, vol. 55, no. 7, pp. 3227–3238, Jul. 2007.
[18] R. Tahmasbi and S. Rezaei, "A soft voice activity detection using GARCH filter and Variance Gamma distribution," *IEEE Trans. Audio Speech Language Process.*, vol. 15, no. 4, pp. 1129–1134, May 2007.

[19] R. Tahmasbi and S. Rezaei, "Change point detection in GARCH models for voice activity detection," *IEEE Trans. Audio Speech Language Process.*, vol. 16, no. 5, pp. 1038–1046, Jul. 2008.

[20] A. Abramson and I. Cohen, "Single-sensor audio source separation using classification and estimation approach and GARCH modeling," *IEEE Trans. Audio Speech Language Process.*, vol. 16, no. 8, pp. 1528–1540, Nov. 2008.

[21] S. Mousazadeh and I. Cohen, "AR-GARCH in presence of noise: Parameter estimation and its application to voice activity detection," *IEEE Trans. Audio Speech Language Process.*, vol. 19, no. 4, pp. 916–926, May 2011.

[22] J. P. Pascual, N. von Ellenrieder, M. Hurtado, and C. H. Muravchik, "Radar detection algorithm for GARCH clutter model," *Digit. Signal Process.*, vol. 23, no. 4, pp. 1255–1264, 2013.

[23] J. P. Pascual, N. von Ellenrieder, M. Hurtado, and C. H. Muravchik, "Adaptive radar detection algorithm based on an autoregressive GARCH-2D clutter model," *IEEE Trans. Signal Process.*, vol. 62, no. 15, pp. 3822–3832, Aug. 2014.

[24] A. Noiboar and I. Cohen, "Two-dimensional GARCH model with application to anomaly detection," in *Proc. Eur. Signal Process. Conf.*, 2005, pp. 1–4.

[25] A. Noiboar and I. Cohen, "Anomaly detection based on wavelet domain GARCH random field modeling," *IEEE Trans. Geosci. Remote Sens*, vol. 45, no. 5, pp. 1361–1373, May 2007.

[26] H. T. Pham and B. Yang, "Estimation and forecasting of machine health condition using ARMA/GARCH model," *Mech. Syst. Signal Process.*, vol. 24, no. 2, pp. 546–558, 2010.

[27] A. Eamaz, F. Yeganegi, and M. Soltanalian, "Covariance recovery for one-bit sampled non-stationary signals with time-varying sampling thresholds," *IEEE Trans. Signal Process.*, vol. 70, pp. 5222–5236, 2022.

[28] Y. Wang, N. Routledge, Y. Zhao, and D. Zhang, "Online muscle activation onset detection using likelihood of conditional heteroskedasticity of electromyography signals," *IEEE Trans. Biomed. Eng.*, vol. 71, no. 5, pp. 1663–1676, May 2024.

[29] M. Amirmazlaghani, H. Amindavar, and A. Moghaddamjoo, "Speckle suppression in SAR images using the 2-D GARCH model," *IEEE Trans. Image Process.*, vol. 18, no. 2, pp. 250–259, Feb. 2009.

[30] H. Kalbkhani, "Robust algorithm for brain magnetic resonance image (MRI) classification based on GARCH variances series," *Biomed. Signal Process. Control*, vol. 8, no. 6, p. 11, 2013.

[31] J. Hong, Y. Yan, E. E. Kuruoglu, and W. K. Chan, "Multivariate time series forecasting with GARCH models on graphs," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 9, pp. 557–568, 2023.

[32] L. Wu and M. Shahidehpour, "A hybrid model for day-ahead price forecasting," *IEEE Trans. Power Syst.*, vol. 25, no. 3, pp. 1519–1530, Aug. 2010.

[33] L. M. Lima, P. Damien, and D. W. Bunn, "Bayesian predictive distributions for imbalance prices with time-varying factor impacts," *IEEE Trans. Power Syst.*, vol. 38, no. 1, pp. 349–357, Jan. 2023.

[34] M. Yu, J. Han, H. Wu, J. Yan, and R. Zeng, "Short-term wind power prediction based on wind2vec-BERT model," *IEEE Trans. Sustain. Energy*, vol. 16, no. 2, pp. 933–944, Apr. 2025.

[35] D. D. Wu, L. Zheng, and D. L. Olson, "A decision support approach for online stock forum sentiment analysis," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 44, no. 8, pp. 1077–1087, Aug. 2014.

[36] Y. Zhang, A. Haghani, and X. Zeng, "Component GARCH models to account for seasonal patterns and uncertainties in travel-time prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 2, pp. 719–729, Apr. 2015.

[37] Q. T. Tran, Z. Ma, H. Li, L. Hao, and Q. K. Trinh, "A multiplicative seasonal ARIMA/GARCH model in EVN traffic prediction," *Int. J. Comput. Sci. Netw. Secur.*, vol. 08, no. 4, pp. 43–49, 2015.

[38] C. Ding, J. Duan, Y. Zhang, X. Wu, and G. Yu, "Using an ARIMA-GARCH modeling approach to improve subway short-term ridership forecasting accounting for dynamic volatility," *IEEE Trans. Intell. Transport. Syst.*, vol. 19, no. 4, pp. 1054–1064, Apr. 2018.

[39] J. Skoglund, "A simple efficient GMM estimator of GARCH models," SSE/EFI Working Paper Series in Economics Finance, 2001. [Online]. Available: https://ssrn.com/abstract=2241031.

[40] P. Arie and S. Giuseppe, "Least-squares estimation for GARCH$(1,1)$ model with heavy tailed errors," Tech. Rep. Feb. 2017. [Online]. Available: https://mpra.ub.uni-muenchen.de/59082/1/MPRA_paper_59082.pdf.

[41] G. Cornuejols, J. Peña, and R. Tütüncü, *Optimization Methods in Finance.* Cambridge, U.K.: Cambridge Univ. Press, 2018.

[42] P. Hall and Q. Yao, "Inference in ARCH and GARCH models with heavy–Tailed errors," *Econometrica*, vol. 71, no. 1, pp. 285–317, 2003.

[43] L. Peng, and Q. Yao, "Least absolute deviations estimation for ARCH and GARCH models," *Biometrika*, vol. 90, no. 4, pp. 967–975, 2003.

[44] G. Calzolari, R. Halbleib, and A. Parrini, "Estimating GARCH-type models with symmetric stable innovations: Indirect inference versus maximum likelihood," *Comput. Statist. Data Anal.*, vol. 76, pp. 158–171, Aug. 2014.

[45] K. L. Lange, R. J. Little, and J. M. Taylor, "Robust statistical modeling using the t distribution," *J. Amer. Statist. Assoc.*, vol. 84, no. 408, pp. 881–896, 1989.

[46] D. Ardia and L. F. Hoogerheide, "Bayesian estimation of the GARCH$(1,1)$ model with Student-t innovations," *R J.*, vol. 2, no. 2, p. 41, 2010.

[47] T. Adrian and H. Kurt, "Tseries: Time series analysis and computational finance," 2020, R Package Version 0, pp. 10–48. [Online]. Available: https://cran.rproject.org/web/packages/tseries/tseries.pdf.

[48] W. Diethelm, "fGarch: Rmetrics - autoregressive conditional heteroskedastic modelling," R Package Version 3042.83.2, 2017. [Online]. Available: https://cran.r-project.org/web/packages/fGarch/fGarch.pdf.

[49] A. Ghalanos, "rugarch: Univariate GARCH models," R Package Version 1.4-4, 2020. [Online]. Available: https://cran.rproject.org/web/packages/rugarch/rugarch.pdf.

[50] K. Sheppard, "Bashtage/Arch," Python Package Version v4.18, 2021. [Online]. Available: https://pypi.org/project/arch/.

[51] "MATLAB econometrics toolbox," MathWorks, Natick, MA, USA, 2022. [Online]. Available: https://ww2.mathworks.cn/help/econ/index.html.

[52] D. B. Nelson, "Stationarity and persistence in the GARCH$(1,1)$ model," *Econ. Theory*, vol. 6, no. 3, pp. 318–334, 1990.

[53] S. T. Jensen and A. Rahbek, "Asymptotic inference for nonstationary GARCH," *Econ. Theory*, vol. 20, no. 6, pp. 1203–1226, 2004.

[54] C. Francq and J. Zakoïan, *GARCH Models: Structure, Statistical Inference and Financial Applications.* Hoboken, NJ, USA: Wiley, 2019.

[55] Y. Sun, P. Babu, and D. P. Palomar, "Majorization-minimization algorithms in signal processing, communications, and machine learning," *IEEE Trans. Signal Process.*, vol. 65, no. 3, pp. 794–816, Feb. 2017.

[56] R. F. Engle, D. M. Lilien, and R. P. Robins, "Estimating time varying risk premia in the term structure: The ARCH-M model," *Econometrica*, vol. 55, no. 2, pp. 391–407, 1987.

[57] T. Bollerslev, R. F. Engle, and J. M. Wooldridge, "A capital asset pricing model with time-varying covariances," *J. Political Econ.*, vol. 96, no. 1, pp. 116–131, 1988.

[58] N. Chan and C. Ng, "Statistical inference for non-stationary GARCH(p, q) models," *Electron. J. Statist.*, vol. 3, pp. 956–992, 2009.

[59] C. Francq and J.-M. Zakoïan, "Strict stationarity testing and estimation of explosive and stationary generalized autoregressive conditional heteroscedasticity models," *Econometrica*, vol. 80, no. 2, pp. 821–861, 2012.

[60] D. P. Palomar, "Convex primal decomposition for multicarrier linear MIMO transceivers," *IEEE Trans. Signal Process.*, vol. 53, no. 12, pp. 4661–4674, Dec. 2005.

[61] C. Liu and D. B. Rubin, "ML estimation of the t distribution using EM and its extensions, ECM and ECME," *Statistica Sinica*, vol. 5, no. 1, pp. 19–39, 1995.

[62] R. P. Brent, *Algorithms for Minimization without Derivatives.* North Chelmsford, MA, USA: Courier Corporation, 2013.

[63] M. Razaviyayn, M. Hong, and Z. Luo, "A unified convergence analysis of block successive minimization methods for nonsmooth optimization," *SIAM J. Optim.*, vol. 23, no. 2, pp. 1126–1153, 2013.

[64] R. Varadhan and C. Roland, "Simple and globally convergent methods for accelerating the convergence of any EM algorithm," *Scand. J. Statist.*, vol. 35, no. 2, pp. 335–353, 2008.

[65] K. Prause et al., "The generalized hyperbolic model: Estimation, financial derivatives, and risk measures," Ph.D. dissertation, Univ. of Freiburg, 1999.

[66] K. Mukherjee, "M-estimation in GARCH models," *Econ. Theory*, vol. 24, no. 6, pp. 1530–1553, 2008.

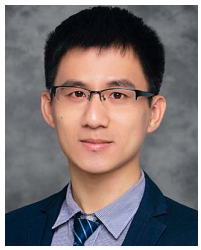
[67] K. Mukherjee, "A review of robust estimation under conditional heteroscedasticity," in *Handbook Statist.*, vol. 30, pp. 123–154, 2012.

[68] A. Panorska, S. Mittnik, and S. Rachev, "Stable GARCH models for financial time series," *Appl. Math. Lett.*, vol. 8, no. 5, pp. 33–37, 1995.

[69] A. Harvey and T. Chakravarty, "Beta-t-(E)GARCH," Faculty of Economics, University of Cambridge, Cambridge Working Papers in Economics, 2008. [Online]. Available: https://files.econ.cam.ac.uk/repec/cam/pdf/cwpe0840.pdf

[70] A. C. Harvey, Dynamic models for volatility and heavy tails: with applications to financial and economic time series, vol. 52, Cambridge, U.K.: Cambridge University Press, 2013.
[71] D. Creal, S. J. Koopman, and A. Lucas, "Generalized autoregressive score models with applications," *J. Appl. Econom.*, vol. 28, no. 5, pp. 777–795, 2013.
[72] T. Bollerslev and J. M. Wooldridge, "Quasi-maximum likelihood estimation and inference in dynamic models with time-varying covariances," *Econometric Rev.*, vol. 11, no. 2, pp. 143–172, 1992.
[73] C. Francq and J.-M. Zakoïan, "Maximum likelihood estimation of pure GARCH and ARMA-GARCH processes," *Bernoulli*, vol. 10, no. 4, pp. 605–637, 2004.
[74] J. Fan, L. Qi, and D. Xiu, "Quasi-maximum likelihood estimation of GARCH models with heavy-tailed likelihoods," *J. Bus. Econ. Statist.*, vol. 32, no. 2, pp. 178–191, 2014.
[75] L. Hentschel, "All in the family nesting symmetric and asymmetric GARCH models," *J. Financial Econ.*, vol. 39, no. 1, pp. 71–104, 1995.
[76] F. Pérez-Cruz, J. A. Afonso-Rodríguez, and J. Giner, "Estimating GARCH models using support vector machines," *Quant. Finance*, vol. 3, no. 3, pp. 163–172, 2003.
[77] M. Rosenblatt, "Remarks on a multivariate transformation," *Ann. Math. Statist.*, vol. 23, no. 3, pp. 470–472, 1952.
[78] S. Kim, N. Shepherd, and S. Chib, "Stochastic volatility: Likelihood inference and comparison with ARCH models," *Rev. Econ. Stud.*, vol. 65, no. 3, pp. 361–393, 1998.
[79] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time Series Analysis: Forecasting and Control.* Hoboken, NJ, USA: Wiley, 2015.
[80] I. Cohen, "Speech spectral modeling and enhancement based on autoregressive conditional heteroscedasticity models," *Signal Process.*, vol. 86, no. 4, pp. 698–709, 2006.
[81] S. Mousazadeh and I. Cohen, "Simultaneous parameter estimation and state smoothing of complex GARCH process in the presence of additive noise," *Signal Process.*, vol. 90, no. 11, pp. 2947–2953, 2010.
[82] S. Haykin, C. Krasnor, T. J. Nohara, B. W. Currie, and D. Hamburger, "A coherent dual-polarized radar for studying the ocean environment," *IEEE Trans. Geosci. Remote Sens.*, vol. 29, no. 1, pp. 189–191, Jan. 1991.
[83] M. A. Richards et al., *Fundamentals of Radar Signal Processing*, vol. 1. New York, NY, USA: McGraw-Hill, 2005.
[84] H. L. Van Trees, *Detection, Estimation, and Modulation Theory.* Hoboken, NJ, USA: Wiley, 2004.
[85] P. Hughes, "A high-resolution radar detection strategy," *IEEE Trans. Aerosp. Electron. Syst.*, no. 5, pp. 663–667, Sep. 1983.
[86] H. Alzer, "On some inequalities for the Gamma and Psi functions," *Math. Comput.*, vol. 66, no. 217, pp. 373–389, 1997.
[87] G. Nemes, "Error bounds for the asymptotic expansion of the Hurwitz zeta function," *Proc. R. Soc. A, Math. Phys. Eng. Sci.*, vol. 473, no. 2203, 2017, Art. no. 20170363.

Chenyu Gao (Graduate Student Member, IEEE) received the B.Sc. degree in Mathematics and Applied Mathematics from Xidian University, Xi'an, China, in 2020, and he is currently working toward the Ph.D. degree in Computer Science and Technology at ShanghaiTech University, Shanghai, China. His research interests include MM/EM algorithms, location-scale models, volatility models, and quantization.



Ziping Zhao (Member, IEEE) received the B.Eng. degree in Electronics and Information Engineering (with highest honor) from the Huazhong University of Science and Technology (HUST), Wuhan, China, in 2014, and the Ph.D. degree in Electronic and Computer Engineering from the Hong Kong University of Science and Technology (HKUST), Hong Kong, in 2019. He has also held several visiting research positions with the University of Minnesota, Twin Cities, MN, USA, and the Hong Kong University of Science and Technology, Hong Kong. Since 2019, he has been an Assistant Professor with the School of Information Science and Technology, ShanghaiTech University, Shanghai, China. His research interests include optimization and statistics with applications in signal processing and machine learning. His work was the recipient of the Best Student Paper Award from IEEE ICASSP 2023 and was a finalist for the Best Student Paper Award from IEEE SAM 2020, IEEE SPAWC 2021, and IEEE SAM 2024. He was the recipient of Hong Kong PhD Fellowship. He is a member of EURASIP and APSIPA.

Daniel P. Palomar (Fellow, IEEE) received the Electrical Engineering and Ph.D. degrees from the Technical University of Catalonia (UPC), Barcelona, Spain, in 1998 and 2003, respectively, and was a Fulbright Scholar with Princeton University from 2004 to 2006. He is a Professor with the Department of Electronic & Computer Engineering at the Hong Kong University of Science and Technology (HKUST), Hong Kong, which he joined in 2006. He had previously held several research appointments, namely, King's College London (KCL), London, UK; Stanford University, Stanford, CA; Telecommunications Technological Center of Catalonia (CTTC), Barcelona, Spain; Royal Institute of Technology (KTH), Stockholm, Sweden; University of Rome "La Sapienza", Rome, Italy; and Princeton University, Princeton, NJ. His current research interests include data analytics, optimization methods, and deep learning in financial systems. Dr. Palomar has been recognized as a EURASIP Fellow (2024), an IEEE Fellow (2012), and, among others, has been awarded with the 2004/06 Fulbright Research Fellowship and the 2004, 2015, and 2020 Young Author Best Paper Awards by the IEEE Signal Processing Society. He is the author of many research articles and books, including *Portfolio Optimization: Theory and Application* and *Convex Optimization in Signal Processing and Communications*. He has been a Guest Editor of the *IEEE Journal of Selected Topics in Signal Processing* 2016 Special Issue on *Financial Signal Processing and Machine Learning for Electronic Trading*, an Associate Editor of IEEE TRANSACTIONS ON INFORMATION THEORY and of IEEE TRANSACTIONS ON SIGNAL PROCESSING, a Guest Editor of the *IEEE Signal Processing Magazine* 2010 Special Issue on *Convex Optimization for Signal Processing*, the *IEEE Journal on Selected Areas in Communications* 2008 Special Issue on *Game Theory in Communication Systems*, and the *IEEE Journal on Selected Areas in Communications* 2007 Special Issue on *Optimization of MIMO Transceivers for Realistic Communication Networks*.

Supplementary Material to "Maximum Likelihood Estimation of Gaussian and Student's t GARCH: A Unified Penalty Method"

Chenyu Gao, *Graduate Student Member, IEEE*, Ziping Zhao, *Member, IEEE*, and Daniel P. Palomar, *Fellow, IEEE*

Appendix B
Proof for Theorem 4

Proof: We rewrite the objective function of problem (21) as $\Upsilon(\nu) = \sum_{t=1}^{n} \Upsilon_t(\nu)$, where $\Upsilon_t(\nu)$ is defined as the log-likelihood function for one sample with respect to the variable ν:

$$\Upsilon_t(\nu) = -\nu \log(\nu-2) + 2\log\Gamma\left(\frac{\nu}{2}\right) - 2\log\Gamma\left(\frac{\nu+1}{2}\right) + (\nu+1)\log(\nu-2+w_t),$$

with $w_t \geq 0$. Then, we have $\Upsilon'(\nu) = \sum_{t=1}^{n} \Upsilon_t'(\nu)$, where

$$\Upsilon_t'(\nu) = -\log(\nu-2) - \frac{\nu}{\nu-2} + \psi\left(\frac{\nu}{2}\right) - \psi\left(\frac{\nu+1}{2}\right) + \log(\nu-2+w_t) + \frac{\nu+1}{\nu-2+w_t}.$$

We analyze the number of zeros of $\Upsilon'(\nu)$ by examining those of $\Upsilon_t'(\nu)$. We first characterize the behavior of $\Upsilon_t'(\nu)$ at the two extremes of the domain of ν, namely $\nu \to 2$ and $\nu \to \infty$, which are given respectively by

$$\lim_{\nu\to 2} \Upsilon_t'(\nu) = -\lim_{\nu\to 2}\left(\log(\nu-2) + \frac{2}{\nu-2}\right) + \psi(1) - \psi\left(\frac{3}{2}\right) + \log w_t + \frac{3}{w_t} = -\infty, \tag{27}$$

and

$$\begin{aligned}\lim_{\nu\to\infty} \Upsilon_t'(\nu) &= \lim_{\nu\to\infty}\left(\log\left(\frac{\nu-2+w_t}{\nu-2}\right) + \psi\left(\frac{\nu}{2}\right) - \psi\left(\frac{\nu+1}{2}\right)\right) \\ &= \lim_{\nu\to\infty}\left(\psi\left(\frac{\nu}{2}\right) - \psi\left(\frac{\nu+1}{2}\right)\right).\end{aligned} \tag{28}$$

To calculate the limits related to the digamma function in (28), we introduce the following result.

Lemma 8 ([86]). *For the digamma function ψ, we have*

$$\log x - \frac{1}{x} \leq \psi(x) \leq \log x - \frac{1}{2x}.$$

By Lemma 8 and the squeeze theorem, we obtain $\lim_{x\to\infty} \psi(x) = \lim_{x\to\infty} \log x$. Hence,

$$\lim_{\nu\to\infty} \Upsilon_t'(\nu) = \lim_{\nu\to\infty} \log\left(\frac{\nu}{\nu+1}\right) = 0. \tag{29}$$

Based on the results in (27) and (29), if $\Upsilon_t'(\nu)$ is monotonically increasing on $(2,\infty)$, it has no zeros; if $\Upsilon_t'(\nu)$ is monotonically decreasing on (ν_t,∞) for some $\nu_t > 2$, it has at least one zero. Hence, we need to describe the monotonicity of $\Upsilon_t'(\nu)$ for $\nu \in (2,\infty)$. Consider the second derivative of Υ_t as follows:

$$\Upsilon_t''(\nu) = -\frac{1}{\nu-2} + \frac{2}{(\nu-2)^2} + \frac{1}{2}\psi'\left(\frac{\nu}{2}\right) - \frac{1}{2}\psi'\left(\frac{\nu+1}{2}\right) + \frac{1}{\nu-2+w_t} + \frac{w_t-3}{(\nu-2+w_t)^2},$$

where $\psi'(x) = -\int_0^1 \frac{y^{x-1}\log y}{1-y}\mathrm{d}y$ denotes the trigamma function. Since the value of ψ' cannot be analyzed directly, we introduce the following result.

Lemma 9 ([87]). *For the trigamma function ψ', we have*

$$\psi'(x) = \frac{1}{x} + \frac{1}{2x^2} + \frac{1}{6x^3} - \frac{1}{30x^5} + O\left(\frac{1}{x^7}\right).$$

Define

$$\psi'_{\text{lb}}(x) = \frac{1}{x} + \frac{1}{2x^2} + \frac{1}{6x^3} - \frac{1}{30x^5},$$

and

$$\psi'_{\text{ub}}(x) = \frac{1}{x} + \frac{1}{2x^2} + \frac{1}{6x^3}.$$

Based on Lemma 9, we have

$$\psi'_{\text{lb}}(x) < \psi'(x) < \psi'_{\text{ub}}(x). \tag{30}$$

We first find the condition for $\Upsilon'_t(\nu)$ being monotonically increasing, i.e., $\Upsilon''_t(\nu) \geq 0$. By (30), we obtain a lower bound of $\Upsilon''_t(\nu)$ as follows:

$$\begin{aligned}
\Upsilon''_t(\nu) &> -\frac{1}{\nu-2} + \frac{2}{(\nu-2)^2} + \frac{1}{2}\psi'_{\text{lb}}\left(\frac{\nu}{2}\right) - \frac{1}{2}\psi'_{\text{ub}}\left(\frac{\nu+1}{2}\right) + \frac{1}{\nu-2+w_t} + \frac{w_t-3}{(\nu-2+w_t)^2} \\
&= -\frac{1}{\nu-2} + \frac{2}{(\nu-2)^2} + \frac{1}{\nu} + \frac{1}{\nu^2} + \frac{2}{3\nu^3} - \frac{8}{15\nu^5} - \frac{1}{\nu+1} - \frac{1}{(\nu+1)^2} - \frac{2}{3(\nu+1)^3} + \frac{1}{\nu-2+w_t} + \frac{w_t-3}{(\nu-2+w_t)^2} \\
&= \frac{\chi_2(\nu-2)}{\chi_1(\nu-2)},
\end{aligned}$$

where χ_1 and χ_2 are two relatively prime polynomial functions, defined respectively as

$$\chi_1(x) = 15x^2(x+w_t)^2(x+2)^5(x+3)^3,$$

and

$$\begin{aligned}
\chi_2(x) = {} & (-15w_t^2 + 90w_t - 45)x^9 \\
& + (-240w_t^2 + 1620w_t - 765)x^8 \\
& + (-1545w_t^2 + 12600w_t - 5498)x^7 \\
& + (-4755w_t^2 + 55334w_t - 21482)x^6 \\
& + (4583w_t^2 + 150536w_t - 48956)x^5 \\
& + (15388w_t^2 + 261368w_t - 64616)x^4 \\
& + (61564w_t^2 + 285488w_t - 45360)x^3 \\
& + (97384w_t^2 + 181440w_t - 12960)x^2 \\
& + (77760w_t^2 + 51840w_t)x + 25920w_t^2.
\end{aligned}$$

Since $\chi_1(\nu-2) > 0$, $\Upsilon''_t(\nu) \geq 0$ implies $\chi_2(\nu-2) \geq 0$. It is sufficient to require all the coefficients in the polynomial χ_2 to be nonnegative, which is achieved when $w_t \in [3-\sqrt{6}, 3+\sqrt{6}]$. It follows that $\Upsilon''(\nu) = \sum_{t=1}^{n} \Upsilon''_t(v) \geq 0$ if $w_t \in [3-\sqrt{6}, 3+\sqrt{6}]$ for $t = 1, \ldots, n$. Then, $\Upsilon'(\nu)$ increases monotonically from $-\infty$ to 0 and therefore has no zeros on $(2, \infty)$.

Then, we study the case where $\Upsilon'_t(\nu)$ is monotonically decreasing, i.e., $\Upsilon''_t(\nu) < 0$ on (ν_t, ∞) for some $\nu_t > 2$. By (30), we get an upper bound of $\Upsilon''_t(\nu)$ as follows:

$$\begin{aligned}
\Upsilon''_t(\nu) &< -\frac{1}{\nu-2} + \frac{2}{(\nu-2)^2} + \frac{1}{2}\psi'_{\text{ub}}\left(\frac{\nu}{2}\right) - \frac{1}{2}\psi'_{\text{lb}}\left(\frac{\nu+1}{2}\right) + \frac{1}{\nu-2+w_t} + \frac{w_t-3}{(\nu-2+w_t)^2} \\
&= -\frac{1}{\nu-2} + \frac{2}{(\nu-2)^2} + \frac{1}{\nu} + \frac{1}{\nu^2} + \frac{2}{3\nu^3} - \frac{1}{\nu+1} - \frac{1}{(\nu+1)^2} - \frac{2}{3(\nu+1)^3} + \frac{8}{15(\nu+1)^5} + \frac{1}{\nu-2+w_t} + \frac{w_t-3}{(\nu-2+w_t)^2} \\
&= \frac{\chi_3(\nu-2)}{\chi_1(\nu-2)},
\end{aligned}$$

where χ_3 is a relatively prime polynomial, defined as

$$\begin{aligned}
\chi_3(x) = {} & (-15w_t^2 + 90w_t - 45)x^9 \\
& + (-270w_t^2 + 1800w_t - 855)x^8 \\
& + (-1980w_t^2 + 15570w_t - 6877)x^7 \\
& + (-7185w_t^2 + 76066w_t - 30227)x^6 \\
& + (10342w_t^2 + 229796w_t - 77874)x^5 \\
& + (15403w_t^2 + 441222w_t - 116486)x^4 \\
& + (93441w_t^2 + 530048w_t - 92340)x^3 \\
& + (172684w_t^2 + 369360w_t - 29160)x^2 \\
& + (155520w_t^2 + 116640w_t)x + 58320w_t^2.
\end{aligned}$$

Since $\chi_1(\nu-2) > 0$, to have $\Upsilon_t''(\nu) < 0$ on (ν_t, ∞) for some $\nu_t > 2$, it is sufficient to require the highest order of $\chi_3(\nu)$ is negative which is achieved when $w_t \in [0, 3-\sqrt{6}) \cup (3+\sqrt{6}, \infty)$. It follows that $\Upsilon''(\nu) = \sum_{t=1}^{n} \Upsilon_t''(\nu) < 0$ on $(\max_{t=1,\ldots,n} \nu_t, \infty)$ if $w_t \in [0, 3-\sqrt{6}) \cup (3+\sqrt{6}, \infty)$ for $t = 1, \ldots, n$. Then, $\Upsilon'(\nu)$ decreases monotonically to 0 on $(\max_{t=1,\ldots,n} \nu_t, \infty)$, and hence it has at least one zero on $(2, \infty)$. ■

APPENDIX C
PROOF FOR LEMMA 5

The second derivative of Υ_1 is given by

$$\Upsilon_1''(\nu) = \frac{n}{2}\psi'\left(\frac{\nu}{2}\right) - \frac{n}{2}\psi'\left(\frac{\nu+1}{2}\right) + \frac{n}{\nu+1} - \frac{n}{\nu-2} + \frac{2n}{(\nu-2)^2}.$$

By (30), we have

$$\begin{aligned}\Upsilon_1''(\nu) &> \frac{n}{2}\psi'_{\text{lb}}\left(\frac{\nu}{2}\right) - \frac{n}{2}\psi'_{\text{ub}}\left(\frac{\nu+1}{2}\right) + \frac{n}{\nu+1} - \frac{n}{\nu-2} + \frac{2n}{(\nu-2)^2} \\ &= n\frac{16\nu^4 + 29\nu^3 + 19\nu^2 + 36\nu + 12}{3\nu^2(\nu+1)^3(\nu-2)^2} \\ &> 0.\end{aligned}$$

We can verify $\Upsilon_1'' > 0$ on $(2, \infty)$, implying Υ_1 is strictly convex.

The second derivative of Υ_2 is computed as

$$\Upsilon_2''(\nu) = -\sum_{t=1}^{n} \frac{(w_t-3)^2}{(\nu+1)(\nu-2+w_t)} \leq 0.$$

We can obtain that $\Upsilon_2''(\nu) \leq 0$ on $(2, \infty)$, i.e., $\Upsilon_2(\nu)$ is concave.

APPENDIX D
PROOF FOR PROPOSITION 6

Proof: We first analyze the two limits of $\bar{\Upsilon}'(\nu)$. When ν approaches 2, we have

$$\begin{aligned}\lim_{\nu\to 2} \bar{\Upsilon}'(\nu) &= \lim_{\nu\to 2}\left(-n\log(\nu-2) - n\frac{2}{\nu-2}\right) + n\psi(1) - n\psi\left(\frac{3}{2}\right) + n\log 3 + n + \sum_{t=1}^{n}\left[\log\left(\frac{\underline{\nu}-2+w_t}{\underline{\nu}+1}\right) + \frac{3-w_t}{\underline{\nu}-2+w_t}\right] \\ &= -\infty.\end{aligned}$$

When ν approaches ∞, we have

$$\begin{aligned}\lim_{\nu\to\infty} \bar{\Upsilon}'(\nu) &= \lim_{\nu\to\infty}\left(n\log\left(\frac{\nu+1}{\nu-2}\right) - n\frac{\nu}{\nu-2} + n + n\psi\left(\frac{\nu}{2}\right) - n\psi\left(\frac{\nu+1}{2}\right)\right) + \sum_{t=1}^{n}\left[\log\frac{\underline{\nu}-2+w_t}{\underline{\nu}+1} + \sum_{t=1}^{n}\frac{3-w_t}{\underline{\nu}-2+w_t}\right] \\ &= \lim_{\nu\to\infty} n\log\left(\frac{\nu}{\nu+1}\right) + \sum_{t=1}^{n}\left[\log\left(\frac{\underline{\nu}-2+w_t}{\underline{\nu}+1}\right) + \frac{3-w_t}{\underline{\nu}-2+w_t}\right] \\ &= \sum_{t=1}^{n}\left[\log\left(\frac{\underline{\nu}-2+w_t}{\underline{\nu}+1}\right) + \frac{3-w_t}{\underline{\nu}-2+w_t}\right],\end{aligned}$$

where the second equation is due to Lemma 8. It is easy to verify that

$$\lim_{\nu\to\infty} \bar{\Upsilon}'(\nu) \begin{cases} = 0, & w_t = 3, \text{ for } t = 1, \ldots, n \\ > 0, & \text{otherwise.}\end{cases}$$

Since $\bar{\Upsilon}$ is strictly convex on $(2, \infty)$, $\bar{\Upsilon}'$ is strictly monotonically increasing on $(2, \infty)$. Hence, if $w_t = 3$ for $t = 1, \ldots, n$, i.e., $\lim_{\nu\to\infty} \bar{\Upsilon}'(\nu) = 0$, then $\bar{\Upsilon}'(\nu)$ has no zeros; if there exists $w_t \neq 3$ for $t = 1, \ldots, n$, i.e., $\lim_{\nu\to\infty} \bar{\Upsilon}'(\nu) > 0$, then $\bar{\Upsilon}'(\nu)$ has a unique zero. ■

APPENDIX E
ESTIMATION OF GARCH VARIANTS

A. Threshold GARCH

To describe the asymmetric clustering effect in volatility, the threshold GARCH [8] was proposed as an extension of the vanilla GARCH, with the volatility equation given by

$$h_t = \omega + \sum_{i=1}^{q}\left(\alpha_i\epsilon_{t-i}^2 + \xi_i I(\epsilon_{t-i} < 0)\epsilon_{t-i}^2\right) + \sum_{j=1}^{p}\beta_j h_{t-j}, \tag{31}$$

where the coefficients ξ_i for $i = 1, \ldots, q$ capture the asymmetric response of volatility to past negative shocks and $I(\cdot)$ denotes the indicator function. The threshold GARCH model is also known as the GJR-GARCH model, as an essentially equivalent formulation was independently introduced by Glosten, Jagannathan, and Runkle in [9]. For the model to be valid, in addition to the condition in (5), it is also required that

$$\xi_i \geq 0, \quad \text{for } i = 1, \ldots, q.$$

Define

$$\boldsymbol{\gamma} = [\alpha_1, \ldots, \alpha_q, \xi_1, \ldots, \xi_q, \beta_1, \ldots, \beta_p]^\top.$$

Similar to the GARCH setting, to ensure strict stationarity with a finite unconditional variance, the following stationarity condition is imposed:

$$\boldsymbol{\gamma}^\top \mathbf{1} < 1.$$

In contrast, when

$$\boldsymbol{\gamma}^\top \mathbf{1} = 1,$$

the process remains strictly stationary but has an infinite unconditional variance. Define

$$\boldsymbol{c}_t = \left[\epsilon_{t-1}^2, \ldots, \epsilon_{t-q}^2, I(\epsilon_{t-1})\epsilon_{t-1}^2, \ldots, I(\epsilon_{t-q})\epsilon_{t-q}^2, h_{t-1}, \ldots, h_{t-p}\right]^\top.$$

The volatility equation in (31) can then be expressed compactly as

$$h_t = \omega + \boldsymbol{\gamma}^\top \boldsymbol{c}_t.$$

Since both the constraint structure and the form of the volatility equation are analogous to those in the GARCH model, the threshold GARCH model can be estimated in a similar manner using the proposed BMM algorithm.

B. GARCH with Exogenous Covariates

The generalized autoregressive conditional heteroskedasticity model with exogenous covariates (GARCHX) model extends the standard GARCH framework by allowing external information to enter the conditional variance equation directly. The volatility equation of a GARCHX(q, p, d) model is given by

$$h_t = \omega + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j h_{t-j} + \sum_{k=1}^{d} \theta_k \varsigma_{t-k},$$

where $\varsigma_{t-1}, \ldots, \varsigma_{t-d}$ denote observable exogenous variables and $\theta_1, \ldots, \theta_d$ are the corresponding coefficients. Define

$$\boldsymbol{\gamma} = [\alpha_1, \ldots, \alpha_q, \beta_1, \ldots, \beta_p, \theta_1, \ldots, \theta_d]^\top,$$

and

$$\boldsymbol{c}_t = \left[\epsilon_{t-1}^2, \ldots, \epsilon_{t-q}^2, h_{t-1}, \ldots, h_{t-p}, \varsigma_{t-1}, \ldots, \varsigma_{t-d}\right]^\top.$$

The conditional variance equation can then be expressed compactly as

$$h_t = \omega + \boldsymbol{\gamma}^\top \boldsymbol{c}_t.$$

The nonnegativity and stationarity constraints for GARCHX models coincide with those for standard GARCH models and are therefore omitted for brevity. Since both the constraint structure and the functional form of the volatility equation are analogous to those in the GARCH model, GARCHX models can be estimated in a similar manner using the proposed BMM algorithm.

APPENDIX F
JOINT ESTIMATION OF GARCH WITH AUTOREGRESSIVE CONDITIONAL MEAN MODELS

Assume that y_t in (1) follows an autoregressive model of order k, denoted as AR(k), which specifies the conditional mean μ_t as

$$\mu_t = \phi_0 + \sum_{i=1}^{k} \phi_i y_{t-i} = \boldsymbol{\phi}^\top \boldsymbol{y}_t,$$

where

$$\boldsymbol{\phi} = [\phi_0, \phi_1, \ldots, \phi_k]^\top,$$

and

$$\boldsymbol{y}_t = [1, y_{t-1}, \ldots, y_{t-k}]^\top.$$

In this section, we demonstrate how the penalty method and the proposed BMM algorithm can be adapted to solve the joint ML estimation of the GARCH(q,p) and AR(k) models. Considering the Gaussian AR(k)-GARCH(q,p) model, we have the following estimation problem

$$\begin{aligned}\min_{\omega,\boldsymbol{\gamma},\boldsymbol{h},\boldsymbol{\phi}} \quad & \sum_{t=1}^{n}\left[\log h_t+\frac{\left(y_t-\boldsymbol{\phi}^\top\boldsymbol{y}_t\right)^2}{h_t}+\frac{\eta}{2}\left(h_t-\omega-\sum_{i=1}^{q}\alpha_i\left(y_{t-i}-\boldsymbol{\phi}^\top\boldsymbol{y}_{t-i}\right)^2-\sum_{j=1}^{p}\beta_j h_{t-j}\right)^2\right]\\ \text{s.t.} \quad & \omega\geq 0_\varepsilon,\ \boldsymbol{\gamma}\geq\mathbf{0},\ \boldsymbol{\gamma}\in\mathcal{S},\end{aligned}$$

where all values of y_t for $t=1-\max\{k,q\},\ldots,0,\ldots,n$, ϵ_t for $t=1-q,\ldots,0$, and h_t for $t=1-p,\ldots,0$ are assumed to be known. The update rules for the variables ω, $\boldsymbol{\gamma}$, and $\boldsymbol{h}$ follow directly from Section (IV). Hence, we focus on the derivation of the update for the $\boldsymbol{\phi}$-block in the remainder of this section.

The subproblem for $\boldsymbol{\phi}$ is given as follows:

$$\min_{\boldsymbol{\phi}} \quad \sum_{t=1}^{n}\left[\frac{1}{\underline{h}_t}\left(y_t-\boldsymbol{\phi}^\top\boldsymbol{y}_t\right)^2+\frac{\eta}{2}\left(\sum_{i=1}^{q}\underline{\alpha}_i\left(y_{t-i}-\boldsymbol{\phi}^\top\boldsymbol{y}_{t-i}\right)^2-b_t\right)^2\right], \tag{32}$$

where

$$b_t=\underline{\boldsymbol{h}}_t^\top\underline{\tilde{\boldsymbol{\beta}}}-\underline{\omega}.$$

Problem (32) is an unconstrained quartic program in $\boldsymbol{\phi}$, for which closed-form solutions are generally unavailable. To address this, we introduce a suitable surrogate function to facilitate efficient optimization.

Define

$$\tilde{\boldsymbol{\phi}}=\left[1,-\boldsymbol{\phi}^\top\right]^\top,$$

and

$$\tilde{\boldsymbol{y}}_t=\left[y_t,\boldsymbol{y}_t^\top\right]^\top.$$

Using Lemma 1, for the first term in the objective function of (32), we have

$$\left(y_t-\boldsymbol{\phi}^\top\boldsymbol{y}_t\right)^2=\tilde{\boldsymbol{\phi}}^\top\left(\tilde{\boldsymbol{y}}_t\tilde{\boldsymbol{y}}_t^\top\right)\tilde{\boldsymbol{\phi}}\leq\|\tilde{\boldsymbol{y}}_t\|^2\tilde{\boldsymbol{\phi}}^\top\tilde{\boldsymbol{\phi}}+2\tilde{\boldsymbol{\phi}}^\top\left(\tilde{\boldsymbol{y}}_t\tilde{\boldsymbol{y}}_t^\top-\|\tilde{\boldsymbol{y}}_t\|^2\boldsymbol{I}\right)\underline{\tilde{\boldsymbol{\phi}}}+\text{const}.$$

For the second term, we write

$$\sum_{i=1}^{q}\underline{\alpha}_i\left(y_{t-i}-\boldsymbol{\phi}^\top\boldsymbol{y}_{t-i}\right)^2=\sum_{i=1}^{q}\underline{\alpha}_i\tilde{\boldsymbol{\phi}}^\top\tilde{\boldsymbol{y}}_{t-i}\tilde{\boldsymbol{y}}_{t-i}^\top\tilde{\boldsymbol{\phi}}=\tilde{\boldsymbol{\phi}}^\top\boldsymbol{Y}_t\tilde{\boldsymbol{\phi}},$$

where $\boldsymbol{Y}_t=\sum_{i=1}^{q}\underline{\alpha}_i\tilde{\boldsymbol{y}}_{t-i}\tilde{\boldsymbol{y}}_{t-i}^\top$. Then we obtain

$$\left(\sum_{i=1}^{q}\underline{\alpha}_i\left(y_{t-i}-\boldsymbol{\phi}^\top\boldsymbol{y}_{t-i}\right)^2-b_t\right)^2=\left(\tilde{\boldsymbol{\phi}}^\top\boldsymbol{Y}_t\tilde{\boldsymbol{\phi}}-b_t\right)^2=\left(\tilde{\boldsymbol{\phi}}^\top\boldsymbol{Y}_t\tilde{\boldsymbol{\phi}}\right)^2-2b_t\tilde{\boldsymbol{\phi}}^\top\boldsymbol{Y}_t\tilde{\boldsymbol{\phi}}+b_t^2. \tag{33}$$

For the term $\left(\tilde{\boldsymbol{\phi}}^\top\boldsymbol{Y}_t\tilde{\boldsymbol{\phi}}\right)^2$ in (33), based on Lemma 1, we have

$$\begin{aligned}\left(\tilde{\boldsymbol{\phi}}^\top\boldsymbol{Y}_t\tilde{\boldsymbol{\phi}}\right)^2&=\operatorname{vec}\left(\tilde{\boldsymbol{\phi}}\tilde{\boldsymbol{\phi}}^\top\right)^\top\operatorname{vec}\left(\boldsymbol{Y}_t\right)\operatorname{vec}\left(\boldsymbol{Y}_t\right)^\top\operatorname{vec}\left(\tilde{\boldsymbol{\phi}}\tilde{\boldsymbol{\phi}}^\top\right)\\&\leq\|\operatorname{vec}\left(\boldsymbol{Y}_t\right)\|^2\operatorname{vec}\left(\tilde{\boldsymbol{\phi}}\tilde{\boldsymbol{\phi}}^\top\right)^\top\operatorname{vec}\left(\tilde{\boldsymbol{\phi}}\tilde{\boldsymbol{\phi}}^\top\right)\\&\quad+2\operatorname{vec}\left(\tilde{\boldsymbol{\phi}}\tilde{\boldsymbol{\phi}}^\top\right)^\top\left(\operatorname{vec}\left(\boldsymbol{Y}_t\right)\operatorname{vec}\left(\boldsymbol{Y}_t\right)^\top-\|\operatorname{vec}\left(\boldsymbol{Y}_t\right)\|^2\boldsymbol{I}\right)\operatorname{vec}\left(\underline{\tilde{\boldsymbol{\phi}}}\underline{\tilde{\boldsymbol{\phi}}}^\top\right)+\text{const}.\\&=\|\operatorname{vec}\left(\boldsymbol{Y}_t\right)\|^2\left\|\tilde{\boldsymbol{\phi}}^\top\tilde{\boldsymbol{\phi}}\right\|^2+2\left(\left(\underline{\tilde{\boldsymbol{\phi}}}^\top\boldsymbol{Y}_t\underline{\tilde{\boldsymbol{\phi}}}\right)\tilde{\boldsymbol{\phi}}^\top\boldsymbol{Y}_t\tilde{\boldsymbol{\phi}}-\|\operatorname{vec}\left(\boldsymbol{Y}_t\right)\|^2\tilde{\boldsymbol{\phi}}^\top\left(\underline{\tilde{\boldsymbol{\phi}}}\underline{\tilde{\boldsymbol{\phi}}}^\top\right)\tilde{\boldsymbol{\phi}}\right)+\text{const}.\end{aligned} \tag{34}$$

Consider the term $-2b_t\tilde{\boldsymbol{\phi}}^\top\boldsymbol{Y}\tilde{\boldsymbol{\phi}}$ in (33). Based on Lemma 1, we have

$$-2b_t\tilde{\boldsymbol{\phi}}^\top\boldsymbol{Y}_t\tilde{\boldsymbol{\phi}}\leq-4b_t\tilde{\boldsymbol{\phi}}^\top\boldsymbol{Y}_t\underline{\tilde{\boldsymbol{\phi}}}+\text{const}.$$

Consider the term $\left\|\tilde{\boldsymbol{\phi}}^\top\tilde{\boldsymbol{\phi}}\right\|^2$ in (34). Based on Lemma 1, we can further derive

$$\begin{aligned}\left\|\tilde{\boldsymbol{\phi}}^\top\tilde{\boldsymbol{\phi}}\right\|^2=\left(\mathbf{1}^\top\left(\tilde{\boldsymbol{\phi}}\odot\tilde{\boldsymbol{\phi}}\right)\right)^2&=\left(\tilde{\boldsymbol{\phi}}\odot\tilde{\boldsymbol{\phi}}\right)^\top\left(\mathbf{1}\mathbf{1}^\top\right)\left(\tilde{\boldsymbol{\phi}}\odot\tilde{\boldsymbol{\phi}}\right)\\&\leq(k+2)\left(\tilde{\boldsymbol{\phi}}\odot\tilde{\boldsymbol{\phi}}\right)^\top\left(\tilde{\boldsymbol{\phi}}\odot\tilde{\boldsymbol{\phi}}\right)+2\left(\tilde{\boldsymbol{\phi}}\odot\tilde{\boldsymbol{\phi}}\right)^\top\left(\mathbf{1}\mathbf{1}^\top-(k+2)\boldsymbol{I}\right)\left(\underline{\tilde{\boldsymbol{\phi}}}\odot\underline{\tilde{\boldsymbol{\phi}}}\right)+\text{const}.,\end{aligned}$$

where $\odot$ denotes the elementwise product. For the term $\tilde{\boldsymbol{\phi}}^\top \boldsymbol{Y}_t \tilde{\boldsymbol{\phi}}$ in (34), based on Lemma 1, we have

$$\tilde{\boldsymbol{\phi}}^\top \boldsymbol{Y}_t \tilde{\boldsymbol{\phi}} \leq \left(\sum_{i=1}^{q} \underline{\alpha}_i \left\|\tilde{\boldsymbol{y}}_{t-i}\right\|^2\right) \tilde{\boldsymbol{\phi}}^\top \tilde{\boldsymbol{\phi}} + 2\tilde{\boldsymbol{\phi}}^\top \left(\boldsymbol{Y}_t - \left(\sum_{i=1}^{q} \underline{\alpha}_i \left\|\tilde{\boldsymbol{y}}_{t-i}\right\|^2\right) \boldsymbol{I}\right) \underline{\tilde{\boldsymbol{\phi}}} + \text{const.}$$

For the term $-\tilde{\boldsymbol{\phi}}^\top \left(\underline{\tilde{\boldsymbol{\phi}}\tilde{\boldsymbol{\phi}}}\right) \tilde{\boldsymbol{\phi}}$ in (34), based on Lemma 1, we have

$$-\tilde{\boldsymbol{\phi}}^\top \left(\underline{\tilde{\boldsymbol{\phi}}\tilde{\boldsymbol{\phi}}}\right) \tilde{\boldsymbol{\phi}} \leq -2\underline{\tilde{\boldsymbol{\phi}}}^\top \underline{\tilde{\boldsymbol{\phi}}\tilde{\boldsymbol{\phi}}}^\top \tilde{\boldsymbol{\phi}} + \text{const.}$$

Combining the above results, the surrogate function for the objective in (32) is given by

$$\begin{aligned}
&\sum_{t=1}^{n} \Bigg[\frac{\eta}{2} \left\|\text{vec}\left(\boldsymbol{Y}_t\right)\right\|^2 (k+2) \left(\tilde{\boldsymbol{\phi}} \odot \tilde{\boldsymbol{\phi}}\right)^\top \left(\tilde{\boldsymbol{\phi}} \odot \tilde{\boldsymbol{\phi}}\right) \\
&+ \eta \left\|\text{vec}\left(\boldsymbol{Y}_t\right)\right\|^2 \left(\tilde{\boldsymbol{\phi}} \odot \tilde{\boldsymbol{\phi}}\right)^\top \left(\boldsymbol{1}\boldsymbol{1}^\top - (k+2)\boldsymbol{I}\right) \left(\underline{\tilde{\boldsymbol{\phi}}} \odot \underline{\tilde{\boldsymbol{\phi}}}\right) + \left(\frac{1}{\underline{h}_t} \|\tilde{\boldsymbol{y}}_t\|^2 + \eta \left(\underline{\tilde{\boldsymbol{\phi}}}^\top \boldsymbol{Y}_t \underline{\tilde{\boldsymbol{\phi}}}\right) \left(\sum_{i=1}^{q} \underline{\alpha}_i \left\|\tilde{\boldsymbol{y}}_{t-i}\right\|^2\right)\right) \left(\tilde{\boldsymbol{\phi}}^\top \tilde{\boldsymbol{\phi}}\right) \\
&+ 2\tilde{\boldsymbol{\phi}}^\top \left(\frac{1}{\underline{h}_t} \left(\tilde{\boldsymbol{y}}_t \tilde{\boldsymbol{y}}_t^\top - \|\tilde{\boldsymbol{y}}_t\|^2 \boldsymbol{I}\right) + \eta \left(\underline{\tilde{\boldsymbol{\phi}}}^\top \boldsymbol{Y}_t \underline{\tilde{\boldsymbol{\phi}}}\right) \left(\boldsymbol{Y}_t - \left(\sum_{i=1}^{q} \underline{\alpha}_i \left\|\tilde{\boldsymbol{y}}_{t-i}\right\|^2\right) \boldsymbol{I}\right) - \eta \left\|\text{vec}\left(\boldsymbol{Y}_t\right)\right\|^2 \underline{\tilde{\boldsymbol{\phi}}\tilde{\boldsymbol{\phi}}}^\top - \eta b_t \boldsymbol{Y}_t\right) \underline{\tilde{\boldsymbol{\phi}}} \Bigg] + \text{const.} \\
=&\sum_{t=1}^{n} \Bigg[\frac{\eta}{2} \left\|\text{vec}\left(\boldsymbol{Y}_t\right)\right\|^2 (k+2) \left(\boldsymbol{\phi} \odot \boldsymbol{\phi}\right)^\top \left(\boldsymbol{\phi} \odot \boldsymbol{\phi}\right) \\
&+ \eta \left\|\text{vec}\left(\boldsymbol{Y}_t\right)\right\|^2 \left(\boldsymbol{\phi} \odot \boldsymbol{\phi}\right)^\top \left[\left(\boldsymbol{1}\boldsymbol{1}^\top - (k+2)\boldsymbol{I}\right) \left(\underline{\tilde{\boldsymbol{\phi}}} \odot \underline{\tilde{\boldsymbol{\phi}}}\right)\right]_{2:k+2} + \left(\frac{1}{\underline{h}_t} \|\tilde{\boldsymbol{y}}_t\|^2 + \eta \left(\underline{\tilde{\boldsymbol{\phi}}}^\top \boldsymbol{Y}_t \underline{\tilde{\boldsymbol{\phi}}}\right) \left(\sum_{i=1}^{q} \underline{\alpha}_i \left\|\tilde{\boldsymbol{y}}_{t-i}\right\|^2\right)\right) \left(\boldsymbol{\phi}^\top \boldsymbol{\phi}\right) \\
&+ 2\boldsymbol{\phi}^\top \Bigg[\left(\frac{1}{\underline{h}_t} \left(\tilde{\boldsymbol{y}}_t \tilde{\boldsymbol{y}}_t^\top - \|\tilde{\boldsymbol{y}}_t\|^2 \boldsymbol{I}\right) + \eta \left(\underline{\tilde{\boldsymbol{\phi}}}^\top \boldsymbol{Y}_t \underline{\tilde{\boldsymbol{\phi}}}\right) \left(\boldsymbol{Y}_t - \left(\sum_{i=1}^{q} \underline{\alpha}_i \left\|\tilde{\boldsymbol{y}}_{t-i}\right\|^2\right) \boldsymbol{I}\right) \\
&- \eta \left\|\text{vec}\left(\boldsymbol{Y}_t\right)\right\|^2 \underline{\tilde{\boldsymbol{\phi}}\tilde{\boldsymbol{\phi}}}^\top - \eta b_t \boldsymbol{Y}_t\Bigg) \underline{\tilde{\boldsymbol{\phi}}} \Bigg]_{2:k+2} \Bigg] + \text{const.},
\end{aligned}$$

where $\boldsymbol{x}_{2:k+2}$ is the subvector of $\boldsymbol{x}$ from the 2nd to the $(k+2)$-th elements. The above function is quartic in $\boldsymbol{\phi}$ and is separable over different elements in $\boldsymbol{\phi}$. The solution of $\boldsymbol{\phi}$ based on the above surrogate function can be obtained based on the cubic formula.