Received: 14 June 2023 | Revised: 17 October 2023 | Accepted: 17 November 2023

DOI: 10.1002/sam.11652

WILEY

RESEARCH ARTICLE

The generalized hyperbolic family and automatic model selection through the multiple-choice LASSO

Luca Bagnato[1] | **Alessio Farcomeni**[2] | **Antonio Punzo**[3]

[1]Department of Economic and Social Sciences, Catholic University of the Sacred Heart, Piacenza, Italy

[2]Department of Economics and Finance, Tor Vergata University of Rome, Rome, Italy

[3]Department of Economics and Business, University of Catania, Catania, Italy

Correspondence
Antonio Punzo, Department of Economics and Business, University of Catania, Catania, Italy.
Email: antonio.punzo@unict.it

Funding information
2022XRHT8R; E63C22002120006

Abstract

We revisit the generalized hyperbolic (GH) distribution and its nested models. These include widely used parametric choices like the multivariate normal, skew-*t*, Laplace, and several others. We also introduce the multiple-choice LASSO, a novel penalized method for choosing among alternative constraints on the same parameter. A hierarchical multiple-choice Least Absolute Shrinkage and Selection Operator (LASSO) penalized likelihood is optimized to perform simultaneous model selection and inference within the GH family. We illustrate our approach through a simulation study and a real data example. The methodology proposed in this paper has been implemented in R functions which are available as supplementary material.

KEYWORDS

EM algorithm, generalized hyperbolic distribution, kurtosis, penalized likelihood, skewness

1 | INTRODUCTION

As Cox [12] stated, "choice of an appropriate family of distributions may be the most challenging phase of analysis.". Researchers always face a trade-off between goodness of fit and simplicity of the distributional assumptions. The generalized hyperbolic (GH) distribution [5] provides a particularly convenient family in this direction. It has flexible tails, spanning from Gaussian to exponential tails, and can handle skewed scatters. Moreover, the family contains as special cases several widely used parametric distributions. Applications of the GH family are widespread [10, 34, 44]. In particular, the GH distribution is rapidly becoming the most popular model for financial applications (e.g., Refs. [6, 9, 14]; S. I. [19, 39, 50]), and this is corroborated by the presence of this distribution in classic textbooks of the financial literature where this model is treated as a reference model (e.g., Ref. [30]).

A contribution of this work indeed is that we outline a precise taxonomy of the GH family and its many nested models (the most famous ones). The main novelty with respect to previous works is that we do not compare the GH and alternatives by separately fitting each model, but we specify a unified penalized likelihood framework that successfully performs simultaneous parameter estimation and model choice.

To proceed in this direction, we introduce the multiple-choice Least Absolute Shrinkage and Selection Operator (LASSO), a new type of LASSO penalty. Indeed, LASSO-type penalties [47] are commonly used to shrink parameters to a single specific value (typically, zero). Nested models within the GH family are selected by fixing certain shape parameters at one of the different alternative values. The multiple-choice LASSO is devised precisely for this purpose: to allow shrinkage of the same parameter toward one of several alternative values. To restrict the possible choices, we will also build on the hierarchical LASSO (as introduced by Ref. [7], see also Ref. [24]) so that certain constraints can be activated only conditionally.

The rest of the paper is as follows: in the next section, we review the GH distribution and provide a map of its nested models. After reviewing LASSO and hierarchical LASSO, we then introduce the multiple-choice LASSO. In Section 3, we use the hierarchical and multiple-choice LASSO to define penalized objective functions that can yield any model within the GH family, and describe how to optimize those in Section 4. In Section 5, we illustrate the results of a simulation study conducted with the aim of investigating the ability of our multiple-choice LASSO procedure in discovering the true data-generating model. We present a data analysis in Section 6. Some concluding remarks are given in Section 7.

The methodology proposed in this paper has been implemented in R [43] functions which are available as supplementary material.

2 | SETUP

2.1 | The generalized hyperbolic distribution and its special cases

The joint probability density function of a d-variate random variable $\boldsymbol{X}$ following the GH distribution can be written as

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \frac{\exp\left[(\boldsymbol{x}-\boldsymbol{\mu})'\Sigma^{-1}\boldsymbol{\gamma}\right]}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}K_{\lambda}(\sqrt{\chi\psi})}\left[\frac{\chi+\delta(\boldsymbol{x};\boldsymbol{\mu},\Sigma)}{\psi+\rho(\boldsymbol{\gamma},\Sigma)}\right]^{\frac{\lambda-\frac{d}{2}}{2}} K_{\lambda-\frac{d}{2}}\left(\sqrt{[\chi+\delta(\boldsymbol{x};\boldsymbol{\mu},\Sigma)][\psi+\rho(\boldsymbol{\gamma},\Sigma)]}\right), \quad (1)$$

where $\boldsymbol{\mu}\in\mathbb{R}^d$ is the location parameter, Σ is a $d\times d$ scale matrix, such that $|\Sigma|=1$ for identifiability purposes (see Ref. [29], for details), $\boldsymbol{\gamma}\in\mathbb{R}^d$ is the skewness parameter, $\lambda\in\mathbb{R}$ is the index parameter, and $\chi,\psi>0$ are concentration parameters; compactly, we adopt the notation $\boldsymbol{X}\sim\mathcal{GH}_d(\boldsymbol{\mu},\Sigma,\boldsymbol{\gamma},\lambda,\chi,\psi)$. In (1), $\boldsymbol{\theta}=\{\boldsymbol{\mu},\Sigma,\boldsymbol{\gamma},\lambda,\chi,\psi\}$ contains all the parameters of the model, $\delta(\boldsymbol{x};\boldsymbol{\mu},\Sigma)=(\boldsymbol{x}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$ is the squared Mahalanobis distance between $\boldsymbol{x}$ and $\boldsymbol{\mu}$ (with covariance matrix Σ), $\rho(\boldsymbol{\gamma},\Sigma)=\boldsymbol{\gamma}'\Sigma^{-1}\boldsymbol{\gamma}$, and K_λ is the modified Bessel function of the third kind with index λ.

It is of practical importance to note that $\boldsymbol{X}\sim\mathcal{GH}_d(\boldsymbol{\mu},\Sigma,\boldsymbol{\gamma},\lambda,\chi,\psi)$ has the normal mean–variance mixture (NMVM) representation

$$\boldsymbol{X}=\boldsymbol{\mu}+W\boldsymbol{\gamma}+\sqrt{W}\boldsymbol{U}, \quad (2)$$

where W has a generalized inverse Gaussian (GIG) distribution, in symbols $W\sim\mathcal{GIG}(\lambda,\chi,\psi)$ (see Appendix A), and $\boldsymbol{U}\sim\mathcal{N}_d(\boldsymbol{0},\Sigma)$, where $\mathcal{N}_d(\boldsymbol{\mu},\Sigma)$ denotes a d-variate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix Σ. As a related alternative, we can refer to the following hierarchical representation of $\boldsymbol{X}\sim\mathcal{GH}_d(\boldsymbol{\mu},\Sigma,\boldsymbol{\gamma},\lambda,\chi,\psi)$ as

$$\begin{aligned} W &\sim \mathcal{GIG}(\lambda,\chi,\psi)\\ \boldsymbol{X}\mid W=w &\sim \mathcal{N}_d(\boldsymbol{\mu}+w\boldsymbol{\gamma},w\Sigma), \end{aligned} \quad (3)$$

where w is a realization of W. The hierarchical representation in (3) is useful for random data generation and for the implementation of the ECME algorithm discussed in Section 4.

Figure 1 gives a hierarchical representation of all the existing models the GH distribution nests as special or limiting cases by varying the values/ranges of $\boldsymbol{\gamma}$, λ, χ, and ψ. Such a hierarchy is easily derived by using the representation of the GH distribution given in (2). Appendix B illustrates how to obtain some of these special and limiting cases, those we believe are more difficult to be derived and about which there is more confusion in the literature due to the use of different identifiability constraints. On the left/right of Figure 1, we have the models related to negative/positive values of λ. Instead, on the bottom (below the dashed line), we have the symmetric models (those with $\gamma=0$); as we can see, the symmetric counterpart of each model on the top is available. The diagram in Figure 1 can be considered as a contribution of this paper. It provides, for the first time to our knowledge, a complete and organized taxonomy of all the models nested within the GH family.

Summarizing we have: 2 possibilities for $\boldsymbol{\gamma}$ ($\boldsymbol{\gamma}$ free or $\boldsymbol{\gamma}=\boldsymbol{0}$), 6 possibilities for λ ($\lambda\to-\infty$, $\lambda<0$, $\lambda=-1/2$, $\lambda=(d+1)/2$, $\lambda=1$ or $\lambda>0$), 3 possibilities for χ (χ free, $\chi\to0$ or $\chi\to\infty$), and 2 possibilities for ψ (ψ free and $\psi\to0$). Combining all these possibilities would generate $2\cdot6\cdot3\cdot2=72$ models. However, many of them are not of practical interest. Just as two examples, the combination $\{\boldsymbol{\gamma}=\boldsymbol{0},\lambda<0,\chi\to0,\psi\to0\}$ would generate a degenerate t distribution on $\boldsymbol{\mu}$, while the combination $\{\boldsymbol{\gamma}=\boldsymbol{0},\lambda=1,\chi\to0,\psi\to0\}$ would generate a degenerate Laplace distribution on $\boldsymbol{\mu}$.

2.2 | Preliminaries about LASSO and hierarchical LASSO

Suppose to be interested to a particular configuration/-value of $\boldsymbol{\theta}$, say $\boldsymbol{\theta}_0$. The LASSO involves specification of an L_1 penalty for (possibly, a subset of) the parameter vector $\boldsymbol{\theta}$, so that the estimate $\widehat{\boldsymbol{\theta}}$ is exactly equal to $\boldsymbol{\theta}_0$ if the likelihood at $\boldsymbol{\theta}_0$ is not too far from the maximum. More formally, given a random sample $S_n=\{\boldsymbol{x}_i;i=1,\dots,n\}$ (observed data) from $\boldsymbol{X}\sim\mathcal{GH}_d(\boldsymbol{\mu},\Sigma,\boldsymbol{\gamma},\lambda,\chi,\psi)$, estimation proceeds through optimization of the penalized log-likelihood

$$\sum_{i=1}^{n}\log\left[f(\boldsymbol{x}_i;\boldsymbol{\theta})\right]-P_h(\boldsymbol{\theta}) \quad (4)$$

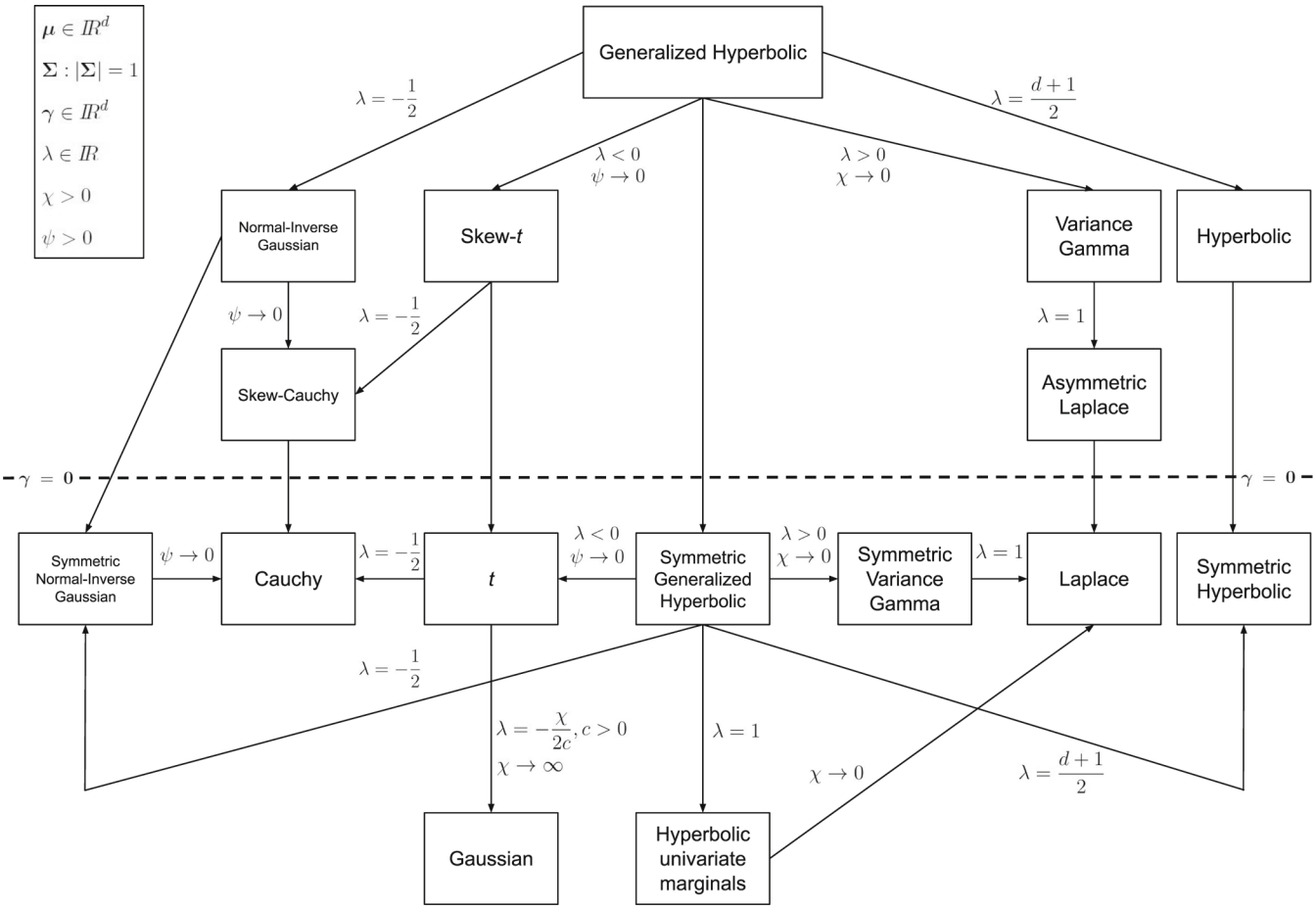


FIGURE 1 Hierarchy of the special and limiting cases of the generalized hyperbolic (GH) distribution in terms of γ, λ, χ and ψ. On the top-left corner, a recap on the values the GH-parameters can assume is provided.

for an appropriate penalty function $P_h(\boldsymbol{\theta})$, with $f(\cdot;\boldsymbol{\theta})$ being defined in (1). In classical LASSO, $P_h(\boldsymbol{\theta}) = h\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_{L_1}$, where $\|\cdot\|_{L_1}$ indicates the L_1-norm (the sum of absolute values) and $h > 0$ is a fixed penalty parameter. In linear models, often times $\boldsymbol{\theta}_0 = \mathbf{0}$.

The resulting estimator is less efficient than the MLE, but superefficient at $\boldsymbol{\theta}_0$ (see, for example, Ref. [48] and references therein). It is well known that any superefficient estimator may improve efficient estimators at most on a subset of the parameter space of zero Lebesgue measure.

In our work, we will also make use of the hierarchical LASSO [7], which is devised for structured sparsity: some constraints can be activated only if others are simultaneously active. Without loss of generality, assume we allow $\theta_c = 0$ only if $\theta_d = 0$, with θ_c and θ_d being two elements of $\boldsymbol{\theta}$. This can be obtained expressing

$$P_h(\boldsymbol{\theta}) = h\left[|\theta_d| + \frac{\max(|\theta_c|, |\theta_d|)}{2}\right].$$

In words, some shrinkage for θ_c is allowed if $|\theta_c| > |\theta_d|$, but the constraint on $|\theta_c|$ can be exactly activated only as soon as $\theta_d = 0$; see Bien et al. [7] on this point.

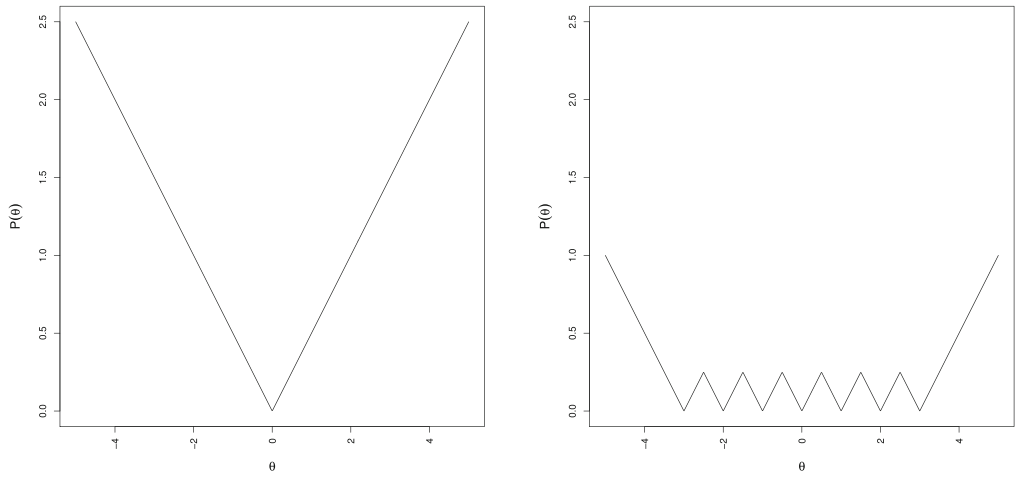
2.3 | The multiple-choice LASSO

We introduce in this section the multiple-choice LASSO, which can be used to enforce one of several constraints on the same parameter. For simplicity, assume we have a one-dimensional parameter θ and several possible constraints on it, that is, we require superefficiency not only at a single point θ_0 in the parameter space, but at a finite collection of points $\{\theta_1, \ldots, \theta_C\}$. Our proposal is to specify

$$P_h(\theta) = h \min(|\theta - \theta_1|, |\theta - \theta_2|, \ldots, |\theta - \theta_C|). \quad (5)$$

In words, only the smallest among all possible L_1 norms contribute to the penalty. The idea is that if the MLE is close enough to θ_j for some $j = 1, \ldots, C$, then $\widehat{\theta} = \theta_j$ as the remaining L_1 norms are simply ignored due to the minimum operator.

For illustration, in Figure 2A, B we show the penalty function for LASSO and multiple-choice LASSO, respectively, for a one-dimensional problem with $h = 0.5$ in both cases. For the LASSO, we set $\theta_0 = 0$, while for multiple-choice LASSO, we set $\theta_0 \in \{-3, -2, -1, 0, 1, 2, 3\}$. The sawtooth shape of the penalty function for the



(A) LASSO (B) Multiple-choice LASSO

FIGURE 2 The penalty function for Least Absolute Shrinkage and Selection Operator (LASSO) (left panel), with $\theta_0 = 0$; and the penalty function for multiple-choice LASSO (right panel), with $\theta_0 \in \{-3, -2, -1, 0, 1, 2, 3\}$.

multiple-choice LASSO is what allows objective functions to be optimized exactly at θ_j, $j = 1, \ldots, C$.

The resulting penalized objective function is clearly non-convex. While in some cases, specific algorithms might be exploited to optimize it, since the parameter space is low dimensional in our context, we propose to simply use a numerical method like the Constrained Optimization BY Linear Approximation (COBYLA) algorithm [40].

3 | SHAPE DETECTION THROUGH PENALIZED LIKELIHOOD MAXIMIZATION

As discussed at the end of Section 2.1, all possible combinations of the discussed constraints on the parameters $\boldsymbol{\gamma}$, λ, χ, and ψ would lead to 72 parametric distributions, nested within the GH distribution. Of these, only 16 have a clear interpretation as outlined in Section 2.1 and Figure 1.

In the following, we show how to specify a multiple-choice LASSO-type penalized likelihood function which can possibly lead to any of the 72 models nested in the GH distribution. We then specify a multiple-choice hierarchical LASSO-type penalized likelihood which restricts the possible solutions only to the 16 models in Figure 1.

The penalized likelihood specification is as in (4). A simple way to proceed is to specify $P_h(\boldsymbol{\gamma}, \lambda, \chi, \psi)$ as a multiple-choice LASSO penalty of the kind

$$P_h(\boldsymbol{\gamma}, \lambda, \chi, \psi) = h\Bigg\{ \min\left[\left|\lambda - \frac{d+1}{2}\right|, \left|\lambda + \frac{1}{2}\right|, |\lambda - 1|, I(\lambda < 0)\left|\frac{1}{\lambda}\right|\right] + \min\left(|\chi|, \left|\frac{1}{\chi}\right|\right) + |\psi| + \|\boldsymbol{\gamma}\|_{L_2}\Bigg\}. \tag{6}$$

We use here a penalty on $\|\boldsymbol{\gamma}\|_{L_2}$ to constrain all d elements of $\boldsymbol{\gamma}$ to be zero, in the spirit of group LASSO (see, for example, Refs. [24, 49]). In case $\lambda \to -\infty$ and $\chi \to \infty$, define $c = -\chi/2\lambda$ as scale parameter of the resulting Gaussian distribution. Note that the constraint $|1/\lambda|$ is satisfied by $\lambda \to \pm\infty$.

Penalty (6) will allow the user to select any of the 72 possible parametric distributions obtained through appropriate constraints. Many of these models might fit well, but do not have a direct interpretation. In order to restrict the list of possible models to the 16 ones listed in Figure 1, we must exclude several possible combinations of constraints on the parameters. To this end, we combine the hierarchical LASSO and the multiple-choice LASSO frameworks and specify the penalty as

$$\begin{aligned} P_h(\gamma, \lambda, \chi, \psi) = h\Bigg\{ & \frac{\|\gamma\|_{L_2}}{\sqrt{d}} + I(\lambda \le 0) \min\left[\left|\lambda + \frac{1}{2}\right|\right] + \\ & + \frac{1}{2}\max\left(\left|\lambda + \frac{1}{2}\right|, |\psi|\right), |\psi| + \frac{1}{2}\max\left(\left|\lambda + \frac{1}{2}\right|, |\psi|\right), \\ & \frac{1}{4}\max\left(\frac{\|\gamma\|_{L_2}}{\sqrt{d}}, \left|\frac{1}{\lambda}\right|, |\psi|, \left|\frac{1}{\chi}\right|\right)\Bigg] + \\ & + I(\lambda > 0) \min\Bigg[\left|\lambda - \frac{d+1}{2}\right|, |\chi| \\ & + \frac{1}{2}\max(|\lambda - 1|, |\chi|), \frac{1}{2}\max\left(\frac{\|\gamma\|_{L_2}}{\sqrt{d}}, |\lambda - 1|\right)\Bigg]\Bigg\}, \end{aligned} \tag{7}$$

where $I(A)$ denotes the indicator function of $A \subseteq \mathbb{R}$ and $h > 0$ is a penalty parameter. In the expression above, we divide by $\sqrt{d}$ to normalize the L_2 norm with respect to the number of elements of the vector involved.

To fix the ideas, we discuss how the GH and Gaussian models are obtained. If the MLE is far from any of the special cases in Figure 1 and the penalty parameter is not too large, no constraint will be activated and the resulting model will be a GH. Suppose now the MLE is close enough

to the case $\boldsymbol{\gamma} = \mathbf{0}$, with sufficiently small λ, large χ, and ψ close to zero. The low $\|\boldsymbol{\gamma}\|_{L_2}$ will make it advantageous to activate the constraint leading to symmetric models. The negative λ will remove the third addend of the penalty, which is multiplied by $I(\lambda > 0)$. For the second addend, the minimum among the three elements listed will be the third, as λ at the MLE will definitely be much smaller than 0.5. Hence, the penalty will essentially reduce to

$$\frac{h}{4}\max\left(\frac{\|\boldsymbol{\gamma}\|_{L_2}}{\sqrt{d}}, \left|\frac{1}{\lambda}\right|, |\psi|, \left|\frac{1}{\chi}\right|\right),$$

and the max operator will lead all the constraints to activate ($\lambda \to -\infty$, $\psi \to 0$, $\chi \to \infty$, $\|\boldsymbol{\gamma}\|_{L_2} \to \mathbf{0}$), leading to the Gaussian model.

4 | PENALIZED MAXIMUM LIKELIHOOD ESTIMATION

We consider a penalized maximum likelihood (ML) approach, with the penalty term given in (6) or (7), to estimate $\boldsymbol{\theta}$ in model (1). Given both the random sample S_n and a value for h, the penalized ML estimation method is based on the maximization of the penalized (observed-data) log-likelihood function

$$\ell_{\text{pen}}(\boldsymbol{\theta}|h) = \sum_{i=1}^{n} \ln f(\boldsymbol{x}_i;\boldsymbol{\theta}) - P_h(\boldsymbol{\gamma},\lambda,\chi,\psi). \tag{8}$$

However, the problem of directly maximizing $\ell_{\text{pen}}(\boldsymbol{\theta}|h)$ over $\boldsymbol{\theta}$ is not particularly easy. The penalized ML fitting is simplified considerably by the application of algorithms based on the expectation–maximization (EM) principle [13]. These algorithms are the classical way to compute ML estimates for parameters of distributions which are defined as a mixture.

Regardless of the particular variant of the EM algorithm used, it is convenient to view the observed data as incomplete. The complete data are $\{(\boldsymbol{x}_i, w_i); i = 1, \dots, n\}$, where the missing variables $w_1, \dots, w_n$ are defined based on the hierarchical representation given in (3)—so that

$$\boldsymbol{X}_i \mid W_i = w_i \sim \mathcal{N}_d(\boldsymbol{\mu} + w_i\boldsymbol{\gamma}, w_i\Sigma),$$

independently for $i \in \{1, \dots, n\}$, and

$$W_i \sim \mathcal{GIG}(\lambda,\chi,\psi).$$

Because of this conditional structure, the penalized complete-data log-likelihood function can be written as

$$\ell_{\text{pen,c}}(\boldsymbol{\theta}|h) = \ell_{1c}(\boldsymbol{\mu},\Sigma,\boldsymbol{\gamma}) + \ell_{2c}(\lambda,\chi,\psi) - P_h(\boldsymbol{\gamma},\lambda,\chi,\psi), \tag{9}$$

where

$$\ell_{1c}(\boldsymbol{\mu},\Sigma,\boldsymbol{\gamma}) = \sum_{i=1}^{n}\left[-\frac{d}{2}\ln(2\pi) - \frac{d}{2}\ln(w_i) - \frac{1}{2}\ln|\Sigma| - \frac{\delta(\boldsymbol{x}_i;\boldsymbol{\mu},\Sigma)}{2w_i} + (\mathbf{x}_i - \boldsymbol{\mu})\Sigma^{-1}\boldsymbol{\gamma} - \frac{w_i}{2}\boldsymbol{\gamma}'\Sigma^{-1}\boldsymbol{\gamma}\right], \tag{10}$$

and

$$\ell_{2c}(\lambda,\chi,\psi) = \sum_{i=1}^{n}\left\{(\lambda-1)\ln(w_i) - \frac{1}{2}\frac{\chi}{w_i} - \frac{1}{2}\psi w_i - \frac{1}{2}\lambda\ln(\chi) + \frac{1}{2}\lambda\ln(\psi) - \ln\left[2K_\lambda(\sqrt{\chi\psi})\right]\right\}. \tag{11}$$

Working on $\ell_{\text{pen,c}}(\boldsymbol{\theta}|h)$, we adopt the expectation-conditional maximization either (ECME) algorithm [25]. The ECME algorithm is an extension of the expectation-conditional maximum (ECM) algorithm which, in turn, is an extension of the EM algorithm [28]. The ECM algorithm replaces the M-step of the EM algorithm with a number of computationally simpler conditional maximization (CM) steps. The ECME algorithm generalizes the ECM algorithm by conditionally maximizing on some or all of the CM-steps, the incomplete-data (penalized) log-likelihood. As for the EM and ECM algorithms, the ECME algorithm monotonically increases the likelihood and reliably converges to a stationary point of the likelihood function [28]. Moreover, Liu and Rubin [25] found the ECME algorithm to be nearly always faster than both the EM and ECM algorithms in terms of number of iterations, and that it can be faster in total computer time by orders of magnitude. In our case, the ECME algorithm iterates between three steps, one E-step and two CM-steps, until convergence. The two CM-steps arise from the partition of $\boldsymbol{\theta}$ as $\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\}$, where $\boldsymbol{\theta}_1 = \{\boldsymbol{\mu}, \Sigma\}$ and $\boldsymbol{\theta}_2 = \{\boldsymbol{\gamma}, \lambda, \chi, \psi\}$. The partition is chosen in such a way that all the parameters in the penalization function $P_h(\cdot)$ belongs to $\boldsymbol{\theta}_2$.

Below, we outline the generic iteration of the ECME algorithm. As in Melnykov and Zhu [32, 33], quantities/-parameters marked with one dot will correspond to the previous iteration and those marked with two dots will represent the estimates at the current iteration.

4.1 | E-Step

The E-step is only needed for the first CM-step of the algorithm—where we update $\boldsymbol{\theta}_1$—and requires the calculation of

$$Q\left(\boldsymbol{\theta}_1, \dot{\boldsymbol{\theta}}_2 | \dot{\boldsymbol{\theta}}\right) = Q_1(\boldsymbol{\mu},\Sigma,\dot{\boldsymbol{\gamma}}|\dot{\boldsymbol{\theta}}) + C, \tag{12}$$

the conditional expectation of $\ell_{\text{pen},c}(\boldsymbol{\theta}|h)$ given the observed data, using the current fit $\dot{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$, with $\boldsymbol{\theta}_2$ fixed at $\dot{\boldsymbol{\theta}}_2$ and where C is a constant not involving parameters inside $\boldsymbol{\theta}_1$. In (12), $Q_1(\boldsymbol{\mu}, \Sigma, \dot{\boldsymbol{\gamma}}|\dot{\boldsymbol{\theta}})$ is the conditional expectation of $\ell_{1c}(\boldsymbol{\mu}, \Sigma, \boldsymbol{\gamma})$ in (9).

To compute $Q\left(\boldsymbol{\theta}_1, \dot{\boldsymbol{\theta}}_2|\dot{\boldsymbol{\theta}}\right)$, we need to replace any function $m(W_i)$ of the latent variable W_i which appears in (10), provided that it is related with either $\boldsymbol{\mu}$ or Σ, by $E_{\dot{\boldsymbol{\theta}}}[m(W_i)|\boldsymbol{X}_i = \boldsymbol{x}_i]$, where the expectation (as it can be noted by the subscript) is taken using the current fit $\dot{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$, $i = 1, \ldots, n$. In particular, the functions satisfying these requirements, involved in (10), are $m_1(w) = w$ and $m_2(w) = 1/w$. To calculate the expectations of m_1 and m_2, we first note that

$$W_i \mid \boldsymbol{X}_i = \boldsymbol{x}_i \sim \mathcal{GIG}\left(\lambda - \frac{d}{2}, \delta(\boldsymbol{x}_i; \boldsymbol{\mu}, \Sigma) + \chi, \boldsymbol{\gamma}' \Sigma^{-1} \boldsymbol{\gamma} + \psi\right).$$

Therefore, according to (A2) and (A3), respectively, we need to compute the following quantities

$$\dot{v}_i := \mathrm{E}_{\dot{\boldsymbol{\theta}}}(W_i|\boldsymbol{X}_i = \boldsymbol{x}_i) = \sqrt{\frac{\delta\left(\boldsymbol{x}_i; \dot{\boldsymbol{\mu}}, \dot{\Sigma}\right) + \dot{\chi}}{\dot{\psi}}} \frac{K_{\dot{\lambda} - \frac{d}{2} + 1}\left\{\sqrt{\dot{\psi}\left[\delta\left(\boldsymbol{x}_i; \dot{\boldsymbol{\mu}}, \dot{\Sigma}\right) + \dot{\chi}\right]}\right\}}{K_{\dot{\lambda} - \frac{d}{2}}\left\{\sqrt{\dot{\psi}\left[\delta\left(\boldsymbol{x}_i; \dot{\boldsymbol{\mu}}, \dot{\Sigma}\right) + \dot{\chi}\right]}\right\}} \tag{13}$$

$$\dot{u}_i := \mathrm{E}_{\dot{\boldsymbol{\theta}}}\left(W_i^{-1}|\boldsymbol{X}_i = \boldsymbol{x}_i\right) = \sqrt{\frac{\dot{\psi}}{\delta\left(\boldsymbol{x}_i; \dot{\boldsymbol{\mu}}, \dot{\Sigma}\right) + \dot{\chi}}} \frac{K_{\dot{\lambda} - \frac{d}{2} + 1}\left\{\sqrt{\dot{\psi}\left[\delta\left(\boldsymbol{x}_i; \dot{\boldsymbol{\mu}}, \dot{\Sigma}\right) + \dot{\chi}\right]}\right\}}{K_{\dot{\lambda} - \frac{d}{2}}\left\{\sqrt{\dot{\psi}\left[\delta\left(\boldsymbol{x}_i; \dot{\boldsymbol{\mu}}, \dot{\Sigma}\right) + \dot{\chi}\right]}\right\}} - \frac{2\left(\dot{\lambda} - \frac{d}{2}\right)}{\delta\left(\boldsymbol{x}_i; \dot{\boldsymbol{\mu}}, \dot{\Sigma}\right) + \dot{\chi}}. \tag{14}$$

Then, by substituting w_i with $\dot{v}_i$ and $1/w_i$ with $\dot{u}_i$ in $\ell_{1c}(\boldsymbol{\mu}, \Sigma, \boldsymbol{\gamma})$, we obtain

$$Q_1(\boldsymbol{\mu}, \Sigma, \dot{\boldsymbol{\gamma}}|\dot{\boldsymbol{\theta}}) = \sum_{i=1}^{n}\left[-\frac{1}{2}\ln|\Sigma| - \frac{\dot{u}_i}{2}\delta(\mathbf{x}_i; \boldsymbol{\mu}, \Sigma) + (\mathbf{x}_i - \boldsymbol{\mu})'\Sigma^{-1}\dot{\boldsymbol{\gamma}} - \frac{\dot{v}_i}{2}\dot{\boldsymbol{\gamma}}'\Sigma^{-1}\dot{\boldsymbol{\gamma}}\right], \tag{15}$$

where we dropped the terms which are constant with respect to $\boldsymbol{\mu}$ and Σ.

4.2 | CM-step 1

The first CM-step requires the calculation of $\ddot{\boldsymbol{\theta}}_1$ as the value of $\boldsymbol{\theta}_1$ that maximizes $Q_1(\boldsymbol{\mu}, \Sigma, \dot{\boldsymbol{\gamma}}|\dot{\boldsymbol{\theta}})$ in (15), with $\boldsymbol{\theta}_2$ fixed at $\dot{\boldsymbol{\theta}}_2$. After simple algebra, we obtain the following updates

$$\ddot{\boldsymbol{\mu}} = \frac{1}{n\dot{\bar{u}}}\left(\sum_{i=1}^{n}\dot{u}_i\boldsymbol{x}_i - \dot{\boldsymbol{\gamma}}\right) \quad \text{and} \quad \ddot{\Sigma} = \left|\ddot{\Sigma}^*\right|^{-\frac{1}{d}}\ddot{\Sigma}^* \tag{16}$$

where

$$\ddot{\Sigma}^* = \frac{1}{n}\sum_{i=1}^{n}\dot{u}_i(\boldsymbol{x}_i - \ddot{\boldsymbol{\mu}})(\boldsymbol{x}_i - \ddot{\boldsymbol{\mu}})' - (\bar{\boldsymbol{x}} - \ddot{\boldsymbol{\mu}})\dot{\boldsymbol{\gamma}}' - \dot{\boldsymbol{\gamma}}(\bar{\boldsymbol{x}} - \ddot{\boldsymbol{\mu}})' + \dot{\bar{v}}\dot{\boldsymbol{\gamma}}\dot{\boldsymbol{\gamma}}', \tag{17}$$

$\dot{\bar{u}} = \sum_{i=1}^{n}\dot{u}_i/n$, $\dot{\bar{v}} = \sum_{i=1}^{n}\dot{v}_i/n$, and $\bar{\mathbf{x}} = \sum_{i=1}^{n}\boldsymbol{x}_i/n$. In (16), the scalar $\left|\ddot{\Sigma}^*\right|^{-\frac{1}{d}}$ is needed to ensure the identifiability constraint $|\ddot{\Sigma}| = 1$.

4.3 | CM-step 2

In the second CM-step, given h, we choose the value of $\boldsymbol{\theta}_2$ that maximizes $\ell_{\text{pen}}(\boldsymbol{\theta}|h)$ in (8), with $\boldsymbol{\theta}_1$ fixed at $\ddot{\boldsymbol{\theta}}_1$. As a closed-form solution for $\ddot{\boldsymbol{\theta}}_2$ is not analytically available, numerical optimization is needed, and any general-purpose optimizer can be used with this aim. Operationally, we perform an unconstrained maximization on $\mathbb{R}^{d+3}$, based on a (log/exp) transformation/back-transformation approach for χ and ϕ, via the general-purpose optimizer `optim()` for R, included in the **stats** package. In analogy with Bagnato and Punzo [4], we try two different commonly used algorithms for maximization: Nelder–Mead, which is derivatives-free, and BFGS which uses (numerical) second-order derivatives. They can be passed to `optim()` via the argument `method`. Once the two algorithms are run, we take the best solution in terms of $\ell_{\text{pen}}(\boldsymbol{\theta}|h)$; see, for example, Punzo and Bagnato [41] for a comparison of the two algorithms, in terms of parameter recovery and computational time, for ML estimation. The choice to run both the algorithms is motivated by two facts: (1) sometimes the algorithms do not provide the same solution, and (2) it can happen that an algorithm does not reach convergence.

4.4 | Selecting the penalty parameter

The choice of the penalty parameter h has direct consequences on the estimation of $\boldsymbol{\theta}$ and, as a sub-product, on the selection of the best model in 1. Optimal penalty/tuning parameters are "difficult to calibrate in

practice" [23]. Specific techniques have their proponents and opponents, making the task even more difficult. Generally speaking, choosing h by trial and error is informative, but it is also convenient to have an objective selection method. This is the reason why data-driven methods are typically preferred, and the literature about them is vast [46]. Among data-driven methods, cross-validation (CV) is a popular choice because it is easy to understand and versatile. However, even with its simplicity and versatility, CV suffers from a heavy computational burden [18]. As a "simplified" variant of the CV method to select h, we consider a simple grid-search partial leave-one-out likelihood cross-validation (LCV) strategy [46]; the term "grid-search" refers to the fact that the LCV statistic is only evaluated on a convenient grid of values, while the term "partial" refers to the fact that we only allow to a proportion p of the sample to be left out one unit at a time. These choices are motivated by the need to speed up the computation that, otherwise, would be too computationally cumbersome.

In detail, we consider the LCV statistic

$$\mathrm{LCV}_p(h) = \frac{1}{\lfloor pn \rfloor} \sum_{\boldsymbol{x}_i \in S_{\lfloor pn \rfloor}} \ln \left[f\left(\boldsymbol{x}_i; \widehat{\boldsymbol{\theta}}_{h, S_n \backslash \{\mathbf{x}_i\}} \right) \right], \qquad (18)$$

where $S_{\lfloor pn \rfloor} \subseteq S_n$ is the sub-sample, of size $\lfloor pn \rfloor$, which is allowed to be left out, and $\widehat{\boldsymbol{\theta}}_{h, S_n \backslash \{\boldsymbol{x}_i\}}$ is the penalized ML estimate of $\boldsymbol{\theta}$, with penalty parameter h, obtained on $S_n \backslash \{\boldsymbol{x}_i\}$ (refer to Section 4). For each value of h in a pre-specified grid G, we first compute $\mathrm{LCV}_p(h)$; then, we select the value of h in correspondence to the maximum value of this statistic.

5 | SIMULATION STUDY

In this section, we describe the results of a simulation study conducted with the aim of investigating the ability of our multiple-choice LASSO procedure in discovering the true data generating model (DGM) among those in Figure 1.

For each of the following DGMs, we consider 50 randomly generated datasets, of size $n = 1000$, with $d = 2$ dimensions. The DGMs considered are: normal (N), t, Cauchy (C), Laplace (L), symmetric generalized hyperbolic (SGH), skew-t (St), variance gamma (VG), and asymmetric Laplace (AL). The DGMs share the same location parameter $\boldsymbol{\mu} = \mathbf{0}$ and scale matrix $\Sigma = \mathbf{I}$, with $\mathbf{I}$ denoting the identity matrix. We fix $\boldsymbol{\gamma} = (-0.5, 0.8)$ for the skewed DGMs (St, VG, and AL). Parameters λ, χ, and ψ vary according to the considered DGM; 1 provides the precise values of these parameters for each Table 1.

TABLE 1 Parameters λ, χ, and ψ of the DGMs used in the simulation study.

	DGM					
Parameter	**N**	***t*, S*t***	**C**	**L, AL**	**SGH**	**VG**
λ	−20	−1	−0.5	1	−1	1.5
χ	100	2	2	0.001	2	0.001
ψ	0.001	0.001	0.001	0.5	3	0.5

We use our penalized ML procedure on each generated dataset. We select the penalty parameter h with the LCV strategy described in Section 4.4, using the grid $G = \{0,5,10,15,20,25,30,35,40,45,50,60,70,80,100\}$ and a proportion $p = 0.1$ of observations which are allowed to be left out one at a time. On each generated dataset, we also compute the Akaike information criterion (AIC; Ref. [1]) and the Bayesian information criterion (BIC; Ref. [45]), and we do that separately for each of the 16 models in the GH family.

Table 2 shows the number of times the three competing approaches (our multiple-choice LASSO method, AIC, and BIC) select each model in our family. Here, there are some models that are fitted to the data but they are not used as DGMs; these models are the normal-inverse Gaussian (NIG), hyperbolic (H), hyperbolic univariate marginals (HUM), symmetric normal-inverse Gaussian (SNIG), symmetric variance gamma (SVG), symmetric hyperbolic (SH), skew-Cauchy (SC), and GH. Results are organized as a contingency table where the true DGM is given by column and the models in the GH-family by row. The shadowed blocks report the true positive count (TPC), measuring the number of times the AIC (on the top), the BIC (on the middle), and the multiple-choice LASSO approach (on the bottom), discover the true DGM. We can note how, regardless of the DGM, our approach is able enough to recognize the true underlying DGM, being the counts mainly concentrated on the shadowed cells. The best results are obtained for the t-DGM, where the TCP is the maximum possible (50). On the opposite side, the worst results are obtained for the N-DGM, where TPC = 42; in the remaining 8 cases, the more general skew-t distribution is selected. Instead, AIC and BIC do not perform so well. When the DGM is the normal one, the BIC performs like our method; instead, the AIC finds the true DGM only half the time (21 times). When the BIC does not select the true N-DGM, it picks the SNIG (5 times) and SVG (3 times) models. Moreover, it seems there are some DGMs AIC and BIC are not able to detect; they are the t, C, SGH, and St. In these cases, while our approach is able to recognize the truth 50, 49, 45, and 49 times, respectively, the corresponding counts for the AIC are 8, 0, 0, and 0, while for the BIC are 1, 0, 0, and 1. If we limit the attention to AIC and BIC

TABLE 2 Number of times AIC, BIC, and multiple-choice LASSO select each model. The true DGM is shown by column, while the models in the GH-family are given by row.

		DGM							
Fitted	**Method**	**N**	***t***	**C**	**L**	**SGH**	**S*t***	**AL**	**VG**
N	AIC	21	0	0	0	0	0	0	0
	BIC	42	0	0	0	0	0	0	0
	LASSO	42	0	0	0	0	0	0	0
t	AIC	0	8	0	0	10	8	0	0
	BIC	0	1	0	0	12	0	0	0
	LASSO	0	50	1	0	4	0	0	0
C	AIC	0	32	0	0	0	32	0	0
	BIC	0	49	0	0	0	49	0	0
	LASSO	0	0	49	0	0	0	0	0
L	AIC	0	0	0	18	0	0	0	0
	BIC	0	0	0	44	0	0	0	0
	LASSO	0	0	0	46	0	0	0	0
SGH	AIC	0	3	7	14	0	3	0	0
	BIC	0	0	0	2	0	0	0	0
	LASSO	0	0	0	0	45	0	0	0
S*t*	AIC	4	0	0	0	0	0	0	0
	BIC	0	0	0	0	0	1	0	0
	LASSO	8	0	0	0	0	49	0	0
AL	AIC	0	0	0	4	0	0	27	0
	BIC	0	0	0	0	0	0	48	0
	LASSO	0	0	0	0	0	0	44	0
VG	AIC	3	0	0	3	5	0	9	38
	BIC	0	0	0	0	0	0	0	39
	LASSO	0	0	0	0	0	0	3	48
NIG	AIC	0	0	8	0	4	0	0	0
	BIC	0	0	0	0	0	0	0	0
	LASSO	0	0	0	0	0	0	0	0
H	AIC	1	0	0	0	1	0	0	10
	BIC	0	0	0	0	0	0	0	11
	LASSO	0	0	0	0	0	0	0	1
HUM	AIC	0	0	0	3	7	0	0	0
	BIC	0	0	0	0	7	0	0	0
	LASSO	0	0	0	1	0	0	0	0
SNIG	AIC	10	0	35	0	8	0	0	0
	BIC	5	0	50	0	11	0	0	0
	LASSO	0	0	0	0	0	0	0	0
SVG	AIC	10	0	0	5	10	0	0	0
	BIC	3	0	0	4	12	0	0	0
	LASSO	0	0	0	3	1	0	0	0

TABLE 2 (Continued)

		DGM							
Fitted	**Method**	**N**	***t***	**C**	**L**	**SGH**	**S*t***	**AL**	**VG**
SH	AIC	1	0	0	0	5	0	0	0
	BIC	0	0	0	0	8	0	0	0
	LASSO	0	0	0	0	0	0	0	0
SC	AIC	0	3	0	0	0	3	0	0
	BIC	0	0	0	0	0	0	0	0
	LASSO	0	0	0	0	0	1	0	0
GH	AIC	0	4	0	3	0	4	14	2
	BIC	0	0	0	0	0	0	2	0
	LASSO	0	0	0	0	0	0	3	1

TABLE 3 Number of times AIC, BIC, and multiple-choice LASSO detect the true DGM.

	DGM									
	N	***t***	**C**	**L**	**SGH**	**S*t***	**AL**	**VG**	**Total**	**% Truth**
AIC	21	8	0	18	0	0	27	38	112	28.00
BIC	42	1	0	44	0	1	48	39	175	43.75
LASSO	42	50	49	46	45	49	44	48	373	93.25
# Replications	50	50	50	50	50	50	50	50	400	

Note: The true DGM is shown by column. Overall total count and overall true positive percentage are reported in the last two columns.

only, we can note, as expected, how the AIC tends to prefer less parsimonious models than the BIC. This can be noted, just as an example, when the DGM is the AL distribution. In this case, the BIC detects the truth 48 times, and the remaining two times it picks the most general GH model. Instead, the AIC detects the truth 27 times, the VG model 9 times, and the GH 14 times, with the AL being nested in the VG model (refer to Figure 1).

To have an overall look at the performance of the 3 competing methods (AIC, BIC, and LASSO), Table 3 reports the TPCs in the shadowed blocks of Table 2 along with the total count and true positive percentage (TPP) over the 400 replications. By looking at the TPPs, it is easy to realize how the LASSO approach performs very well (with a TPP of 93.25%); moreover, it works much better than the BIC (TPP = 43.75) which, in turn, works much better than the AIC (TPP = 28.00).

6 | DATA ANALYSIS

In finance, one of the main challenges is modeling the joint distribution of stock prices and asset returns. The considered models are inherently multivariate, as stressed by McNeil et al. [30], with the multivariate normal (MN) playing a special rule (Refs. [21, 42], Chapter 14). However, many empirical financial studies show that the MN distribution is not appropriate [15, 26]. A possible motivation for this inappropriateness concerns its thin tails [8], which are not consistent with the empirical heavy tails of the distribution of returns. This has motivated numerous proposals for alternative parametric multivariate heavy-tailed distributions. In this direction, the GH family of distributions represents one of the most famous and widespread proposals (Ref. [30], Chapter 6).

Motivated by the above considerations, we fit the GH distribution, and its special and limiting cases, on real financial data. The data we consider are related to two cryptocurrencies: TRON EUR (TRX-EUR) and FUNToken EUR (FUN-EUR). We downloaded the daily adjusted close prices (in Euro) from https://finance.yahoo.com/cryptocurrencies. The period under investigation goes from January 1, 2021–September 1, 2023. We work with daily log-returns, computed by taking logarithmic differences; this leads to $n = 973$ observations.

The scatter plot of the data is displayed in Figure 3. Table 4 provides some descriptive statistics (and Jarque-Bera normality tests) for the TRX-EUR and FUN-EUR series, separately considered. All the series are leptokurtic and the Jarque-Bera statistic confirms the

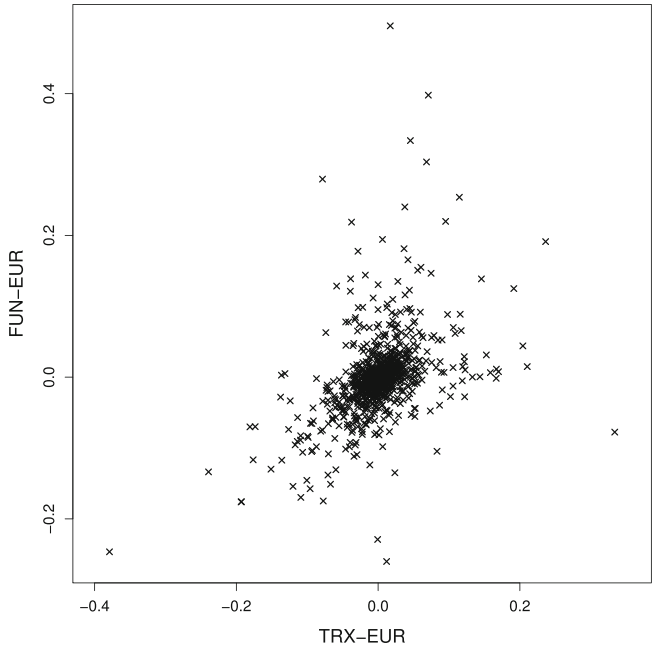


FIGURE 3 Scatter plot of daily log-returns of the pair (TRX-EUR,FUN-EUR) spanning the period from September 1, 2019–September 1, 2023.

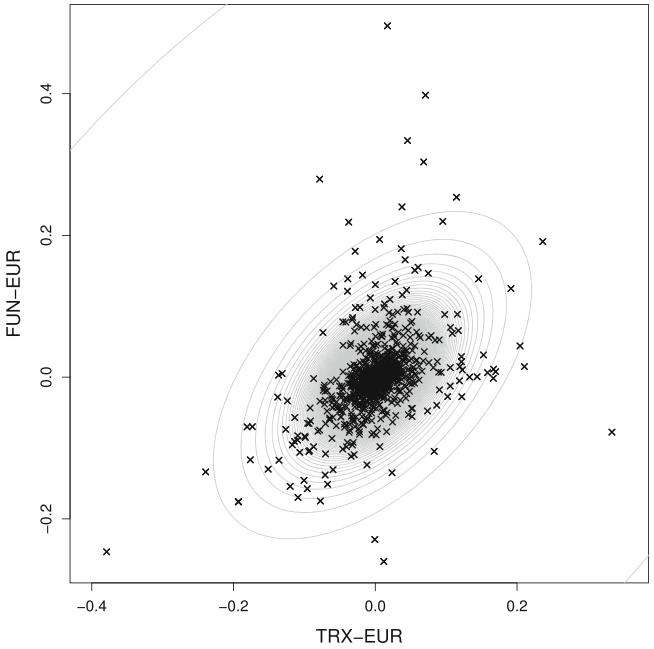


FIGURE 4 Scatterplot of TRX-EUR and FUN-EUR series, with isodensities from the LASSO-selected bivariate skew-*t* distribution.

departure from univariate normality at the 1‰ level. As concerns the dependence between cryptocurrencies, the scatter plot in Figure 3 displays a positive correlation between the series; such a sample correlation is 0.440, with a *p*-value from the test of uncorrelation (computed via the `cor.test()` function of the **stats** package) which is lower than 1‰. This confirms the need for a bivariate model allowing for correlation between series. The scatter plot also shows a certain degree of skewness and the need for a heavy-tailed bivariate distribution due to some outlying points.

We first proceed by fitting each of the sixteen models in Figure 1 separately. Having the competing models a differing number of parameters, we compare their goodness-of-fit, as usual, via AIC and BIC. AIC and BIC picked the symmetric GH and symmetric NIG, respectively, as the best model for the data. Instead, our penalized estimation method selected a NIG distribution. We selected the penalty parameter $h = 0.1$ with the LCV strategy described in Section 4.4, using the grid $G = \{0,0.1,0.2,0.3, \ldots ,9.9,10\}$ and a proportion $p = 0.05$ of observations which are allowed to be left out one at a time. However, as noted by looking at the scatter plot in Figure 3, the data seems to be skewed. To corroborate such a visual insight, we performed a test of elliptical symmetry. Among the various tests of elliptical symmetry available in the literature, we considered the MPQ test of Manzotti, Pérez, and Quiroz [27], which is implemented by the `MPQ()` function of the **ellipticalsymmetry** package [3]. We use this test because it preserves the claimed nominal significance level [2]. The resulting *p*-value is 0.032, suggesting the rejection of the null hypothesis of elliptical symmetry, the type of symmetry underlying the symmetric GH and NIG models, at the common 5% significance level. This is some evidence in favor of the choice of the NIG model, as chosen by our method. To further corroborate the choice, Figure 4 displays the scatter plot of the data together with the estimated isodensities for the bivariate NIG distribution selected by our approach.

7 | CONCLUDING REMARKS

In this work, we have put forward a taxonomy of the GH family, and showed how one can perform simultaneous

TABLE 4 Descriptive statistics, and Jarque-Bera normality tests (with *p*-values in brackets), for the TRX-EUR and FUN-EUR series.

	Mean	SD	Excess kurtosis	Jarque-Bera test (*p*-value)
TRX-EUR	0.001	0.047	10.851	4776.236 (<0.001)
FUN-EUR	0.000	0.055	15.170	9832.570 (<0.001)

estimation and selection of nested models within the family. We argue that the GH family is flexible enough to fit well a wide range of distributions in real applications, and that the model selection procedure is effective in providing a simple and interpretable model class without sacrificing goodness of fit. We also have introduced the multiple-choice LASSO. We believe adaptive choice of the shape parameters within the GH family is only one of the possible applications of the multiple-choice LASSO, and that its theoretical properties deserve further investigation. Given the strong connections between penalized methods and Bayesian approaches, and specifically the equivalence of LASSO and use of a Laplace prior (e.g., Ref. [22]), we can speculate that a similar representation should be available for the multiple-choice LASSO in the form of a prior mixture of Laplace distributions.

Additionally, there are other flexible and general parametric families of distributions that might benefit from an approach similar to the one proposed in this work (e.g., Ref. [16]).

ACKNOWLEDGMENTS

This study was partially supported/funded by: (i) MUR, grant number 2022XRHT8R—The SMILE project: Statistical Modeling and Inference to Live the Environment, and (ii) the European Union—NextGenerationEU, in the framework of the "GRINS—Growing Resilient, INclusive and Sustainable" project (GRINS PE00000018—CUP E63C22002120006). As for item (ii), the views and opinions expressed are solely those of the authors and do not necessarily reflect those of the European Union, nor can the European Union be held responsible for them.

CONFLICT OF INTEREST STATEMENT

The authors declare no potential conflict of interests.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Luca Bagnato https://orcid.org/0000-0003-3129-5385
Alessio Farcomeni https://orcid.org/0000-0002-7104-5826
Antonio Punzo https://orcid.org/0000-0001-7742-1821

REFERENCES

1. H. Akaike, *A new look at the statistical model identification*, IEEE Trans. Autom. Control 19 (1974), no. 6, 716–723.
2. S. Babić, C. Ley, and M. Palangetić. *Elliptical symmetry tests in* R (arXiv.org e-print No. 2011.12560v1). 2011 http://arxiv.org/abs/2011.12560v1.
3. Babić, S., Palangetić, M., & Ley, C. (2020). *ellipticalsymmetry*: Elliptical symmetry tests [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=ellipticalsymmetry R package version 0.1.
4. L. Bagnato and A. Punzo, *Unconstrained representation of orthogonal matrices with application to common principal components*, Comput. Stat. 36 (2021), no. 2, 1177–1195.
5. O. Barndorff-Nielsen, *Exponentially decreasing distributions for the logarithm of particle size*, Proc. R. Soc. Lond. A Math. Phys. Sci. 353 (1977), no. 1674, 401–419.
6. A. BenSaïda and S. Slim, *Highly flexible distributions to fit multiple frequency financial returns*, Phys. A Stat. Mech. Appl. 442 (2016), 203–213.
7. J. Bien, J. Taylor, and R. Tibshirani, *A LASSO for hierarchical interactions*, Ann. Stat. 41 (2013), 1111–1141.
8. N. H. Bingham, R. Kiesel, et al., *Semi-parametric modelling in finance: Theoretical foundations*, Quant. Finance 2 (2002), no. 4, 241–250.
9. J. R. Birge and L. Chavez-Bedoya, *Portfolio optimization under the generalized hyperbolic distribution: Optimal allocation, performance and tail behavior*, Quant. Finance 21 (2021), no. 2, 199–219.
10. R. P. Browne and P. D. McNicholas, *A mixture of generalized hyperbolic distributions*, Canadian J. Stat. 43 (2015), no. 2, 176–198.
11. C. R. B. Cabral, V. H. Lachos, and M. O. Prates, *Multivariate mixture modeling using skew-normal independent distributions*, Comput. Stat. Data Anal. 56 (2012), no. 1, 126–142.
12. D. R. Cox, *Role of models in statistical analysis*, Stat. Sci. 5 (1990), 169–174.
13. A. P. Dempster, N. M. Laird, and D. B. Rubin, *Maximum likelihood from incomplete data via the em algorithm*, J. R. Stat. Soc. Ser. B (Stat. Methodol.) 39 (1977), no. 1, 1–38.
14. E. Eberlein and U. Keller, *Hyperbolic distribution in finance*, Bernoulli 1 (1995), 281–299.
15. E. F. Fama, *Portfolio analysis in a stable Paretian market*, Manag. Sci. 11 (1965), no. 3, 404–419.
16. M. Geraci and A. Farcomeni, *A family of linear mixed-effects models using the generalized Laplace distribution*, Stat. Methods Med. Res. 29 (2020), 2665–2682.
17. W. Hu, *Calibration of multivariate generalized hyperbolic distributions using the em algorithm, with applications in risk management, portfolio optimization and portfolio credit risk*, The Florida State University, Florida, United States, 2005.
18. J. Kim and S. Lee, *A convenient approach for penalty parameter selection in robust lasso regression*, Commun. Stat. Appl. Methods 24 (2017), no. 6, 651–662.
19. S. I. Kim, *Arma–garch model with fractional generalized hyperbolic innovations*, Financ. Innov. 8 (2022), no. 1, 48.
20. T. J. Kozubowski and K. Podgórski, *A multivariate and asymmetric generalization of Laplace distribution*, Comput. Stat. 15 (2000), no. 4, 531–540.
21. S. Kring, S. T. Rachev, H. Markus, and F. Fabozzi, "*Estimation of α-stable sub-gaussian distributions for asset returns*," *Risk assessment: Decisions in banking and finance*, G. Bol, S. T. Rachev, and R. Würth Physica-Verlag, Eidelberg, 2008, pp. 111–152.
22. M. Kyung, J. Gill, M. Ghosh, and G. Casella, *Penalized regression, standard errors, and Bayesian lassos*, Bayesian Anal. 5 (2010), 369–412.

23. Lederer, J., & Müller, C. (2015). *Don't fall for tuning parameters: Tuning-free variable selection in high dimensions with the TREX*. In Proceedings of the AAAI Conference on Artificial Intelligence *(Vol. 29)* AAAI, Texas, United States.
24. M. Lim and T. Hastie, *Learning interactions via hierarchical group-lasso regularization*, J. Comput. Graph. Stat. 24 (2015), 627–654.
25. C. Liu and D. B. Rubin, *The ECME algorithm: A simple extension of EM and ECM with faster monotone convergence*, Biometrika 81 (1994), no. 4, 633–648.
26. B. Mandelbrot, *The variation of certain speculative prices*, J. Bus. 36 (1963), no. 4, 394–419.
27. A. Manzotti, F. J. Pérez, and A. J. Quiroz, *A statistic for testing the null hypothesis of elliptical symmetry*, J. Multivar. Anal. 81 (2002), no. 2, 274–285.
28. G. J. McLachlan and T. Krishnan, *The EM algorithm and extensions*, JohnWiley & Sons, New York, 2007.
29. A. J. McNeil, R. Frey, and P. Embrechts, *Quantitative risk management: Concepts, techniques and tools*, Princeton University Press, New Jersey, United States, 2005.
30. A. J. McNeil, R. Frey, and P. Embrechts, *Quantitative risk management: Concepts, techniques and tools—revised edition*, Princeton University Press, New Jersey, United States, 2015 https://books.google.it/books?id=l2yYDwAAQBAJ.
31. P. D. McNicholas, *Mixture model-based classification*, Chapman & Hall/CRC Press, Boca Raton, 2016.
32. V. Melnykov and X. Zhu, *On model-based clustering of skewed matrix data*, J. Multivar. Anal. 167 (2018), 181–194.
33. V. Melnykov and X. Zhu, *Studying crime trends in the USA over the years 2000–2012*, ADAC 13 (2019), no. 1, 325–341.
34. K. Morris and P. D. McNicholas, *Clustering, classification, discriminant analysis, and dimension reduction via generalized hyperbolic mixtures*, Comput. Stat. Data Anal. 97 (2016), 133–150.
35. K. Morris, A. Punzo, P. D. McNicholas, and R. P. Browne, *Asymmetric clusters and outliers: Mixtures of multivariate contaminated shifted asymmetric Laplace distributions*, Comput. Stat. Data Anal. 132 (2019), 145–166.
36. P. M. Murray, R. P. Browne, and P. D. McNicholas, *Mixtures of skew-t factor analyzers*, Comput. Stat. Data Anal. 77 (2014), 326–335.
37. T. Nitithumbundit and J. S. Chan, *Ecm algorithm for auto-regressive multivariate skewed variance gamma model with unbounded density*, Methodol. Comput. Appl. Prob. 22 (2020), no. 3, 1169–1191.
38. A. O'Hagan, T. B. Murphy, I. C. Gormley, P. D. McNicholas, and D. Karlis, *Clustering with the multivariate normal inverse gaussian distribution*, Comput. Stat. Data Anal. 93 (2016), 18–30.
39. D. Pal, *The distribution of commodity futures: A test of the generalized hyperbolic process*, Appl. Econ. (2023), 1–21.
40. M. J. D. Powell, "*A direct search optimization method that models the objective and constraint functions by linear interpolation*," *Advances in optimization and numerical analysis*, S. Gomez and J.-P. Hennart Kluwer Academic Publishers, Dordrecht, 1994, pp. 51–67.
41. A. Punzo and L. Bagnato, *The multivariate tail-inflated normal distribution and its application in finance*, J. Stat. Comput. Simul. 91 (2021), no. 1, 1–36.
42. S. T. Rachev, M. Hoechstoetter, F. J. Fabozzi, and S. M. Focardi, *Probability and statistics for finance*, John Wiley & Sons, New York, 2010 https://books.google.it/books?id=MT3aBgAAQBAJ.
43. R Core Team, *R: A language and environment for statistical computing [computer software manual]*, Austria, Vienna, 2020 https://www.R-project.org/.
44. P. A. Rowińska, A. E. Veraart, and P. Gruet, *A multi-factor approach to modeling the impact of wind energy on electricity spot prices*, Energy Econ. 104 (2021), 105640.
45. G. Schwarz, *Estimating the dimension of a model*, Ann. Stat. 6 (1978), no. 2, 461–464.
46. M. Stone, *Cross-validatory choice and assessment of statistical predictions*, J. R. Stat. Soc. Ser. B (Methodological) 36 (1974), no. 2, 111–133.
47. R. Tibshirani, *Regression shrinkage and selection via the LASSO*, J. R. Stat. Soc. Ser. B 58 (1996), 267–288.
48. X. Wu and X. Zhou, *On Hodges' superefficiency and merits of oracle property in model selection*, Ann. Inst. Stat. Math. 71 (2019), 1093–1119.
49. M. Yuan and Y. Lin, *Model selection and estimation in regression with grouped variables*, J. R. Stat. Soc. Ser. B 68 (2006), 49–67.
50. Y. Zhang, J. Chu, S. Chan, and B. Chan, *The generalised hyperbolic distribution and its subclass in the analysis of a new era of cryptocurrencies: Ethereum and its financial risk*, Phys. A Stat. Mech. Appl. 526 (2019), 120900.

How to cite this article: L. Bagnato, A. Farcomeni, and A. Punzo, *The generalized hyperbolic family and automatic model selection through the multiple-choice LASSO*, Stat. Anal. Data Min.: ASA Data Sci. J. **17** (2024), e11652. https://doi.org/10.1002/sam.11652

APPENDIX A. GENERALIZED INVERSE GAUSSIAN DISTRIBUTION

The random variable W has a GIG distribution if its pdf is

$$f_{\text{GIG}}(w;\lambda,\chi,\psi)=\left(\frac{\psi}{\chi}\right)^{\frac{\lambda}{2}}\frac{w^{\lambda-1}}{2K_{\lambda}(\sqrt{\psi\chi})}\exp\left[-\frac{1}{2}\left(\psi w+\frac{\chi}{w}\right)\right],\quad w>0,\quad \text{(A1)}$$

where the parameters satisfy the conditions: $\chi>0$ and $\psi\geq 0$, if $\lambda<0$; $\chi>0$ and $\psi>0$, if $\lambda=0$; $\chi\geq 0$ and $\psi>0$, if $\lambda>0$. If W has the pdf in (A1), then we simply write $W\sim\mathcal{GIG}(\lambda,\chi,\psi)$. The expectations of W and $1/W$, used in Section 4.1, are

$$\text{E}(W)=\sqrt{\frac{\chi}{\psi}}\frac{K_{\lambda+1}\left(\sqrt{\psi\chi}\right)}{K_{\lambda}\left(\sqrt{\psi\chi}\right)}\quad \text{(A2)}$$

and

$$\mathrm{E}\left(\frac{1}{W}\right) = \sqrt{\frac{\psi}{\chi}}\frac{K_{\lambda+1}\left(\sqrt{\psi\chi}\right)}{K_{\lambda}\left(\sqrt{\psi\chi}\right)} - \frac{2\lambda}{\chi}. \tag{A3}$$

APPENDIX B. SPECIAL AND LIMITING CASES OF THE GH DISTRIBUTION

B.1 GH → Skew-t → t → Gaussian

If $\lambda < 0$ and $\psi \to 0$, then $W \sim \mathcal{GIG}(\lambda, \chi, \psi)$ tends to $W \sim \mathcal{IG}\left(-\lambda, \frac{\chi}{2}\right)$, where $\mathcal{IG}(\cdot)$ denotes the inverse gamma distribution. Therefore, the NMVM representation in (2) becomes

$$\boldsymbol{X} = \boldsymbol{\mu} - V\frac{\chi}{2\lambda}\boldsymbol{\gamma} + \sqrt{V}\overline{\boldsymbol{U}},$$

where $V = -\frac{2\lambda}{\chi}W \sim \mathcal{IG}(-\lambda, -\lambda)$ and $\overline{\boldsymbol{U}} \sim \mathcal{N}_d\left(\boldsymbol{0}, -\frac{\chi}{2\lambda}\Sigma\right)$, with $|\Sigma| = 1$. Note that, thanks to the multiplicative factor $-\chi/(2\lambda)$, $|\mathrm{Cov}(\overline{\boldsymbol{U}})| = [-\chi/(2\lambda)]^d|\Sigma| = [-\chi/(2\lambda)]^d$ can be any positive real number. Under this setting, $\boldsymbol{X} \sim \mathcal{S}t_d\left(\boldsymbol{\mu}, -\frac{\chi}{2\lambda}\Sigma, -\frac{\chi}{2\lambda}\boldsymbol{\gamma}, -2\lambda\right)$, which represents a skew-t distribution with location parameter $\boldsymbol{\mu}$, scale matrix $-\frac{\chi}{2\lambda}\Sigma$, skewness parameter $-\frac{\chi}{2\lambda}\boldsymbol{\gamma}$, and $\nu = -2\lambda$ degrees of freedom [17, 36]. Compared to the GH-parametrization adopted by McNicholas [31], in our case, because of the identifiability constraint $|\Sigma| = 1$, there is no reason to force χ and λ to be related as $\chi = \nu = -2\lambda$. In other words, with our parametrization, χ is unconstrained. Indeed, if we impose the constraint $\chi = \nu = -2\lambda$ with our parametrization, then we would get $|\mathrm{Cov}(\overline{\boldsymbol{U}})| = 1$. If, in addition, $\boldsymbol{\gamma} = \boldsymbol{0}$, then $\boldsymbol{X} \sim t_d\left(\boldsymbol{\mu}, -\frac{\chi}{2\lambda}\Sigma, -2\lambda\right)$, which represents a t distribution with location parameter $\boldsymbol{\mu}$, scale matrix $-\frac{\chi}{2\lambda}\Sigma$, and $\nu = -2\lambda$ degrees of freedom. Finally, if we further consider $\lambda = -\chi/(2c)$, with $c > 0$, and $\chi \to \infty$, then we obtain $\boldsymbol{X} \sim \mathcal{N}_d(\boldsymbol{0}, c\Sigma)$ as a limiting case.

B.2 GH → Variance Gamma → Asymmetric Laplace → Laplace

If $\lambda > 0$ and $\chi \to 0$, then $W \sim \mathcal{GIG}(\lambda, \chi, \psi)$ tends to $W \sim \mathcal{G}\left(\lambda, \frac{\psi}{2}\right)$, where $\mathcal{G}(\cdot)$ denotes the gamma distribution. Then, the NMVM representation in (2) becomes

$$\boldsymbol{X} = \boldsymbol{\mu} + V\frac{\psi}{2\lambda}\boldsymbol{\gamma} + \sqrt{V}\overline{\boldsymbol{U}},$$

where $V = \frac{2\lambda}{\psi}W \sim \mathcal{G}(\lambda, \lambda)$ and $\overline{\boldsymbol{U}} \sim \mathcal{N}_d\left(\boldsymbol{0}, \frac{\psi}{2\lambda}\Sigma\right)$, with $|\Sigma| = 1$. Note that, thanks to the multiplicative factor $\psi/(2\lambda)$, $|\mathrm{Cov}(\overline{\boldsymbol{U}})| = [\psi/(2\lambda)]^d|\Sigma| = [\psi/(2\lambda)]^d$ can be any positive real number. Under this setting, $\boldsymbol{X} \sim \mathcal{VG}_d\left(\boldsymbol{\mu}, \frac{\psi}{2\lambda}\Sigma, \frac{\psi}{2\lambda}\boldsymbol{\gamma}, \lambda\right)$, which represents a variance gamma distribution with location parameter $\boldsymbol{\mu}$, scale matrix $\frac{\psi}{2\lambda}\Sigma$, skewness parameter $\frac{\psi}{2\lambda}\boldsymbol{\gamma}$, and shape parameter λ [37]. Compared to the VG-parametrization adopted by [37] and McNicholas [31], in our case, because of the identifiability constraint $|\Sigma| = 1$, there is no reason to force ψ and λ to be related as $\psi = 2\lambda$. In other words, with our parametrization, ψ is unconstrained. Indeed, if we impose the constraint $\psi = 2\lambda$ with our parametrization, then we would get $|\mathrm{Cov}(\overline{\boldsymbol{U}})| = 1$. If, in addition, $\lambda = 1$, then $V \sim \mathcal{E}(1)$, which is a standard exponential distribution, and $\boldsymbol{X} \sim \mathcal{AL}_d\left(\boldsymbol{\mu}, \frac{\psi}{2}\Sigma, \frac{\psi}{2}\boldsymbol{\gamma}\right)$, which represents an asymmetric Laplace distribution with location parameter $\boldsymbol{\mu}$, scale matrix $\frac{\psi}{2}\Sigma$, and skewness parameter $\frac{\psi}{2}\boldsymbol{\gamma}$; see Kozubowsi and Podgórski [20] and Morris, Punzo, McNicholas, and Browne [35]. Finally, if we further consider $\boldsymbol{\gamma} = \boldsymbol{0}$, then $\boldsymbol{X} \sim \mathcal{L}_d\left(\boldsymbol{\mu}, \frac{\psi}{2}\Sigma\right)$, which represents a Laplace distribution with location parameter $\boldsymbol{\mu}$ and scale matrix $\frac{\psi}{2}\Sigma$; see Kozubowski and Podgórski [20].

B.3 GH → Normal-Inverse Gaussian → Skew-Cauchy → Cauchy

If $\lambda = -1/2$, then $\boldsymbol{X} \sim \mathcal{NIG}_d(\boldsymbol{\mu}, \Sigma, \boldsymbol{\gamma}, \chi, \psi)$, which denotes the normal-inverse Gaussian distribution with location parameter $\boldsymbol{\mu}$, scale matrix Σ, skewness parameter $\boldsymbol{\gamma}$, and concentration parameters χ and ψ [38]. If, in addition, $\psi \to 0$, then $\boldsymbol{X} \sim \mathcal{SC}_d(\boldsymbol{\mu}, \chi\Sigma, \chi\boldsymbol{\gamma})$, which represents the skew-Cauchy distribution with location parameter $\boldsymbol{\mu}$, scale matrix $\chi\Sigma$, and skewness parameter $\chi\boldsymbol{\gamma}$ [11]. Note that, $\mathcal{SC}_d(\boldsymbol{\mu}, \chi\Sigma, \chi\boldsymbol{\gamma})$ can be also obtained as a special case of $\mathcal{S}t_d\left(\boldsymbol{\mu}, -\frac{\chi}{2\lambda}\Sigma, -\frac{\chi}{2\lambda}\boldsymbol{\gamma}, -2\lambda\right)$ when $\lambda = -1/2$; refer to Section B.1. Finally, if we further consider $\boldsymbol{\gamma} = \boldsymbol{0}$, then $\boldsymbol{X} \sim \mathcal{C}_d(\boldsymbol{\mu}, \chi\Sigma)$, which represents a Cauchy distribution with location parameter $\boldsymbol{\mu}$ and scale matrix $\chi\Sigma$.