Advanced Applications of Generalized Hyperbolic Distributions in Portfolio Allocation and Measuring Diversification

A Dissertation presented

by

Xiang Shi

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Applied Mathematics and Statistics

(Quantitative Finance)

Stony Brook University

May 2016

Stony Brook University

The Graduate School

Xiang Shi

We, the dissertation committe for the above candidate for the

Doctor of Philosophy degree, hereby recommend

acceptance of this dissertation

Aaron Kim
Assistant Professor, College of Bussiness

Svetlozar Rachev
Research Professor, Department of Applied Mathematics and Statistics

Raphael Douady
Professor, Department of Applied Mathematics and Statistics
length.

Keli Xiao
Assistant Professor, College of Bussiness

This dissertation is accepted by the Graduate School

Charles Taber
Dean of the Graduate School

Abstract of the Dissertation

Advanced Applications of Generalized Hyperbolic Distributions in Portfolio Allocation and Measuring Diversification

by

Xiang Shi

Doctor of Philosophy

in

Applied Mathematics and Statistics

(Quantitative Finance)

Stony Brook University

2016

This thesis consists of two parts. The first part addresses the parameter estimation and calibration of the Generalized Hyperbolic (GH) distributions. In this part we review the classical expectation maximization (EM) algorithm and factor analysis for the GH distribution. We also propose a simple shrinkage estimator driven from the penalized maximum likelihood. In addition an on-line EM algorithm is implemented to the GH distribution; and its regret for general exponential family can be represented as a mixture of Kullback-Leibler divergence. We compute the Hellinger distance of the joint GH distribution to measure the performances of all the estimators numerically. Empirical studies for long-term and short-term predictions are also performed to evaluate the algorithms.

In the second part we applied the GH distribution to portfolio optimization and risk allocation. We show that the mean-risk portfolio optimization problem of a certain type of normal mixture distributions including the GH distribution can be reduced to a two dimensional problem by fixing the location parameter and the skewness parameter. In addition, we show that the

efficient frontier of the mean-risk optimization problem can be extended to the three dimensional space. We also proposed a simple algorithm to deal with the transaction costs. The first and second derivatives of the CVaR are computed analytically when the underlying distribution is GH. With these results we are able to extend the effective number of bets (ENB) to general risk measures with the GH distribution. By diagonalizing the Hessian matrix of a risk measure we are able to extract locally independent marginal contributions to the risk. The minimal torsion approach can still be applied to get the local coordinators of the marginal contributions.

To the memory of my grandmother

Contents

List of Figures

List of Tables

List of Abbreviations

AD	Anderson-Darling (test)
EM	Expectation maximization
ENB	Effective number of bets
GH	Generalized hyperbolic distribution
GIG	Generalized inverse Gaussian distribution
IG	Inverse Gaussian distribution
KS	Kolmogorov-Smirnov (test)
MCECM	Multi-cycle expectation conditional maximiza- tion
MLE	Maximum likelihood estimator
NIG	Normal inverse Gaussian distribution
NTS	Normal tempered stable distribution
VG	Variance gamma distribution

Acknowledgements

I would first like to thank my advisor Prof. Aaron Kim for his consistent support of my research throughout my 5 years' study. Without his knowledge, guidance and responsibility I would not be able to finish this dissertation. I am very fortunate to be able to work with Prof. Rachev, a world-class expert in heavy-tailed distributions. I am always grateful for his valuable advices and challenging research topics in financial modeling and risk management. I also want to thank Prof. Andrew Mullhaupt for his deep insights in statistics, information theory and portfolio optimization, and above all his excellent lectures in matrix analysis and linear time invariant system. I would like to thank Dr. Attilio Meucci for inventing the effective number of bets and his wonderful ARPM bootcamp which helps me to see the big picture of modern quantitative finance. I am also grateful to Prof. James Glimm, Prof. Raphael Douady, Prof. Keli Xiao, Prof. Haipeng Xing and Prof Jiaoqiao Hu, from whom I learnt a lot about mathematics and finance.

I also want to express my gratitude to my colleagues at Stony Brook University. They are not limited to followings: Yu Mu, Rong Lin, Tiantian Li, Ke Zhang, Xiaoping Zhou, Tianyu Lu, Xiao Yu, Riyu Yu, Fangfei Dong, Yuzhong Zhang, Barret Shao, Hua Mo, Jianzhao Yang, Xu Dong, Tengjie Jia, Jaehyung Choi, Naoshi Tsuchida, Ning Ma, Chi Kong, Jiazhou Wang, Ruibo Yang and Si Wen. Above all, I would like to thank Yikang Chai, Angela Tsao and Lihua Zhang whom I have closely worked with on several research projects, for their inspiration, encouragement and insightful discussions.

Finally, my most sincere thanks go to my family. I could not imagine that I can make any accomplishment without their love and support.

Vita, Publications and/or Fields of Study

I received B.S in applied math at Shanghai University in 2009 and M.Sc in mathematics and finance at Imperial College London in 2011. I was admitted to the Ph.D program in applied math at Stony Brook University in 2011. My research concentrates on quantitative finance, especially on portfolio optimization, risk management and heavy-tailed distributions. The list of publications and working papers is following:

1. Xiang Shi, Aaron Kim. Coherent Risk Measure and Normal Mixture Distributions with Application in Portfolio Optimization and Risk Allocation. SSRN 2548057 (2015).

2. Xiang Shi. Marginal Contribution to Risk and Generalized Effective Number of Bets. SSRN 2642408 (2015).

3. Xiang Shi, Lihua Zhang, Aaron Kim. A Markov Chain Approximation for American Option Pricing in Tempered Stable-GARCH Models. Frontiers in Applied Mathematics and Statistics. (2015)

4. Angela Tsao, Xiang Shi, Alexander Melnikov. CVaR hedging under stochastic interest rate. Frontiers in Applied Mathematics and Statistics. (2015).

1 Introduction

The quantitative financial modeling can be traced back to [24] who assumes that the stock price follows the Brownian motion with zero drift. The most influential option pricing theory developed by [8] applies the geometric Brownian motion to model stock price; and derive the famous Black-Scholes formula by deducting a unique risk-neutral measure. In the geometric Brownian motion stock price model, the log returns are assumed to be normally or Gaussian distributed. However, this assumption was rejected by the empirical studies from [25], who conjectured that the log returns of most financial instruments are well described by a class of stable distributions. Further solid evidences were discovered by [16]. Stable distributions can be viewed as a generalization of the normal distribution; but they are heavy-tailed in nature. The "tails" of a distribution can be viewed as the probability of the occurrence of extreme values. It can be measured by kurtosis of a distribution. And a continuous probability distribution is heavy-tailed if its kurtosis is greater than the one of the normal distribution.

The major fallacy of the normal distribution in finance is that it wrongly underestimates the frequency of extreme events such as financial crisis. It is well-known that the geometric Brownian motion based Black-Scholes formula failed on Black Monday, October 19, 1987. Today it is widely recognized that the financial time series has three important stylized facts: (i) they have heavy tails; (ii) they are skewed; (iii) they exhibits volatility clustering. There are branch of studies in heavy-tailed stable distributions and their application to finance, see [33], [35], [38], [37] and [46]. However a drawback of stable distributions is that they do not have the second moment, or equivalently, finite variance. Some stable distributions do not have even the first moment. Thus they may be too heavy-tailed to fit finance data properly. A simple way to fix this problem to truncated the tails of stable distributions a bit, see [29] for example. Another more mathematically beautiful approach is to multiply the Lévy measure of stable distributions by an function with exponential decay. The new distribution constructed by this approach is called the tempered stable distribution. We refer readers to [42] for detailed construction of the tempered stable distribution. [22], [36] and [41] applied the tempered stable distribution with time series models to option pricing.

Despite the fact that financial market exhibits heavy-tails, the normal distribution is still the most widely-used distribution in financial industry and literatures, for two major reasons. First, there is a natural multivari-

ate version of the normal distribution; and the linear transformation of any multivariate normally distributed random vectors still follows the normal distribution. Copula is often used to model the dependence structures of heavy-tailed distributions that lacks multivariate extension. However the copula-based structure would often be destroyed by linear transformation. For many heavy-tailed distributions, problems such as portfolio allocation, stress testing, measuring diversification and risk contribution can only be done via Monte Carlo numerically; while the normal distribution can provide analytical solutions to most of these problems.

Secondly, the normal distribution has better statistical properties. It belongs to the exponential family, so distance measures like the Kullback-Leibler divergence or Hellinger distance are trivial. Unfortunately, an exponential family defined on the whole real line cannot be heavy-tailed. Some heavy-tailed distributions such as stable or tempered stable distribution do not even have analytical representations of their density functions, which can only be computed numerically via fast Fourier transform (FFT), see [28], [32] and [7] for example. Furthermore, traditional unbiased estimators usually have a poor performance in modeling high dimensional noisy financial data. A branch of biased estimators have been invented for parameter estimation for Gaussian models. Typical examples include the principle component analysis (PCA); factor analysis with the expectation maximization (EM) algorithm; Lasso regression proposed by [44]; James-Stein shrinkage estimator proposed by [19] and Ledoit-Wolf shrinkage estimator proposed by [23]. Applying these techniques to heavy-tailed distributions is not trivial and often numerically intractable.

In order to solve the first problem, a class of normal mixture distributions are introduced to financial modeling. The idea is to multiply a Gaussian random vector by an independent positive heavy-tailed random number, so that the mixture distribution is closed under linear transformation, and still has heavier tails than the normal distribution. The positive heavy-tailed random number is sometimes called a subordinator, since it corresponds to the randomized time under stochastic process framework. For example, if the subordinator follows the inverse Gaussian (IG) distribution, then the mixture distribution is called the normal inverse Gaussian (NIG) distribution, see [6]. Its continuous time counter-party is sometimes called the Carr-Geman-Madan-Yor (CGMY) process proposed by [10]. A generalization of the NIG distribution is called the normal tempered stable (NTS) distribution, whose subordinator is given by a positive tempered stable random number. [21]

studies the portfolio optimization problem based on the NTS distribution. The IG distribution is also a special case of the generalized inverse Gaussian (GIG) distribution. The mixture distribution with a GIG subordinator is called the generalized hyperbolic (GH) distribution, first introduced by [5]. The applications of the GH distribution to finance can be found in [13], [14] and [18]. The skewed multivariate t distribution and the variance gamma (VG) distribution are the limiting cases of the GH distribution.

The goal of this paper is to address the second problem: efficient parameter estimation, portfolio optimization and risk allocation approaches for normal mixture distributions. We choose the GH distribution as an example to illustrate these approaches for two reasons. First, the majority of normal mixture distributions applied to finance can be viewed as subclasses of the GH distribution. Secondly, although the GH distribution does not belong to exponential family itself; it is the marginal distribution of a strict exponential family. Thus it shares some nice statistical properties with exponential families.

This paper is organized as follows. In chapter 2 we review the definitions and statistical properties of the GIG and the GH distribution. In chapter 3 we first review the expectation maximization (EM) algorithm for the GH distribution and the regularization of the GH parameters. Then we introduce three potentially advanced approaches: a shrinkage approach with penalized likelihood, the factor analysis and the on-line EM algorithm. These approaches are tested both numerically and empirically using U.S equity data. In the last chapter we investigate the general mean-risk portfolio optimization problem with normal mixture distributions. We find that Markowitz's mean-variance efficient frontier can be extended to the three dimensional space. Then we compute the first and second derivatives of the conditional value-at-risk (CVaR) for normal mixture distributions. The results can be applied to a fast mean-CVaR portfolio algorithm with transaction costs; and measuring portfolio diversification based on the generalized effective number of bets (ENB) proposed by [40].

2 Preliminary

2.1 The Generalized Inverse Gaussian Distributions

In this section we review the definition and statistical properties of the generalized inverse Gaussian (GIG) distributions.

Definition 1. *The generalized inverse Gaussian distribution is a continuous probability distribution with the density function:*

$$p(y|\lambda,\chi,\psi) = \frac{(\psi/\chi)^{\frac{\lambda}{2}}}{2K_\lambda(\sqrt{\chi\psi})} y^{\lambda-1} \exp\Big(-\frac{1}{2}(\chi y^{-1} + \psi y)\Big), y > 0, \qquad (1)$$

where $K_\lambda(\cdot)$ is the modified Bessel function of the second kind and the parameters (λ,χ,ψ) satisfies:

$$\begin{cases} \chi > 0, \psi \geq 0, \text{ if } \lambda < 0 \\ \chi > 0, \psi > 0, \text{ if } \lambda = 0 \\ \chi \geq 0, \psi > 0, \text{ if } \lambda > 0 \end{cases}.$$

Throughout this paper we assume that $\chi > 0$ and $\psi > 0$ for simplicity. Another useful way to parameterize the GIG distribution is to set $\delta = \sqrt{\chi/\psi}$, $\eta = \sqrt{\chi\psi}$. In that case the density function can be written as:

$$p(y|\lambda,\delta,\eta) = \frac{\delta^\lambda}{2K_\lambda(\eta)} y^{\lambda-1} \exp\Big(-\frac{\eta}{2}(\delta y^{-1} + \delta^{-1} y)\Big), y > 0. \qquad (2)$$

Without the loss of generality we will denote the triple (λ,χ,ψ) for the GIG distribution with density (1) and (λ,δ,η) for (2). It is easy to compute the moment generating function of a GIG distributed random variable Y is given by:

$$E[e^{uY}] = \left(\sqrt{\frac{\psi}{\psi-2u}}\right)^\lambda \frac{K_\lambda(\sqrt{\chi(\psi-2u)})}{K_\lambda(\sqrt{\chi\psi})} = \left(\sqrt{\frac{\eta}{\eta-2\delta u}}\right)^\lambda \frac{K_\lambda(\sqrt{\eta^2-2\delta u})}{K_\lambda(\eta)}.$$

From the second equation one can observe that δ is served as a scale parameter of the GIG distribution. This is even clear when we look at the moments of Y:

$$E[Y^\alpha] = \left(\sqrt{\frac{\chi}{\psi}}\right)^\alpha \frac{K_{\lambda+\alpha}(\sqrt{\chi\psi})}{K_\lambda(\sqrt{\chi\psi})} = \delta^\alpha \frac{K_{\lambda+\alpha}(\eta)}{K_\lambda(\eta)}. \qquad (3)$$

The GIG distribution is an exponential family with natural parameters (λ, χ, ψ). On the other side, the corresponding expectation parameters are given by:

$$\begin{aligned} s_1 &= E[Y^{-1}] = \sqrt{\frac{\psi}{\chi}} \frac{K_{\lambda-1}(\sqrt{\chi\psi})}{K_\lambda(\sqrt{\chi\psi})}, \\ s_2 &= E[Y] = \sqrt{\frac{\chi}{\psi}} \frac{K_{\lambda+1}(\sqrt{\chi\psi})}{K_\lambda(\sqrt{\chi\psi})}, \\ s_3 &= E[\log(Y)] = \frac{\partial}{\partial\alpha} E[Y^\alpha]|_{\alpha=0}. \end{aligned} \tag{4}$$

Unfortunately we do not have an analytical formula for s_3. In practice it can be only approximated numerically. On the other side, given (s_1, s_2, s_3) computing (λ, χ, ψ) by solving the above equations is proved to be a hard problem. First note that computing the natural parameters from the expectation parameters of an exponential family is basically the same as computing the maximum likelihood given sufficient statistics. Let $y_1, y_2, \ldots, y_n$ be a sequence of sample data, then the maximum likelihood estimator (MLE) of GIG are given by:

$$(\hat{\lambda}, \hat{\chi}, \hat{\psi}) = \arg\max_{\lambda,\chi,\psi} L_{GIG}(\lambda, \chi, \psi | \hat{s}_1, \hat{s}_2, \hat{s}_3), \tag{5}$$

where L_{GIG} is the log-likelihood function excluding constants:

$$L_{GIG}(\lambda, \chi, \psi | s_1, s_2, s_3) := \tag{6}$$

$$-\frac{1}{2}\chi\hat{s}_1 - \frac{1}{2}\psi\hat{s}_2 + \lambda\hat{s}_3 + \frac{\lambda}{2}\log(\psi/\chi) - \log(K_\lambda(\sqrt{\chi\psi})), \tag{7}$$

and $\hat{s}_1 = \frac{1}{n}\sum_{k=1}^n y_k^{-1}$, $\hat{s}_2 = \frac{1}{n}\sum_{k=1}^n y_k$ and $\hat{s}_3 = \frac{1}{n}\sum_{k=1}^n \log(y_k)$. One can check that the optimal solution $(\hat{\lambda}, \hat{\chi}, \hat{\psi})$ must satisfies (4) where (s_1, s_2, s_3) are replaced by $(\hat{s}_1, \hat{s}_2, \hat{s}_3)$.

As far as we know, there is no analytical expression of $\hat{\lambda}$ or even its partial derivatives. Most literatures, for example [20], suggests to fix λ when we maximize the log likelihood function. Even λ is fixed, [18] reports that when $|\lambda|$ is large, say above 10, there might be no solution for the first two equations in (4).

To test the maximum likelihood approach we first need to find a proper way to measure the estimation errors. One good choice is to Hellinger distance between the true and estimated parameters.

Proposition 1. *Let* $(\lambda_1, \chi_1, \psi_1)$ *and* $(\lambda_2, \chi_2, \psi_2)$ *be the parameters of two GIG distributions. The squared Hellinger distance between the two distributions is given by:*

$$H^2_{GIG}(\lambda_1, \chi_1, \psi_1 \| \lambda_2, \chi_2, \psi_2) = 1 - \frac{(\psi_1/\chi_1)^{\frac{\lambda_1}{4}}(\psi_2/\chi_2)^{\frac{\lambda_2}{4}}}{\sqrt{K_{\lambda_1}(\sqrt{\chi_1\psi_1})K_{\lambda_2}(\sqrt{\chi_2\psi_2})}} \frac{K_{\bar{\lambda}}(\sqrt{\bar{\chi}\bar{\psi}})}{(\bar{\psi}/\bar{\chi})^{\frac{\bar{\lambda}}{2}}},$$

where $\bar{\lambda} = \frac{1}{2}\lambda_1 + \frac{1}{2}\lambda_2$, $\bar{\chi} = \frac{1}{2}\chi_1 + \frac{1}{2}\chi_2$ *and* $\bar{\psi} = \frac{1}{2}\psi_1 + \frac{1}{2}\psi_2$.

Proof. The Hellinger affinity is given by:

$$\begin{aligned}
&\int_0^\infty \sqrt{p(y|\lambda_1, \delta_1, \eta_1)p(y|\lambda_2, \delta_2, \eta_2)}dy \\
&= \frac{(\psi_1/\chi_1)^{\frac{\lambda_1}{4}}(\psi_2/\chi_2)^{\frac{\lambda_2}{4}}}{2\sqrt{K_{\lambda_1}(\sqrt{\chi_1\psi_1})K_{\lambda_2}(\sqrt{\chi_2\psi_2})}} \int_0^\infty y^{\bar{\lambda}-1} \exp\Big(-\frac{1}{2}(\bar{\chi}y^{-1} + \bar{\psi}y)\Big)dy \\
&= \frac{(\psi_1/\chi_1)^{\frac{\lambda_1}{4}}(\psi_2/\chi_2)^{\frac{\lambda_2}{4}}}{\sqrt{K_{\lambda_1}(\sqrt{\chi_1\psi_1})K_{\lambda_2}(\sqrt{\chi_2\psi_2})}} \frac{K_{\bar{\lambda}}(\sqrt{\bar{\chi}\bar{\psi}})}{(\bar{\psi}/\bar{\chi})^{\frac{\bar{\lambda}}{2}}}.
\end{aligned}$$

□

Now we are ready to analyze the optimization problem (5). In order to see how ill-conditioned it might be, we first set $\lambda = -10$, $\chi = 1$ and $\psi = 10^{-5}$. Then we compute the expectation parameters using (4) and get: $s_1 = 20.0000$, $s_2 = 0.0556$ and $s_3 = -2.9449$. Instead of fixing λ, we use Matlab function `fmincon` together with the interior-point algorithm to solve (4) directly. The partial derivatives of χ and ψ have analytical expressions while the derivative of λ is computed numerically. The results are given by Table 1.

	True parameters	Estimated parameters	Relative errors
λ	-10	-9.9538	0.0046
χ	1	0.9975	0.0025
ψ	10^{-5}	0.7612	7.6119×10^4

Table 1: Relative errors of the GIG parameters

We can observe that the estimated λ and χ are quite accurate, but ψ is not even close to the true one. However the log-likelihoods of the two sets of

parameters are almost the same, which is about -3.3537. In fact the difference are approximately 10^{-7}. This fact indicts that the log-likelihood function is very flat with respect to ψ. Figure 1 plots the log-likelihood function against ψ when $\lambda = -10$ and $\chi = 1$. The x-axis is plotted in the log scale.

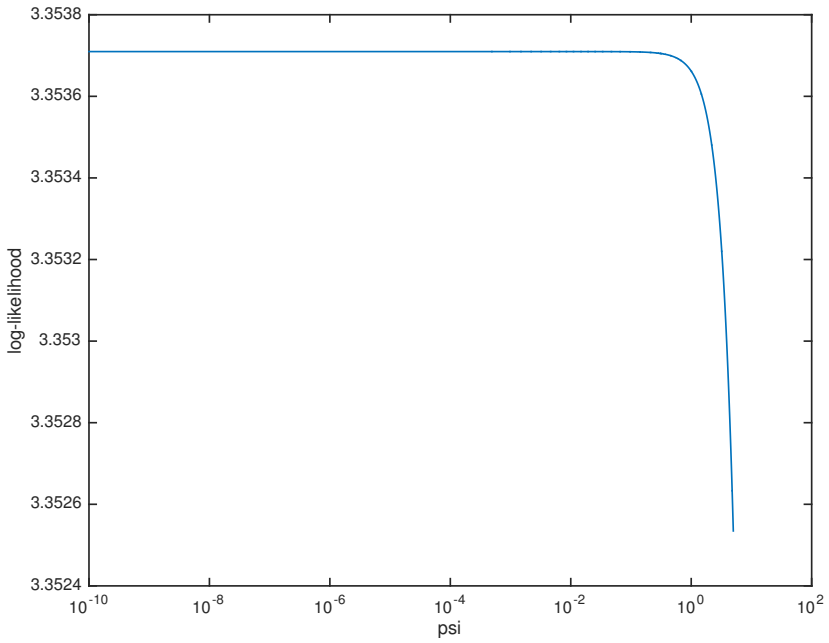


Figure 1: Log-likelihood function of GIG

The Hellinger distance between two sets of parameters are also small, which is approximately 6.1809×10^{-8}. This implies that the two distributions are almost the same.

It is easy to explain this phenomena. Let us write $\theta_k = (\lambda_k, \chi_k, \psi_k), k = 1, 2$, $\bar{\theta} = (\theta_1 + \theta_2)/2$ and $s = (s_1, s_2, s_3)$ for simplicity. Note that the GIG distribution is an exponential family, it is easy to see that the squared Hellinger distance satisfies:

$$H^2_{GIG}(\theta_1 \| \theta_2) = 1 - \exp\left(\frac{L_{GIG}(\theta_1|s) + L_{GIG}(\theta_2|s)}{2} - L_{GIG}(\bar{\theta}|s) \right),$$

for any choice of s. Since L_{GIG} is concave, we always have $(L_{GIG}(\theta_1|s) + L_{GIG}(\theta_2|s))/2 - L_{GIG}(\bar{\theta}|s) \leq 0$ where the equality holds if and only if $\theta_1 = \theta_2$.

On the other side, we can apply Talyor expansion to $L_{GIG}(\bar{\theta})$:

$$L_{GIG}(\bar{\theta}) \leq \frac{1}{2}(L_{GIG}(\theta_1|s) + L_{GIG}(\theta_2|s)) + \frac{1}{4}(\theta_2 - \theta_1)^{\mathsf{T}}(\nabla L_{GIG}(\theta_1|s) - \nabla L_{GIG}(\theta_2|s)).$$

It follows immediately that:

$$H_{GIG}(\theta_1\|\theta_2) \leq \frac{1}{2}\sqrt{(\theta_2 - \theta_1)^{\mathsf{T}}(\nabla L_{GIG}(\theta_1|s) - \nabla L_{GIG}(\theta_2|s))}. \tag{8}$$

Thus if the difference between each element in θ_1 and θ_2 is bounded and both θ_1 and θ_2 reaches the "local" optimal in the sense of each element $\nabla L_{GIG}(\theta_k|s), k = 1, 2$ is close to zero, then the Hellinger distance would be small even θ_1 and θ_2 are not close to each other. This fact can be applied to most exponential families. We will observe the same phenomena later for the GH distribution.

There are other special cases of the GIG distribution that are widely applied to finance: the inverse Gaussian (IG) distribution (when $\lambda = -1/2$), the gamma distribution (when $\lambda > 0$ and $\chi = 0$), and the inverse gamma distribution (when $\lambda < 0$ and $\psi = 0$).

2.2 The Generalized Hyperbolic Distributions

In this section we briefly review the basic properties of the GH distribution.

Definition 2. *Let Y be a GIG random variable with parameters (λ, χ, ψ), Z be an independent Gaussian random vector with zero mean and covariance Σ, Then the random vector:*

$$X \overset{d}{=} \mu + \gamma Y + \sqrt{Y}Z \tag{9}$$

follows the generalized hyperbolic distribution with parameters $(\mu, \gamma, \Sigma, \lambda, \chi, \psi)$, where $\mu, \gamma \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d\times d}$ is a positive definite matrix.

In the above definition we can see that μ is the location parameter, γ is the skewness parameter, Σ models the dependency structure of the multivariate distribution, and λ, χ, ψ serves to the heavy-tailness. In general many multivariate heavy-tailed distributions can be defined by (9) given some non-negative random variable Y. These distributions are usually called normal

mixture or Gaussian mixture distributions. In many literatures normal mixture means a discrete mixture of normal densities. In this paper we always refer the normal mixture distributions to any random variables that can be expressed by (9).

The joint distribution of X and Y is crucial in analyzing the GH distribution. In this paper we will call the distribution of X and Y as the *joint-GH distribution*. Its density function is as follows:

$$p(x,y|\mu,\gamma,\Sigma,\lambda,\chi,\psi) = \frac{1}{\sqrt{(2\pi)^d|\Sigma|}}\frac{(\psi/\chi)^{\frac{\lambda}{2}}}{2K_\lambda(\sqrt{\chi\psi})}y^{\lambda-1-\frac{d}{2}}$$
$$\exp\Big(-\frac{1}{2}(x-\mu-\gamma y)^\mathsf{T}\Sigma^{-1}(x-\mu-\gamma y)y^{-1}-\frac{1}{2}(\chi y^{-1}+\psi y)\Big), y>0, \quad (10)$$

from which one get the marginal distribution of x which is the GH density function:

$$p(x|\mu,\gamma,\Sigma,\lambda,\chi,\psi) = \int_0^\infty p(x,y|\mu,\gamma,\Sigma,\lambda,\chi,\psi)dy$$
$$= c\frac{K_{\lambda-\frac{d}{2}}(\sqrt{(\chi+(x-\mu)^\mathsf{T}\Sigma^{-1}(x-\mu))(\psi+\gamma^\mathsf{T}\Sigma^{-1}\gamma)})}{(\sqrt{(\chi+(x-\mu)^\mathsf{T}\Sigma^{-1}(x-\mu))(\psi+\gamma^\mathsf{T}\Sigma^{-1}\gamma)})^{\frac{d}{2}-\lambda}}e^{(x-\mu)^\mathsf{T}\Sigma^{-1}\gamma}, \quad (11)$$

where

$$c = \frac{(\psi/\chi)^{\frac{\lambda}{2}}(\psi+\gamma^\mathsf{T}\Sigma^{-1}\gamma)^{\frac{d}{2}-\lambda}}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}K_\lambda(\sqrt{\chi\psi})}. \quad (12)$$

Similar as before one can use another parameterization δ and η instead of χ and ψ:

$$p(x|\mu,\gamma,\Sigma,\lambda,\delta,\eta) =$$
$$c\frac{K_{\lambda-\frac{d}{2}}(\sqrt{(\eta+(x-\mu)^\mathsf{T}(\delta\Sigma)^{-1}(x-\mu))(\eta+(\delta\gamma)^\mathsf{T}(\delta\Sigma)^{-1}\delta\gamma)})}{(\sqrt{(\eta+(x-\mu)^\mathsf{T}(\delta\Sigma)^{-1}(x-\mu))(\eta+(\delta\gamma)^\mathsf{T}(\delta\Sigma)^{-1}\delta\gamma)})^{\frac{d}{2}-\lambda}}e^{(x-\mu)^\mathsf{T}(\delta\Sigma)^{-1}\delta\gamma},$$

where

$$c = \frac{(\eta+(\delta\gamma)^\mathsf{T}(\delta\Sigma)^{-1}\delta\gamma)^{\frac{d}{2}-\lambda}}{(2\pi)^{\frac{d}{2}}|\delta\Sigma|^{\frac{1}{2}}K_\lambda(\eta)}.$$

From the above representation one can observe that the GH model is not regular since the parameter sets $(\mu, \gamma/c, \Sigma/c, \lambda, c\delta, \eta)$ have the same distributionx for any $c > 0$. So the Fisher information matrix of the GH distribution would be singular. There are several ways to regularize the GH family. The simplest way is to set $\delta = 1$. [34] sets $\chi = 1$ in the EM-algorithm; [18] suggests to fix χ when $\lambda > -1$ and fix ψ when $\lambda < 1$, due to the ill-condition of the GIG optimization problem discussed in the previous section; and [26] suggests to fix the determinant of Σ, for example, set $|\Sigma| = 1$.

There is a good reason to fix $|\Sigma| = 1$ when the dimension of the problem is high. Note that $|\Sigma/c| = |\Sigma|/c^d$, any small perturbation of the matrix scale will make $|\Sigma|$ change dramatically when d is large. The inversion of the matrix would be intractable if $|\Sigma|$ is too large or too small to be computed numerically. And the matrix inversion is the key step in the EM algorithm that we will discuss in following sections.

Let us return to the joint-GH distribution (10). It is clear that it is also an exponential family with the expectation parameters:

$$\begin{aligned}
s_1 &:= E[Y^{-1}] = \sqrt{\frac{\psi}{\chi}}\frac{K_{\lambda-1}(\sqrt{\chi\psi})}{K_\lambda(\sqrt{\chi\psi})}, \\
s_2 &:= E[Y] = \sqrt{\frac{\chi}{\psi}}\frac{K_{\lambda+1}(\sqrt{\chi\psi})}{K_\lambda(\sqrt{\chi\psi})}, \\
s_3 &:= E[\log(Y)] = \frac{\partial}{\partial\alpha}\left(\sqrt{\frac{\chi}{\psi}}\right)^\alpha \frac{K_{\lambda+\alpha}(\sqrt{\chi\psi})}{K_\lambda(\sqrt{\chi\psi})}\bigg|_{\alpha=0}, \\
s_4 &:= E[X] = \mu + \gamma s_2, \\
s_5 &:= E[XY^{-1}] = \mu s_1 + \gamma, \\
s_6 &:= E[XX^\mathsf{T}Y^{-1}] = \Sigma + \mu\mu^\mathsf{T} s_1 + \gamma\gamma^\mathsf{T} s_2 + \mu\gamma^\mathsf{T} + \gamma\mu^\mathsf{T},
\end{aligned} \tag{13}$$

where $s_1, s_2, s_3 \in \mathbb{R}$, $s_4, s_5 \in \mathbb{R}^d$ and $s_6 \in \mathbb{R}^{d\times d}$. Note that s_1, s_2, s_3 are exactly the expectation parameters of GIG random variable Y. On the other side, given all the expectation parameters we can get the original parameter

as follows:

$$\begin{aligned}
\mu &= \frac{s_4 - s_2 s_5}{1 - s_1 s_2}, \\
\gamma &= \frac{s_5 - s_1 s_4}{1 - s_1 s_2}, \\
\Sigma &= s_6 - s_5\mu^\mathsf{T} - \mu s_5^\mathsf{T} + s_1\mu\mu^\mathsf{T} - s_2\gamma\gamma^\mathsf{T}, \\
(\lambda, \chi, \psi) &= \arg\max_{\lambda,\chi,\psi} L_{GIG}(\lambda, \chi, \psi | s_1, s_2, s_3),
\end{aligned} \tag{14}$$

where L_{GIG} is given by (6). As we already know that the above equations are exactly the solutions of the optimization problem:

$$\max_{\mu,\gamma,\Sigma,\lambda,\chi,\psi} L_{GH}(\mu, \gamma, \Sigma, \lambda, \chi, \psi | s_1, s_2, s_3, s_4, s_5, s_6),$$

where

$$\begin{aligned}
&L_{GH}(\mu, \gamma, \Sigma, \lambda, \chi, \psi | s_1, s_2, s_3, s_4, s_5, s_6) \\
&= -\frac{1}{2}\mu^\mathsf{T}\Sigma^{-1}\mu s_1 - \frac{1}{2}\gamma^\mathsf{T}\Sigma^{-1}\gamma s_2 + \gamma^\mathsf{T}\Sigma^{-1}s_4 + \mu^\mathsf{T}\Sigma^{-1}s_5 \\
&- \frac{1}{2}tr(\Sigma^{-1}s_6) - \mu^\mathsf{T}\Sigma^{-1}\gamma - \frac{1}{2}\log|\Sigma| + L_{GIG}(\lambda, \chi, \psi | s_1, s_2, s_3)
\end{aligned}$$

is the log-likelihood function of the joint-GH distribution excluding some constants. We will see later that (13) and (14) forms the E-step and the M-step in the EM algorithm, except that the expectations in (13) are replaced by the conditional ones.

As far as we know, there is no analytical formulation of the Hellinger distance between two GH distributions. However the Hellinger distance of the joint-GH distributions is easy to compute:

Proposition 2. *Let $\theta_1 = (\mu_1, \gamma_1, \Sigma_1, \lambda_1, \chi_1, \psi_1)$ and $\theta_2 = (\mu_2, \gamma_2, \Sigma_2, \lambda_2, \chi_2, \psi_2)$ be the parameters of two joint-GH distributions. The squared Hellinger distance between the two distributions is given by:*

$$\begin{aligned}
H^2_{JGH}(\theta_1 \| \theta_2) =& 1 - \frac{|\Sigma_1\Sigma_2|^{\frac{1}{4}}}{|\bar{\Sigma}|^{\frac{1}{2}}} \frac{(\psi_1/\chi_1)^{\frac{\lambda_1}{4}}(\psi_2/\chi_2)^{\frac{\lambda_2}{4}}}{\sqrt{K_{\lambda_1}(\sqrt{\chi_1\psi_1})K_{\lambda_2}(\sqrt{\chi_2\psi_2})}} \\
& \frac{K_{\bar{\lambda}}(\sqrt{(\bar{\chi} + \frac{1}{4}\Delta\mu^\mathsf{T}\bar{\Sigma}^{-1}\Delta\mu)(\bar{\psi} + \frac{1}{4}\Delta\gamma^\mathsf{T}\bar{\Sigma}^{-1}\Delta\gamma)})}{((\bar{\psi} + \frac{1}{4}\Delta\gamma^\mathsf{T}\bar{\Sigma}^{-1}\Delta\gamma)/(\bar{\chi} + \frac{1}{4}\Delta\mu^\mathsf{T}\bar{\Sigma}^{-1}\Delta\mu))^{\frac{\bar{\lambda}}{2}}} e^{-\frac{1}{4}\Delta\mu^\mathsf{T}\bar{\Sigma}^{-1}\Delta\gamma},
\end{aligned}$$

where $\Delta\mu = \mu_1 - \mu_2$, $\Delta\gamma = \gamma_1 - \gamma_2$, $\bar{\Sigma} = \frac{1}{2}\Sigma_1 + \frac{1}{2}\Sigma_2$, $\bar{\lambda} = \frac{1}{2}\lambda_1 + \frac{1}{2}\lambda_2$, $\bar{\chi} = \frac{1}{2}\chi_1 + \frac{1}{2}\chi_2$ *and* $\bar{\psi} = \frac{1}{2}\psi_1 + \frac{1}{2}\psi_2$.

Proof. The Hellinger affinity is given by:

$$
\begin{aligned}
&\int\int \sqrt{p(x,y|\theta_1)p(x,y|\theta_2)}dxdy \\
&= \int\int \sqrt{p(x|y,\theta_1)p(x|y,\theta_2)}dx\sqrt{p(y|\theta_1)p(y|\theta_2)}dy \\
&= \int \frac{|\Sigma_1\Sigma_2|^{\frac{1}{4}}}{|\bar{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{8}(\Delta\mu + \Delta\gamma y)^{\mathsf{T}}\bar{\Sigma}^{-1}(\Delta\mu + \Delta\gamma y)y^{-1}\right) \\
&\sqrt{p(y|\lambda_1,\chi_1,\psi_1)p(y|\lambda_2,\chi_2,\psi_2)}dy \\
&= \frac{|\Sigma_1\Sigma_2|^{\frac{1}{4}}}{|\bar{\Sigma}|^{\frac{1}{2}}} \frac{(\psi_1/\chi_1)^{\frac{\lambda_1}{4}}(\psi_2/\chi_2)^{\frac{\lambda_2}{4}}}{\sqrt{K_{\lambda_1}(\sqrt{\chi_1\psi_1})K_{\lambda_2}(\sqrt{\chi_2\psi_2})}} \frac{K_{\bar{\lambda}}(\sqrt{\bar{\chi}\bar{\psi}})}{(\bar{\psi}/\bar{\chi})^{\frac{\bar{\lambda}}{2}}} \\
&\int \exp\left(-\frac{1}{8}(\Delta\mu + \Delta\gamma y)^{\mathsf{T}}\bar{\Sigma}^{-1}(\Delta\mu + \Delta\gamma y)y^{-1}\right) p(y|\bar{\lambda},\bar{\chi},\bar{\psi})dy \\
&= \frac{|\Sigma_1\Sigma_2|^{\frac{1}{4}}}{|\bar{\Sigma}|^{\frac{1}{2}}} \frac{(\psi_1/\chi_1)^{\frac{\lambda_1}{4}}(\psi_2/\chi_2)^{\frac{\lambda_2}{4}}}{\sqrt{K_{\lambda_1}(\sqrt{\chi_1\psi_1})K_{\lambda_2}(\sqrt{\chi_2\psi_2})}} \\
&\frac{K_{\bar{\lambda}}(\sqrt{(\bar{\chi} + \frac{1}{4}\Delta\mu^{\mathsf{T}}\bar{\Sigma}^{-1}\Delta\mu)(\bar{\psi} + \frac{1}{4}\Delta\gamma^{\mathsf{T}}\bar{\Sigma}^{-1}\Delta\gamma)})}{((\bar{\psi} + \frac{1}{4}\Delta\gamma^{\mathsf{T}}\bar{\Sigma}^{-1}\Delta\gamma)/(\bar{\chi} + \frac{1}{4}\Delta\mu^{\mathsf{T}}\bar{\Sigma}^{-1}\Delta\mu))^{\frac{\bar{\lambda}}{2}}} e^{-\frac{1}{4}\Delta\mu^{\mathsf{T}}\bar{\Sigma}^{-1}\Delta\gamma}.
\end{aligned}
$$

□

It is easy to see that if $\mu_1 = \mu_2$, $\gamma_1 = \gamma_2$ and $\Sigma_1 = \Sigma_2$ then $H_{JGH}(\theta_1\|\theta_2) = H_{GIG}(\lambda_1,\chi_1,\psi_1\|\lambda_2,\chi_2,\psi_2)$. Although H_{JGH} is different with the Hellinger distance of GH, it gives an upper bound of the latter so we are able to use it the measure how close two GH distributions are.

Now suppose that $\theta = (\mu,\gamma,\Sigma,\lambda,\chi,\psi)$ are the "true" parameters and $(s_1,s_2,s_3,s_4,s_5,s_6)$ are the corresponding expectation parameters. Then if we do (14) in computer, we would get a different set of parameters, say $\tilde{\theta} = (\tilde{\mu},\tilde{\gamma},\tilde{\Sigma},\tilde{\lambda},\tilde{\chi},\tilde{\psi})$. Since the formula for μ are arithmetic, we must have $\|\tilde{\mu} - \mu\|/\|\mu\| = O(\epsilon)$ where ϵ denotes the machine epsilon. This is true for γ and Σ. Thus we can believe that $H_{JGH}(\theta\|\tilde{\theta}) \approx H_{GIG}(\lambda,\chi,\psi\|\tilde{\lambda},\tilde{\chi},\tilde{\psi})$. Then we are able to apply (8) and conclude that the $H_{JGH}(\theta\|\tilde{\theta})$ would be small if $\tilde{\lambda},\tilde{\chi},\tilde{\psi}$ reaches to "local" optimal of the GIG log-likelihood function.

To illustrate this point, we first use EM algorithm to fit the S&P 500 index daily return from 2010 to 2015. λ is fixed to be 10 in order to make the problem ill-conditioned. The output parameters are assumed to be the "true" ones; then we compute (13) and (14) via Matlab. Table 2 shows the results.

	True parameters	Estimated parameters	Relative errors
μ	0.2646	0.2646	1.2587e-15
γ	-0.3013	-0.3013	-1.4740e-15
σ	1	1	0
λ	10	9.9897	0.0010
χ	8.4438e-05	0.0099	115.9873
ψ	24.1022	24.0932	0.0004

Table 2: Relative errors of the GH parameters

It is clear that the relative errors of μ and γ are around the machine epsilon. σ is set to be 1 in order to regularize the GH parameters, which we will discuss in the next chapter. The relative error of χ is tremendous. However the Hellinger distance between the true and the estimated parameters is small, which is about 1.5359×10^{-8}. And the GIG Hellinger distance is almost the same as the GH distance. To conclude, the computation of (14) is not stable in terms of relative error under certain conditions; but it is relatively stable in terms of the Hellinger distance.

3 Parameter Estimation of the Generalized Hyperbolic Distributions

3.1 Expectation-Maximization Algorithm for the Generalized Hyperbolic Distributions

The EM algorithm is a classical iterative method for fitting data with hidden values. [12] shows that the EM algorithm would converge to the traditional MLE. There are several different types of the EM algorithms for the GH distribution. In this paper we follow the EM framework in [18]. However, our approach is slightly different with the previous works. First we use general convex optimization algorithms to solve (5) directly without fixing the parameter λ. The numerical tests in the previous chapter show that the computation of λ is relatively precise. In addition, constraints on χ or ψ are also unnecessary since the optimization problem (5) is stable under the Hellinger distance. The best way to regularize the GH parameters is to fix the determinant of Σ, as suggested by [26]. Unlike the algorithm in [26] however, we show that the regularization can be done directly at the end of each EM iteration without affecting the convergence.

Let $\theta = (\mu, \gamma, \Sigma, \lambda, \chi, \psi)$ be the parameter set of the GH distribution for simplicity. Recall that a GH random vector X can be expressed as (9) and the joint distribution of X and Y has the density (10). We are able to compute the conditional density of Y given X:

$$
\begin{aligned}
p(y|x,\theta) &= \frac{p(x,y|\theta)}{p(x|\theta)} \\
&\sim y^{\lambda-1-\frac{d}{2}} \exp\Big(-\frac{1}{2}(x-\mu-\gamma y)^{\mathsf{T}}\Sigma^{-1}(x-\mu-\gamma y)y^{-1} - \frac{1}{2}(\chi y^{-1}+\psi y)\Big) \\
&\sim y^{\lambda-1-\frac{d}{2}} \exp\Big(-\frac{1}{2}(\chi + (x-\mu)^{\mathsf{T}}\Sigma^{-1}(x-\mu))y^{-1} - \frac{1}{2}(\psi^{\mathsf{T}}+\gamma\Sigma^{-1}\gamma)y\Big),
\end{aligned}
$$

where the operator $\sim$ means "proportional" since (x, θ) are regarded as constants in the conditional distribution. It is clear that the above density is the GIG distribution with parameters $(\lambda - d/2, \chi + (x-\mu)^{\mathsf{T}}\Sigma^{-1}(x-\mu), \psi +$

$\gamma^\mathsf{T}\Sigma^{-1}\gamma$). Apply (3) we obtain:

$$E[Y^\alpha|X=x,\theta] = \left(\sqrt{\frac{\chi+(x-\mu)^\mathsf{T}\Sigma^{-1}(x-\mu)}{\psi+\gamma^\mathsf{T}\Sigma^{-1}\gamma}}\right)^\alpha \frac{K_{\lambda-\frac{d}{2}+\alpha}(\sqrt{(\chi+(x-\mu)^\mathsf{T}\Sigma^{-1}(x-\mu))(\psi+\gamma^\mathsf{T}\Sigma^{-1}\gamma)})}{K_{\lambda-\frac{d}{2}}(\sqrt{(\chi+(x-\mu)^\mathsf{T}\Sigma^{-1}(x-\mu))(\psi+\gamma^\mathsf{T}\Sigma^{-1}\gamma)})}, \tag{15}$$

$$E[\log Y|X=x,\theta] = \frac{\partial}{\partial\alpha}E[Y^\alpha|X=x,\theta]\Big|_{\alpha=0}.$$

Like before the derivative in the second equation can be computed numerically.

Now let $x_1,\ldots,x_n \in \mathbb{R}^d$ be a sequence of i.i.d sample data. The EM algorithm is an iterative approach for maximizing the likelihood. Each iteration consists two steps: the expectation or the E-step, and the maximization or the M-step. Given a set of initial parameters $\theta_0 = (\mu_0, \gamma_0, \Sigma_0, \lambda_0, \chi_0, \psi_0)$, the $k+1$-th E-step of the EM algorithm computes the average of the conditional expectation of the sufficient statistics:

$$\hat{s}_1^{(k)} = \frac{1}{n}\sum_{j=1}^n E[Y^{-1}|X=x_j,\theta_k],$$

$$\hat{s}_2^{(k)} = \frac{1}{n}\sum_{j=1}^n E[Y|X=x_j,\theta_k],$$

$$\hat{s}_3^{(k)} = \frac{1}{n}\sum_{j=1}^n E[\log Y|X=x_j,\theta_k],$$

$$\hat{s}_4^{(k)} = \frac{1}{n}\sum_{j=1}^n E[X|X=x_j,\theta_k] = \sum_{j=1}^n x_j,$$

$$\hat{s}_5^{(k)} = \frac{1}{n}\sum_{j=1}^n E[XY^{-1}|X=x_j,\theta_k] = \frac{1}{n}\sum_{j=1}^n x_j E[Y^{-1}|X=x_j,\theta_k],$$

$$\hat{s}_6^{(k)} = \frac{1}{n}\sum_{j=1}^n E[XX^\mathsf{T}Y^{-1}|X=x_j,\theta_k] = \frac{1}{n}\sum_{j=1}^n x_j x_j^\mathsf{T} E[Y^{-1}|X=x_j,\theta_k],$$

where the expectations can be computed by (15). The M-step is to solve the

optimization problem:

$$\theta_{k+1} = \arg\max_{\theta} \sum_{j=1}^{n} E[\log p(X, Y|\theta)|X = x_j, \theta_k].$$

And this is exactly equivalent to

$$\theta_{k+1} = \arg\max_{\theta} L_{GH}(\mu, \gamma, \Sigma, \lambda, \chi, \psi|\hat{s}_1^{(k)}, \hat{s}_2^{(k)}, \hat{s}_3^{(k)}, \hat{s}_4^{(k)}, \hat{s}_5^{(k)}, \hat{s}_6^{(k)}), \tag{16}$$

whose solutions are given by (14). Rewrite (14) we have the M-step of the algorithm:

$$\begin{aligned}
\mu_{k+1} &= \frac{\hat{s}_4^{(k)} - \hat{s}_2^{(k)}\hat{s}_5^{(k)}}{1 - \hat{s}_1^{(k)}\hat{s}_2^{(k)}}, \\
\gamma_{k+1} &= \frac{\hat{s}_5^{(k)} - \hat{s}_1^{(k)}\hat{s}_4^{(k)}}{1 - \hat{s}_1^{(k)}\hat{s}_2^{(k)}}, \\
\Sigma_{k+1} &= \hat{s}_6^{(k)} - \hat{s}_5^{(k)}\mu^{\mathsf{T}} - \mu(\hat{s}_5^{(k)})^{\mathsf{T}} + \hat{s}_1^{(k)}\mu\mu^{\mathsf{T}} - \hat{s}_2^{(k)}\gamma\gamma^{\mathsf{T}}, \\
(\lambda_{k+1}, \chi_{k+1}, \psi_{k+1}) &= \arg\max_{\lambda,\chi,\psi} L_{GIG}(\lambda, \chi, \psi|\hat{s}_1^{(k)}, \hat{s}_2^{(k)}, \hat{s}_3^{(k)}).
\end{aligned} \tag{17}$$

There are some numerical issues in the above algorithm when the dimension d is high. The first one is the computation of the modified Bessel functions in (15). For large d, say 500 for example, $K_{\lambda+\frac{d}{2}+\alpha}$ might turns to zero or infinity in Matlab for some data points x. This problem is addressed in the appendix.

The second problem is to compute Σ^{-1}. As we have discussed in the previous chapter, a good way to regulate GH distribution is to set its determinant to be 1. Instead of adding an extra constraint such that $|\Sigma| = 1$ in the optimization problem (16), we rescale the parameters at the end of M-step:

$$(\mu_k, \gamma_k, \Sigma_k, \lambda_k, \chi_k, \psi_k) \to (\mu_k, |\Sigma_k|^{-\frac{1}{d}}\gamma_k, |\Sigma_k|^{-\frac{1}{d}}\Sigma_k, \lambda_k, |\Sigma_k|^{\frac{1}{d}}\chi_k, |\Sigma_k|^{-\frac{1}{d}}\psi_k).$$

As we know these two sets of parameters are equivalent for the GH distribution. And a simple calculation will show that such rescaling will not change the result and the convergency of the EM algorithm:

Proposition 3. *If we write the $k+1$-th iteration of the EM algorithm as a function f, i.e.*

$$(\mu_{k+1}, \gamma_{k+1}, \Sigma_{k+1}, \lambda_{k+1}, \chi_{k+1}, \psi_{k+1}) = f(\mu_k, \gamma_k, \Sigma_k, \lambda_k, \chi_k, \psi_k),$$

then for any $c > 0$

$$(\mu_{k+1}, c\gamma_{k+1}, c\Sigma_{k+1}, \lambda_{k+1}, \chi_{k+1}/c, c\psi_{k+1}) = f(\mu_k, c\gamma_k, c\Sigma_k, \lambda_k, \chi_k/c, c\psi_k).$$

Proof. Let $\theta_k = (\mu_k, \gamma_k, \Sigma_k, \lambda_k, \chi_k, \psi_k)$ and $\tilde{\theta}_k = (\mu_k, c\gamma_k, c\Sigma_k, \lambda_k, \chi_k/c, c\psi_k)$. A direct computation using (15) shows that:

$$\begin{aligned} E[Y^\alpha | X = x, \tilde{\theta}_k] &= E[Y^\alpha | X = x, \theta_k]/c^\alpha, \\ E[\log Y | X = x, \tilde{\theta}_k] &= E[\log Y | X = x, \theta_k] - \log c. \end{aligned}$$

Thus if $\hat{s}_1^{(k)}, \hat{s}_2^{(k)}, \hat{s}_3^{(k)}, \hat{s}_4^{(k)}, \hat{s}_5^{(k)}, \hat{s}_6^{(k)}$ are the outputs of the E-step given θ_k, then $c\hat{s}_1^{(k)}, \hat{s}_2^{(k)}/c, \hat{s}_3^{(k)} - \log c, \hat{s}_4^{(k)}, c\hat{s}_5^{(k)}, c\hat{s}_6^{(k)}$ would be the corresponding outputs given $\tilde{\theta}_k$. We finish the proof by applying these parameters to (17). □

Another approach is the multi-cycle expectation conditional maximization (MCECM) algorithm proposed by [26]. Unlike the EM-algorithm which updates all parameters via (17), the MCECM algorithm first computes μ_{k+1}, γ_{k+1} and Σ_{k+1} according to the first three equations in (17) and set $\Sigma_{k+1} \leftarrow \Sigma_{k+1}/|\Sigma_{k+1}|^{1/d}$ so that it has unite determinate while μ_{k+1} and γ_{k+1} are unchanged. Then it sets $\tilde{\theta}_{k+1} := (\mu_{k+1}, \gamma_{k+1}, \Sigma_{k+1}, \lambda_k, \chi_k, \psi_k)$ and computes:

$$\begin{aligned} \tilde{s}_1^{(k+1)} &= \frac{1}{n}\sum_{j=1}^n E[Y^{-1} | X = x_j, \tilde{\theta}_{k+1}], \\ \tilde{s}_2^{(k+1)} &= \frac{1}{n}\sum_{j=1}^n E[Y | X = x_j, \tilde{\theta}_{k+1}], \\ \tilde{s}_3^{(k+1)} &= \frac{1}{n}\sum_{j=1}^n E[\log Y | X = x_j, \tilde{\theta}_{k+1}]. \end{aligned}$$

Finally the updates of the rest of the parameters are given by:

$$(\lambda_{k+1}, \chi_{k+1}, \psi_{k+1}) = \arg\max_{\lambda,\chi,\psi} L_{GIG}(\lambda, \chi, \psi | \tilde{s}_1^{(k+1)}, \tilde{s}_2^{(k+1)}, \tilde{s}_3^{(k+1)}).$$

Both the MCECM and the EM algorithm that rescales parameters in each iteration converge to the MLE. In terms of efficiency they are also very similar to our experience. The simpler EM algorithm might have a bit computation advantage, but the difference is not significant.

To illustrate, we test both algorithms numerically. First we fit the standardized daily returns of S&P 500 and FTSE indices from 2008 to 2016 by a two dimensional GH distribution via the EM algorithm with fixed λ. According to [18] the accuracy of the M-step would be affected by λ, especially under extreme cases. Thus We set λ to be $-10, -9, \ldots, 10$ and get corresponding 21 sets of parameters given by the EM algorithm. Then for each set of parameters we generate 5000 i.i.d multivariate GH random samples. Then we fit the sample by the EM algorithm and compute the relative error of the estimated parameters. We define the relative errors as $\|\hat{\theta} - \theta\|_2 / \|\theta\|_2$ where $\|\cdot\|_2$ is the 2-norm. The results are given by table 3.

	Relative errors						H^2
λ	μ	γ	Σ	λ	χ	ψ	
-10	0.3207	0.3559	0.0108	1.3591	0.8636	4.3056×10^5	0.0344
-9	0.0728	0.0899	0.0071	1.4072	0.8687	9.9812×10^6	0.0166
-8	0.4742	0.4723	0.0232	1.1493	0.6154	5.5261×10^5	0.0137
-7	0.4419	0.3541	0.0089	1.3486	0.8194	1.1266×10^7	0.0223
-6	0.3038	0.4189	0.0439	0.1366	0.1683	71.3657	0.0086
-5	0.5297	0.6145	0.0140	1.1887	0.7586	1.0123×10^7	0.0239
-4	0.5717	0.7755	0.0048	0.8570	0.5648	3.9045×10^5	0.0166
-3	0.1336	0.0950	0.0117	0.0125	0.0179	92.9025	0.0013
-2	0.2183	0.2638	0.0156	0.0088	0.0591	5.2433×10^3	0.0013
-1	0.0650	0.3661	0.0135	0.0416	0.0397	0.3508	0.0008
0	0.3208	0.1452	0.0029	2.2441	0.0916	0.0377	0.0017
1	0.0618	0.0762	0.0155	0.0662	0.6320	0.0789	0.0005
2	0.0714	0.4336	0.0432	0.2634	4.2627×10^5	0.0571	0.0055
3	0.2339	0.3040	0.0195	0.5092	1.4997×10^5	0.2553	0.0059
4	0.4386	0.5570	0.0049	0.0748	1.9663×10^5	0.0160	0.0137
5	0.7064	0.8024	0.0096	0.3068	1.4570×10^5	0.3109	0.0282
6	0.2725	0.3564	0.0143	0.2359	1.9352×10^3	0.2478	0.0113
7	0.2371	0.2265	0.0175	0.0741	1.0459×10^3	0.0617	0.0029
8	0.2232	0.2190	0.0057	0.1029	6.4067×10^3	0.0948	0.0041
9	0.2340	0.1965	0.0062	0.1255	5.7621×10^3	0.1232	0.0039
10	0.2487	0.2793	0.0233	0.0624	1.7391×10^4	0.0572	0.0069

Table 3: Errors of the GH EM algorithm

H^2 in the last column denotes for the squared Hellinger distance. As we can see that the error of ψ is huge if λ is negative and the error χ is huge if λ is positive. This corresponds to our discussion in section 2.2. Even the estimated parameters are not close to the "ture" ones, small Hellinger distance implies that the two distributions are close. Figure 2 and 3 compares two marginal density functions of the "ture" distribution to the fitted distribution when $\lambda = 10$. We can see that even the relative errors of χ are tremendous, the fitted distributions is still quite close to the true ones.

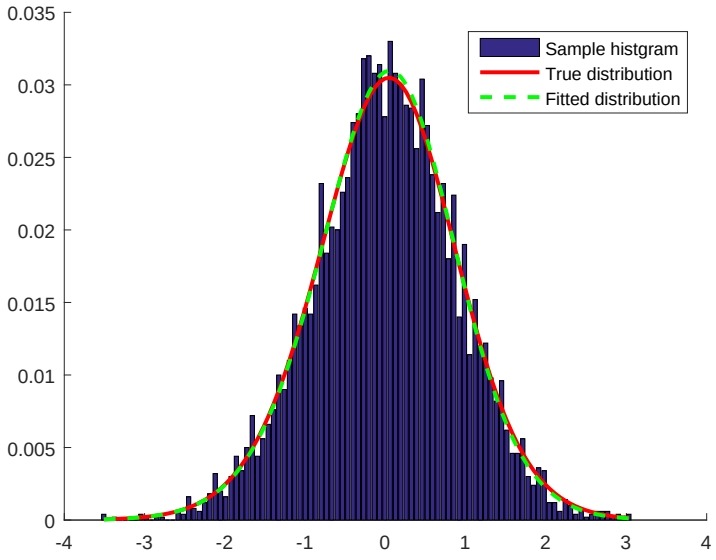


Figure 2: Marginal distribution of S&P 500

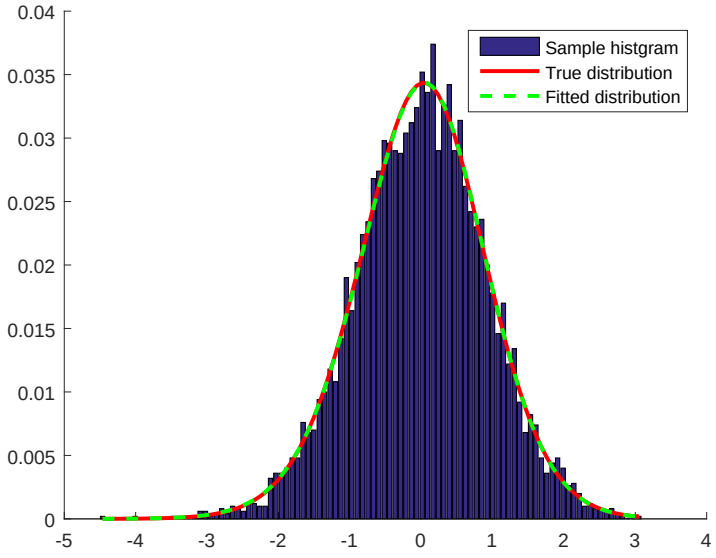


Figure 3: Marginal distribution of FTSE

Now we fit the same samples via the MCECM algorithm. The comparison between two algorithms are shown by table 4. It is not surprising to see that the log-likelihoods of two algorithms are almost the same, since both algorithms converge to the MLE. The computation times are also similar. The EM algorithm takes about 82 seconds and the MCECM algorithm takes about 96 seconds to compute all of these results on a laptop with 1.8 GHz Intel Core i7 processor and 4 GB 1333 MHz DDR3 memory.

	EM algorithm		MCECM algorithm	
	Log-likelihood	H^2	Log-likelihood	H^2
-10	-0.4582	0.0344	-0.4582	0.0346
-9	-0.4310	0.0166	-0.4310	0.0166
-8	-0.4080	0.0137	-0.4080	0.0140
-7	-0.4197	0.0223	-0.4197	0.0228
-6	-0.3734	0.0086	-0.3735	0.0090
-5	-0.3534	0.0239	-0.3534	0.0246
-4	-0.3079	0.0166	-0.3079	0.0171
-3	-0.3432	0.0013	-0.3432	0.0012
-2	-0.3293	0.0013	-0.3293	0.0012
-1	-0.4024	0.0008	-0.4025	0.0008
0	-0.4149	0.0017	-0.4149	0.0017
1	-0.4037	0.0005	-0.4037	0.0005
2	-0.3661	0.0055	-0.3660	0.0051
3	-0.4407	0.0059	-0.4407	0.0057
4	-0.4388	0.0137	-0.4388	0.0137
5	-0.5328	0.0282	-0.5328	0.0284
6	-0.5199	0.0113	-0.5199	0.0115
7	-0.5341	0.0029	-0.5341	0.0029
8	-0.5422	0.0041	-0.5422	0.0042
9	-0.5451	0.0039	-0.5451	0.0038
10	-0.5542	0.0069	-0.5542	0.0066

Table 4: Comparison between the EM algorithm and the MECEM algorithm

3.2 Shrinkage with the Penalized Likelihood

By letting $|\Sigma| = 1$ we ensure that the matrix is numerically invertible in each iteration of the EM algorithm. But this does not guarantee that matrix inversion is well-conditioned. The third formula in (17) has the same problem as the sample covariance: the condition number of Σ_k would be huge when the sample size is relatively small. Thus some biased estimators such as shrinkage are necessary to improve the condition number of Σ. In this section we introduce a simple shrinkage approach based on the penalized likelihood.

Let us first consider a general exponential family with density:

$$p(x, y|\theta) = h(x, y) \exp(\theta^\mathsf{T} S(x, y) - G(\theta)), \tag{18}$$

where

$$G(\theta) = \log \int h(x, y) \exp(\theta^\mathsf{T} S(x, y)) dx dy.$$

And recall that the Kullback-Leibler divergence between two exponential families are given by

$$D_{KL}(\theta_1 \| \theta_2) = E\left[\log \frac{p(X, Y|\theta_1)}{p(X, Y|\theta_2)} \Big| \theta_1 \right] = G(\theta_2) - G(\theta_1) + s_1^\mathsf{T}(\theta_2 - \theta_1),$$

where $s_1 = E[S(X, Y)|\theta_1]$. This also corresponds to the Bregman divergence with potential function G.

Now let us assume that x is observable while y is hidden. Given an sequence of sample $x_1, x_2, \ldots, x_n$ and some parameter θ_0, our goal is to maximize the following penalized likelihood:

$$\max_\theta \frac{1}{n} \sum_{j=1}^{n} \log p(x_j|\theta) - \tau D_{KL}(\theta_0 \| \theta),$$

where $p(x|\theta) = \int p(x, y|\theta) dy$ is the marginal density of X and $\tau \geq 0$ represents the amount of "shrinkage". If $\tau = 0$ then this is the same as the original maximum likelihood. As τ grows larger the optimal solution $\hat{\theta}$ would be closer to θ_0.

This problem can also be solved iteratively by the following EM-algorithm:

$$\theta_{k+1} = \arg\max_\theta \frac{1}{n} \sum_{j=1}^{n} E[\log p(X, Y|\theta)|x_j, \theta_k] - \tau D_{KL}(\theta_0 \| \theta),$$

where the E-step computes the conditional expectation of the log-likelihood and the M-step solves the optimization problem. One can easily show that the penalized likelihood would increase as the number of iterations grows:

$$
\begin{aligned}
&\frac{1}{n}\sum_{j=1}^{n}\log p(x_j|\theta_{k+1}) - \tau D_{KL}(\theta_0\|\theta_{k+1}) - \frac{1}{n}\sum_{j=1}^{n}\log p(x_j|\theta_k) + \tau D_{KL}(\theta_0\|\theta_k) \\
&= \frac{1}{n}\sum_{j=1}^{n} E[\log p(X,Y|\theta_{k+1}) - \log p(Y|X,\theta_{k+1})|X = x_j,\theta_k] - \tau D_{KL}(\theta_0\|\theta_{k+1}) \\
&- \frac{1}{n}\sum_{j=1}^{n} E[\log p(X,Y|\theta_k) - \log p(Y|X,\theta_k)|X = x_j,\theta_k] - \tau D_{KL}(\theta_0\|\theta_k) \\
&\geq \frac{1}{n}\sum_{j=1}^{n} E[\log p(X,Y|\theta_{k+1})|X = x_j,\theta_k] - \tau D_{KL}(\theta_0\|\theta_{k+1}) \\
&- \frac{1}{n}\sum_{j=1}^{n} E[\log p(X,Y|\theta_k)|X = x_j,\theta_k] - \tau D_{KL}(\theta_0\|\theta_k) \geq 0,
\end{aligned}
$$

since $E[\log p(Y|X,\theta_k) - \log p(Y|X,\theta_{k+1})|X = x_j,\theta_k] \geq 0$ is the Kullback-Leibler divergence between two conditional distributions. Using the representation (18) we find that the above optimization problem is equivalent to:

$$
\theta_{k+1} = \arg\max_{\theta} \frac{1}{n}\sum_{j=1}^{n} \theta^{\mathsf{T}}\big(E[S(X,Y)|x_j,\theta_k] + \tau s_0\big) - (1+\tau)G(\theta),
$$

where $s_0 = E[S(X,Y)|\theta_0]$. As a result, in the k-th E-step we computes the

shrinked sufficient statistcs:

$$\hat{s}_1^{(k)} = \frac{1}{(1+\tau)n}\sum_{j=1}^{n} E[Y^{-1}|X = x_j, \theta_k] + \frac{\tau}{1+\tau}E[Y^{-1}|\theta_0],$$

$$\hat{s}_2^{(k)} = \frac{1}{(1+\tau)n}\sum_{j=1}^{n} E[Y|X = x_j, \theta_k] + \frac{\tau}{1+\tau}E[Y^{|}\theta_0],$$

$$\hat{s}_3^{(k)} = \frac{1}{(1+\tau)n}\sum_{j=1}^{n} E[\log Y|X = x_j, \theta_k] + \frac{\tau}{1+\tau}E[\log Y|\theta_0],$$

$$\hat{s}_4^{(k)} = \frac{1}{(1+\tau)n}\sum_{j=1}^{n} x_j + \frac{\tau}{1+\tau}E[X|\theta_0],$$

$$\hat{s}_5^{(k)} = \frac{1}{(1+\tau)n}\sum_{k=1}^{n} x_j E[Y^{-1}|X = x_j, \theta_k] + \frac{\tau}{1+\tau}E[XY^{-1}|\theta_0],$$

$$\hat{s}_6^{(k)} = \frac{1}{(1+\tau)n}\sum_{j=1}^{n} x_j x_j^{\mathsf{T}} E[Y^{-1}|X = x_j, \theta_k] + \frac{\tau}{1+\tau}E[XX^{\mathsf{T}}Y^{-1}|\theta_0],$$

where

$$E[Y^{\alpha}|\theta_0] = \left(\sqrt{\frac{\chi_0}{\psi_0}}\right)^{\alpha} \frac{K_{\lambda_0+\alpha_0}(\sqrt{\chi_0\psi_0})}{K_{\lambda_0}(\sqrt{\chi_0\psi_0})},$$

$$E[\log Y|\theta_0] = \frac{\partial}{\partial\alpha}E[Y^{\alpha}|\theta_0]\Big|_{\alpha=0},$$

$$E[X|\theta_0] = \mu_0 + \gamma_0 E[Y|\theta_0],$$

$$E[XY^{-1}|\theta_0] = \mu_0 E[Y^{-1}|\theta_0] + \gamma_0,$$

$$E[XX^{\mathsf{T}}Y^{-1}|\theta_0] = \Sigma_0 + \mu_0\mu_0^{\mathsf{T}}E[Y^{-1}|\theta_0] + \gamma_0\gamma_0^{\mathsf{T}}E[Y|\theta_0] + \mu_0\gamma_0^{\mathsf{T}} + \gamma_0\mu_0^{\mathsf{T}}.$$

Thus penalized maximum likelihood turns out to be a linear shrinkage of the conditional expectation parameters. The M-step is the same as the original EM-algorithm (17). Furthermore the linear relationship between Σ_{k+1} and $\hat{s}_6^{(k)}$ suggests that Σ_{k+1} can be shrinked to Σ_0 directly. By setting proper Σ_0 we are able to improve the condition of Σ_{k+1} for each k.

3.3 Factor Analysis for the Generalized Hyperbolic Distributions

Another way to improve the condition of Σ is the factor analysis proposed by [45]. Here we assume that Σ follows a certain structure: $\Sigma = FF^{\mathsf{T}} + D$ where $F \in \mathbb{R}^{d\times r}$, $r < d$ and $D \in \mathbb{R}^{d\times d}$ is a positive definite diagonal matrix. This is equivalent to say that a GH random vector X can be expressed as:

$$X \overset{d}{=} \mu + \gamma Y + \sqrt{Y}(FZ + \epsilon),$$

where $Z \in \mathbb{R}^r \sim N(0, I)$ and $\epsilon \sim N(0, D)$. The conditional distribution of X given Y and Z follows $N(\mu+\gamma Y+\sqrt{Y}FZ, DY)$. Thus the joint distribution of (X, Y, Z) is given by:

$$\begin{aligned} &p(x, y, z|\mu, \gamma, F, D, \lambda, \chi, \psi) = p(x|y, z, \mu, \gamma, F, D)p(y|\lambda, \chi, \psi)p(z) \\ &= \frac{1}{\sqrt{(2\pi)^{d+r}|D|}} \frac{(\sqrt{\chi/\psi})^\lambda}{2K_\lambda(\sqrt{\chi\psi})} y^{\lambda-1-\frac{d}{2}} \exp\Big(-\frac{1}{2}z^{\mathsf{T}}z - \frac{1}{2}(\chi y^{-1} + \psi y) \\ &- \frac{1}{2}(x - \mu - \gamma y - F\sqrt{y}z)^{\mathsf{T}} D^{-1}(x - \mu - \gamma y - F\sqrt{y}z)y^{-1}\Big), y > 0. \end{aligned} \tag{19}$$

One can observe that the above density function belongs to a curved exponential family, i.e. which has the expression:

$$p(x, y, z) = h(x, y, z) \exp(\theta(u)^{\mathsf{T}} S(x, y, z) - G(\theta(u))),$$

where $\theta(\cdot)$ is a nonlinear function that projects u to a higher dimension space. In (19) the function $S(x, y, z)$ can be written as a composition of eight elements: y^{-1}, y, $\log y$, x, xy^{-1}, xx^{T}, $xz^{\mathsf{T}}y^{-1/2}$, $zy^{-1/2}$, $zy^{1/2}$ and zz^{T}. Similar as before let us replace these functions of x, y and z by $s = (s_1, s_2, \ldots, s_8)$ respectively, and denote the set of parameters as $u = (\mu, \gamma, F, D, \lambda, \chi, \psi)$ for simplicity. Then the log-likelihood function of factor analysis can be defined as:

$$\begin{aligned} L_{FA}(u|s) = &-\frac{1}{2}\log|D| - \frac{1}{2}\mu^{\mathsf{T}}D^{-1}\mu s_1 - \frac{1}{2}\gamma^{\mathsf{T}}D^{-1}\gamma s_2 + \gamma^{\mathsf{T}}D^{-1}s_4 + \mu^{\mathsf{T}}D^{-1}s_5 \\ &- \frac{1}{2}tr(D^{-1}s_6) + tr(F^{\mathsf{T}}D^{-1}s_7) - \mu^{\mathsf{T}}D^{-1}Fs_8 - \gamma^{\mathsf{T}}D^{-1}Fs_9 \\ &- \frac{1}{2}tr(F^{\mathsf{T}}D^{-1}Fs_{10}) - \mu^{\mathsf{T}}D^{-1}\gamma + L_{GIG}(\lambda, \chi, \psi|s_1, s_2, s_3). \end{aligned}$$

Unlike a full exponential family that always has one to one projection between the natural parameters θ and the expectation parameters s, maximizing the log-likelihood function of a curved exponential family projects the expectation parameters to a lower dimension space. We refer [2] for detail.

The partial derivatives of L_{FA} with respect to μ, γ, F, D are:

$$\begin{aligned}
\frac{\partial L_{FA}}{\partial \mu} &= D^{-1}(-s_1\mu + s_5 - Fs_8 - \gamma),\\
\frac{\partial L_{FA}}{\partial \gamma} &= D^{-1}(-s_2\gamma + s_4 - Fs_9 - \mu),\\
\frac{\partial L_{FA}}{\partial F} &= D^{-1}(s_7 - \mu(s_8)^\mathsf{T} - \gamma s_9^\mathsf{T} - Fs_{10}),\\
\frac{\partial L_{FA}}{\partial D^{-1}} &= \frac{1}{2}\big(- s_1\mu\mu^\mathsf{T} - s_2\gamma\gamma^\mathsf{T} + s_4\gamma^\mathsf{T} + \gamma(s_4)^\mathsf{T} + s_5\mu^\mathsf{T} + \mu s_5^\mathsf{T} - s_6 + s_7F^\mathsf{T}\\
&\quad + Fs_7^\mathsf{T} - Fs_8\mu^\mathsf{T} - \mu Fs_8^\mathsf{T} - Fs_9\gamma^\mathsf{T} - \gamma Fs_9^\mathsf{T} - Fs_{10}F^\mathsf{T} - \mu\gamma^\mathsf{T} - \gamma\mu^\mathsf{T} - D\big).
\end{aligned}$$

By setting the derivatives to be zero we get the analytic formulas for the projection:

$$\begin{aligned}
\mu &= \frac{t_2t_5 - t_3t_4}{t_2^2 - t_1t_3},\\
\gamma &= \frac{t_2t_4 - t_1t_5}{t_2^2 - t_1t_3},,\\
F &= \big(s_7 - \mu s_8^\mathsf{T} - \gamma s_9^\mathsf{T}\big)s_{10}^{-1},\\
D &= diag\big(s_1\mu\mu^\mathsf{T} + s_2\gamma\gamma^\mathsf{T} - s_4\gamma^\mathsf{T} - \gamma s_4^\mathsf{T}\\
&\quad - s_5\mu^\mathsf{T} - \mu s_5^\mathsf{T} + s_6 - s_7F^\mathsf{T} - Fs_7^\mathsf{T},\\
&\quad + Fs_8\mu^\mathsf{T} + \mu(Fs_8)^\mathsf{T} + Fs_9\gamma^\mathsf{T} + \gamma(Fs_9)^\mathsf{T},\\
&\quad + Fs_{10}F^\mathsf{T} + \mu\gamma^\mathsf{T} + \gamma\mu^\mathsf{T}\big),\\
(\lambda, \chi, \psi) &= \arg\max_{\lambda,\chi,\psi} L_{GIG}(\lambda, \chi, \psi | s_1, s_2, s_3),
\end{aligned} \tag{20}$$

where

$$\begin{aligned}
t_1 &= s_8^\mathsf{T}s_{10}^{-1}s_8 - s_1,\\
t_2 &= s_9^\mathsf{T}s_{10}^{-1}s_8 - 1,\\
t_3 &= s_9^\mathsf{T}s_{10}^{-1}s_9 - s_2,\\
t_4 &= s_7^\mathsf{T}s_{10}^{-1}s_8 - s_5,\\
t_5 &= s_7^\mathsf{T}s_{10}^{-1}s_9 - s_4.
\end{aligned} \tag{21}$$

Now let us consider the M-step of the EM algorithm. It is clear that if we integrate (19) over z we can find that the joint distribution of x and y is given by (10) where $\Sigma = FF^{\mathsf{T}} + D$. As a result, the conditional distribution of y given x follows the GIG with parameter $(\lambda - d/2, \chi + (x-\mu)^{\mathsf{T}}(FF^{\mathsf{T}} + D)^{-1}(x-\mu), \psi + \gamma^{\mathsf{T}}(FF^{\mathsf{T}} + D)^{-1}\gamma)$. So similar as before we have:

$$E[Y^{\alpha}|X,u] = \left(\sqrt{\frac{\chi + (x-\mu)^{\mathsf{T}}(FF^{\mathsf{T}} + D)^{-1}(x-\mu)}{\psi + \gamma^{\mathsf{T}}(FF^{\mathsf{T}} + D)^{-1}\gamma}}\right)^{\alpha} \frac{K_{\lambda-\frac{d}{2}+\alpha}(\sqrt{(\chi + (X-\mu)^{\mathsf{T}}(FF^{\mathsf{T}} + D)^{-1}(X-\mu))(\psi + \gamma^{\mathsf{T}}(FF^{\mathsf{T}} + D)^{-1}\gamma)})}{K_{\lambda-\frac{d}{2}}(\sqrt{(\chi + (X-\mu)^{\mathsf{T}}(FF^{\mathsf{T}} + D)^{-1}(X-\mu))(\psi + \gamma^{\mathsf{T}}(FF^{\mathsf{T}} + D)^{-1}\gamma)})}, \tag{22}$$

$$E[\log Y|X,u] = \frac{\partial}{\partial\alpha}E[Y^{\alpha}|X,u]\Big|_{\alpha=0}.$$

On the other side, the joint distribution of X and Z conditional on Y follows the Gaussian distribution:

$$\begin{pmatrix} (X-\mu-\gamma Y)/\sqrt{Y} \\ Z \end{pmatrix}\Bigg|_{Y,u} \sim N\Bigg(0, \begin{pmatrix} FF^{\mathsf{T}} + D & F \\ F^{\mathsf{T}} & I \end{pmatrix}\Bigg).$$

Thus it is easy to obtain:

$$E[Z|X,Y,u] = \beta\left(\frac{X-\mu-\gamma Y}{\sqrt{Y}}\right),$$

and

$$E[ZZ^{\mathsf{T}}|X,Y,u] = I - \beta F + \beta(X-\mu-\gamma Y)(X-\mu-\gamma Y)^{\mathsf{T}}\beta^{\mathsf{T}}Y^{-1},$$

where

$$\beta = F^{\mathsf{T}}(FF^{\mathsf{T}} + D)^{-1}.$$

So we solve the rest of conditional expectations that we need in the EM-algorithm:

$$\begin{aligned} E[ZY^{-\frac{1}{2}}|X,u] &= \beta(X-\mu)E[Y^{-1}|X,u] - \beta\gamma, \\ E[ZY^{\frac{1}{2}}|X,u] &= \beta(X-\mu) - \beta\gamma E[Y|X,u], \\ E[ZZ^{\mathsf{T}}|X,u] &= I - \beta F + \beta(X-\mu)(X-\mu)^{\mathsf{T}}\beta^{\mathsf{T}}E[Y^{-1}|X,u] \\ &\quad - \beta(X-\mu)\gamma^{\mathsf{T}}\beta^{\mathsf{T}} - \beta\gamma(X-\mu)^{\mathsf{T}}\beta^{\mathsf{T}} + \beta\gamma\gamma^{\mathsf{T}}\beta^{\mathsf{T}}E[Y|X,u]. \end{aligned}$$

Now suppose that we have i.i.d samples: $x_1, \ldots, x_n$. The k-th E-step computes the conditional expectation of all ten sufficient statistics given $x_1, \ldots, x_n$. The first six is the same as before:

$$
\begin{aligned}
s_1^{(k)} &= \frac{1}{n}\sum_{j=1}^{n} E[Y^{-1}|X = x_j, u_k], \\
s_2^{(k)} &= \frac{1}{n}\sum_{j=1}^{n} E[Y|X = x_j, u_k], \\
s_3^{(k)} &= \frac{1}{n}\sum_{j=1}^{n} E[\log Y|X = x_j, u_k], \\
s_4^{(k)} &= \frac{1}{n}\sum_{j=1}^{n} x_j, \\
s_5^{(k)} &= \frac{1}{n}\sum_{j=1}^{n} x_j E[Y^{-1}|X = x_j, u_k], \\
s_6^{(k)} &= \frac{1}{n}\sum_{j=1}^{n} x_j x_j^{\mathsf{T}} E[Y^{-1}|X = x_j, u_k],
\end{aligned}
$$

where $u_k = (\mu_k, \gamma_k, F_k, D_k, \lambda_k, \chi_k, \psi_k)$ is the result of the previous iteration. The rest four sufficient statistics are all determined by the first six ones:

$$
\begin{aligned}
s_7^{(k)} &= \frac{1}{n}\sum_{j=1}^{n} E[XZ^{\mathsf{T}}Y^{-\frac{1}{2}}|X = x_j, u_k] = (s_6^{(k)} - s_5^{(k)}\mu_k^{\mathsf{T}} - s_4^{(k)}\gamma_k^{\mathsf{T}})\beta_k^{\mathsf{T}}, \\
s_8^{(k)} &= \frac{1}{n}\sum_{j=1}^{n} E[ZY^{-\frac{1}{2}}|X = x_j, u_k] = \beta_k(s_5^{(k)} - \mu_k s_1^{(k)} - \gamma_k), \\
s_9^{(k)} &= \frac{1}{n}\sum_{j=1}^{n} E[ZY^{\frac{1}{2}}|X = x_j, u_k] = \beta_k(s_4^{(k)} - \mu_k - \gamma_k S_2^{(k)}), \\
s_{10}^{(k)} &= \frac{1}{n}\sum_{j=1}^{n} E[ZZ^{\mathsf{T}}|X = x_j, u_k] \\
&= I - \beta_k F_k + \beta_k\Big(S_6^{(k)} - S_5^{(k)}\mu_k^{\mathsf{T}} - \mu_k(s_5^{(k)})^{\mathsf{T}} + \mu_k\mu_k^{\mathsf{T}}s_1^{(k)} \\
&\quad - (s_4^{(k)} - \mu_k)\gamma_k^{\mathsf{T}} - \gamma_k(s_4^{(k)} - \mu)^{\mathsf{T}} + \gamma_k\gamma_k^{\mathsf{T}}s_2^{(k)}\Big)\beta_k^{\mathsf{T}}.
\end{aligned}
$$

The M-step maximizes the following log-likelihood function:

$$u_{k+1} = \arg\max_{u} \sum_{j=1}^{n} L_{FA}(u_k | s^{(k)}),$$

where $s^{(k)} = (s_1^{(k)}, \ldots, s_8^{(k)})$. The solution of the optimization problem is given by (20) and (21).

To test the factor analysis for GH, we first fit standardized daily returns of Dow Jones 30 companies from 2008 to 2015 via the original EM algorithm where λ is fixed to be $-10, -9, \ldots, 9, 10$. Like before we generate 5000 i.i.d samples for each set of these estimated parameters. Then we fit the samples by the EM algorithms with and without factor analysis. In addition, we apply the principle component analysis (PCA) to the positive definite matrix Σ without factor analysis. Given the singular value decomposition (SVD): $\Sigma = USU^{\mathsf{T}}$ where $U \in \mathbb{R}^{d\times d}$ is unitary and $S \in \mathbb{R}^{d\times d}$ is diagonal such that the singular values $S_{11} \geq S_{22} \geq \cdots \geq S_{dd}$, the PCA of Σ returns a structured positive definite matrix:

$$\Sigma_{PCA} = FF^{\mathsf{T}} + D,$$

where

$$F = U_r S_r^{\frac{1}{2}},$$
$$D = diag(\Sigma - FF^{\mathsf{T}}),$$

$U_r \in \mathbb{R}^{d\times r}$ are the first r columns of U and $S_r \in \mathbb{R}^{r\times r}$ is the first r by r block of S. By doing this we obtain an approximation Σ_{PCA} with improved condition numbers. This is a naive approach to impose the structure to the matrix Σ. Table 5 compares this method to the factor analysis. The columns "EM", "FA" and "PCA" denote the original EM algorithm, the EM algorithm for factor analysis and the PCA after the original EM algorithm. Columns 2-3 list the squared Hellinger distance between the estimated distribution and the ture one. Columns 4-6 are the condition numbers of Σ generated by three algorithms. The last three columns are the average log-likelihood of these parameters.

	H^2			Condition number			Log-likelihood		
λ	EM	FA	PCA	EM	FA	PCA	EM	FA	PCA
-10	0.1000	0.1729	0.3754	1994.20	867.10	691.08	4.7370	4.4345	3.4113
-9	0.0705	0.1393	0.3452	1974.33	910.35	652.46	5.0906	4.7829	3.7616
-8	0.0802	0.1595	0.3652	1954.36	848.52	699.37	5.1395	4.8390	3.7902
-7	0.0530	0.1240	0.3275	2087.17	908.11	697.45	4.9642	4.6519	3.6912
-6	0.0413	0.1181	0.3264	1880.44	825.69	652.00	4.9823	4.6854	3.7247
-5	0.0527	0.1262	0.3405	1941.21	857.07	674.09	4.7375	4.4521	3.4715
-4	0.0328	0.1044	0.3363	1865.89	889.10	647.46	4.8930	4.6235	3.5839
-3	0.0273	0.1035	0.3349	2045.80	880.74	687.54	4.2625	3.9536	2.9333
-2	0.0255	0.0975	0.3301	1957.20	892.15	677.09	3.1803	2.9007	1.8659
-1	0.0260	0.1018	0.3185	1906.74	858.30	670.68	2.7241	2.4332	1.4935
0	0.0233	0.0968	0.3311	1913.15	866.64	663.03	2.6073	2.3293	1.3293
1	0.0236	0.0921	0.3191	2109.98	938.91	699.25	2.1358	1.8556	0.8953
2	0.0286	0.1093	0.3157	2030.69	851.70	698.08	2.1970	1.8906	0.9731
3	0.0343	0.1028	0.3288	1989.78	908.03	673.02	1.7883	1.5097	0.5330
4	0.0365	0.1135	0.3217	1935.65	825.38	667.19	1.1949	0.8893	-0.0374
5	0.0470	0.1209	0.3320	1990.66	925.24	687.06	1.4294	1.1380	0.1686
6	0.0524	0.1261	0.3353	2033.83	893.15	681.75	1.2452	0.9458	-0.0203
7	0.0525	0.1216	0.3369	2031.71	939.93	705.08	1.4742	1.1971	0.2129
8	0.0662	0.1487	0.3414	2004.60	876.78	689.10	1.5345	1.2367	0.2912
9	0.0917	0.1545	0.3685	2023.68	914.30	688.21	1.4006	1.1069	0.0266
10	0.0908	0.1543	0.3631	2057.76	959.67	677.09	1.7179	1.4378	0.4026

Table 5: Analysis of three methods

First it is clear that the factor analysis that maximize the likelihood with the constraint $\Sigma = FF^{\mathsf{T}} + D$ always has a larger likelihood than the PCA. The original EM algorithm achieves the maximum likelihood without any constraint on Σ. In addition it also has the smallest Hellinger distances among three columns while the factor analysis has a better performance than the naive PCA. On the other side the condition number of the original EM algorithm is huge, this corresponds to our discussion that the MLE of the parameter Σ might be ill-conditioned just like the sample covariance matrix. Σ given by the factor analysis has much smaller condition numbers. Although they are larger than the ones of PCA they are still significant improvements of the matrix conditions.

There are two problems in the above test. First the parameters which we used to generate samples are estimated via the EM algorithm without factor analysis; thus the "true" matrix Σ has relative large condition number itself. Secondly, the factor analysis for Gaussian distribution is usually known as a better choice than the sample covariance when sample size is small. Usually we do not have 5000 sample size in low frequency finance.

So we modify the above test to see under what situations the factor analysis might outperform the MLE. Instead of generating random samples right after fitting the 30 stocks via the EM algorithm, we first shrink the singular values of the estimated Σ so that it is well-conditioned. Then we rescale it so that its determinant is one again. 1000 i.i.d samples are generated based on this shrinked Σ. The results are shown by table 6, from which one can observe that the condition numbers of all three approaches are much smaller than before. While the Hellinger distances of the factor analysis are overall smaller than the ones in the "EM" column, which corresponds to our conjecture. This does not mean, of course, that factor analysis would always be better than MLE. As the sample size increases, the MLE will converges to the true distribution while the factor analysis which is a biased estimator may never reach to the true one even it can be quite close.

	H^2			Condition number			Log-likelihood		
λ	EM	FA	PCA	EM	FA	PCA	EM	FA	PCA
-10	0.3032	0.2981	0.5138	68.6334	65.0819	67.6515	5.1644	5.0416	4.0143
-9	0.3565	0.3306	0.5311	64.8407	62.5625	65.7507	5.4475	5.3448	4.4495
-8	0.2914	0.2723	0.5091	63.5038	60.1229	64.1358	5.2004	5.0930	4.0253
-7	0.3417	0.3123	0.5174	57.6966	56.7443	61.2092	4.9359	4.8376	3.9381
-6	0.1645	0.1474	0.3998	66.3506	64.1371	66.9552	5.0843	4.9768	4.0433
-5	0.1876	0.1718	0.4303	63.7664	61.4185	69.9162	4.9568	4.8572	3.9241
-4	0.1599	0.1494	0.4023	60.6704	59.3149	62.8858	4.6953	4.6183	3.6512
-3	0.1182	0.1032	0.3583	63.1663	61.1375	64.5315	4.4342	4.3238	3.4138
-2	0.1268	0.1149	0.3881	63.4210	61.3255	66.8355	3.3633	3.2623	2.3230
-1	0.1345	0.1220	0.3828	68.9667	66.8575	69.8567	2.8859	2.7807	1.8908
0	0.1094	0.0954	0.3512	65.5393	61.3081	65.5617	3.0521	2.9281	2.0972
1	0.1344	0.1232	0.3625	62.5816	59.9246	65.0075	1.9517	1.8444	0.9857
2	0.1295	0.1128	0.3855	67.3400	65.0696	68.7618	2.6876	2.5931	1.7133
3	0.1232	0.1131	0.3780	57.1453	55.0849	58.3601	1.9194	1.8123	0.8495
4	0.1811	0.1654	0.4030	61.6218	60.0264	62.0237	1.9479	1.8484	0.8972
5	0.1524	0.1476	0.3882	78.3277	72.5415	75.3105	1.7491	1.6246	0.6563
6	0.2114	0.2063	0.4424	66.5814	63.5592	67.6542	1.6017	1.5143	0.5993
7	0.2685	0.2748	0.4749	60.4194	58.2861	62.7466	1.5260	1.4329	0.5920
8	0.3047	0.3015	0.4987	66.5735	63.8823	68.2195	1.8319	1.7368	0.7560
9	0.4188	0.4221	0.5804	71.9542	70.3681	73.6883	1.6385	1.5272	0.5244
10	0.3389	0.3093	0.5182	62.8962	60.3928	65.6060	2.0215	1.9075	0.9179

Table 6: Analysis of three methods with shrinked Σ

3.4 On-line EM algorithm for the Exponential Families

In this section we review the on-line EM algorithm proposed by [9]. In addition we define and compute the regret of the algorithm. Let us consider the exponential families whose density function has the form:

$$p(x, y|\theta) = h(x, y) \exp(\theta^{\mathsf{T}} S(x, y) - G(\theta)), \tag{23}$$

where x is observable, y is hidden and

$$G(\theta) = \log \int h(x, y) \exp(\theta^{\mathsf{T}} S(x, y)) dx dy.$$

The expectation parameter or the dual-parameter is given by $\eta = \nabla G(\theta)$; the dual of G is given by $F(\eta) = \eta^{\mathsf{T}}\theta - G(\theta)$ and $\theta = \nabla F(\eta)$. Furthermore one can show that $D_G(\theta_1 \| \theta_2) = D_F(\eta_2 \| \eta_1)$ where D_G and D_F are the Bregman divergences with potential functions G and F.

We can think the process of parameter estimation as a game. Suppose that at time $t-1$ we make a decision or prediction denoted by θ_{t-1}. Then at time t the environment reveals an observation x_t. The loss of our decision θ_{t-1} is then given by the function $l(x_t, \theta_{t-1})$. Our goal is to find a sequence of strategies θ_t that minimize the regret from $t = 1, \ldots, T$:

$$\sum_{t=1}^{T} l(x_t, \theta_{t-1}) - \min_{\theta} \sum_{t=1}^{T} l(x_t, \theta).$$

We refer [11] for a detailed description of the above setup. If we define loss function l as the minus log-likelihood function, then we transfer the parameter estimation to an on-line optimization problem. In fact [4] defines the regret of online density estimation for an exponential family as:

$$-\sum_{t=1}^{T} \log p(x_t, y_t | \theta_{t-1}) - \min_{\theta} \Big(-\sum_{t=1}^{T} \log p(x_t, y_t | \theta) + \tau_0 D_G(\theta \| \theta_0) \Big),$$

where $D_G(\theta \| \theta_0) = G(\theta) - G(\theta_0) + \nabla G(\theta_0)^{\mathsf{T}}(\theta - \theta_0)$ is the Bregman divergence with the potential function G, which is also the Kullback-Leibler divergence between $p(x, y|\theta_0)$ and $p(x, y|\theta)$. The Bregman divergence in the above equation denotes the penalty function as we discussed in section 3.2.

However y is assumed to be unobservable. Hence we have to replace the joint density in the regret function by the marginal density of x:

$$p(x|\theta) = \int p(x, y|\theta) dy,$$

and the regret is defined as:

$$\begin{aligned} &r_T(\theta_0, \ldots, \theta_{T-1}) := \\ &-\sum_{t=1}^{T} \log p(x_t | \theta_{t-1}) - \min_{\theta} \Big(-\sum_{t=1}^{T} \log p(x_t | \theta) + \tau_0 D_G(\theta \| \theta_0) \Big). \end{aligned} \tag{24}$$

[9] proposes an online EM algorithm for exponential families with hidden data:

$$\begin{aligned}
&\eta_0 = \nabla G(\theta_0), \\
&\eta_t = \eta_{t-1} + \tau_t^{-1}(\bar{S}(x_t|\theta_{t-1}) - \eta_{t-1}), \\
&\theta_t = \nabla F(\eta_t),
\end{aligned} \tag{25}$$

where

$$\bar{S}(x_t|\theta_{t-1}) := E[S(X,Y)|X = x_t, \theta_{t-1}],$$

F is the Legendre dual function of G, ∇G and ∇F are the gradients of G and F respectively.

Theorem 1. *Let $\tau_t = \tau_0 + t$, then the regret defined by (24) of the online EM algorithm (25) is given by:*

$$\begin{aligned}
&r_T(\theta_0, \ldots, \theta_{T-1}) \\
&= \sum_{t=1}^{T} \tau_t D_G(\theta_{t-1} \| \theta_t) + \sum_{t=1}^{T} D_{KL}(x_t, \theta_{t-1} \| x_t, \theta_{ML}) - \tau_T D_F(\eta_T \| \eta_{ML}),
\end{aligned} \tag{26}$$

where

$$\theta_{ML} := \arg\min_{\theta} \Big(- \sum_{t=1}^{T} \log p(x_t|\theta) + \tau_0 D_G(\theta \| \theta_0) \Big),$$

is the MLE given sample $x_1, \ldots, x_T$, $\eta_{ML} = \nabla G(\theta_{ML})$ and $D_{KL}(x_t, \theta_{t-1} \| x_t, \theta_{ML})$ denotes the Kullback-Leibler divergence between $p(y|x_t, \theta_{t-1})$ and $p(y|x_t, \theta_{ML})$.

Proof. First note that:

$$\begin{aligned}
-\sum_{t=1}^{T} \log p(x_t|\theta_{t-1}) = &- \sum_{t=1}^{T} E[\log p(X,Y|\theta_{t-1})|X = x_t, \theta_{t-1}] \\
&+ \sum_{t=1}^{T} E[\log p(Y|X, \theta_{t-1})|X = x_t, \theta_{t-1}],
\end{aligned}$$

since $p(y|x,\theta) = p(x,y|\theta)/p(x|\theta)$. The conditional expectation of the log of the joint density is:

$$
\begin{aligned}
&-\sum_{t=1}^{T} E[\log p(X,Y|\theta_{t-1})|X = x_t, \theta_{t-1}] \\
&= \sum_{t=1}^{T} (G(\theta_{t-1}) - \theta_{t-1}^{\mathsf{T}} \bar{S}(x_t|\theta_{t-1})) - \sum_{t=1}^{T} E[\log h(X,Y)|X = x_t, \theta_{t-1}],
\end{aligned}
$$

where

$$
\begin{aligned}
&\sum_{t=1}^{T} (G(\theta_{t-1}) - \theta_{t-1}^{\mathsf{T}} \bar{S}(x_t|\theta_{t-1})) \\
&= \sum_{t=1}^{T} (-F(\eta_{t-1}) - \theta_{t-1}^{\mathsf{T}} (\bar{S}(x_t|\theta_{t-1}) - \eta_{t-1})) \\
&= \sum_{t=1}^{T} (-F(\eta_{t-1}) - \tau_t \theta_{t-1}^{\mathsf{T}} (\eta_t - \eta_{t-1})) \\
&= \sum_{t=1}^{T} (\tau_t (F(\eta_t) - F(\eta_{t-1}) - \theta_{t-1}^{\mathsf{T}} (\eta_t - \eta_{t-1})) - \tau_t F(\eta_t) + (\tau_t - 1) F(\eta_{t-1})) \\
&= \sum_{t=1}^{T} (\tau_t D_F(\eta_t \| \eta_{t-1}) - \tau_t F(\eta_t) + \tau_{t-1} F(\eta_{t-1})) \\
&= \sum_{t=1}^{T} \tau_t D_G(\theta_{t-1} \| \theta_t) + \tau_0 F(\eta_0) - \tau_T F(\eta_T).
\end{aligned}
$$

Here we use the fact that $F(\eta) = \theta^{\mathsf{T}} \eta - G(\theta)$. So we have:

$$
\begin{aligned}
&-\sum_{t=1}^{T} \log p(x_t|\theta_{t-1}) = \sum_{t=1}^{T} \tau_t D_G(\theta_{t-1} \| \theta_t) + \tau_0 F(\eta_0) - \tau_T F(\eta_T) \\
&-\sum_{t=1}^{T} E[\log h(X,Y)|X = x_t, \theta_{t-1}] + \sum_{t=1}^{T} E[\log p(Y|X, \theta_{t-1})|X = x_t, \theta_{t-1}],
\end{aligned}
$$

as the first part of the regret. The second part can be written as:

$$\min_{\theta}\Big(-\sum_{t=1}^{T}\log p(x_t|\theta)+\tau_0 D_G(\theta\|\theta_0)\Big)=-\sum_{t=1}^{T}E[\log p(X,Y|\theta_{ML})|X=x_t,\theta_{t-1}]$$
$$+\sum_{t=1}^{T}E[\log p(Y|X,\theta_{ML})|X=x_t,\theta_{t-1}]+\tau_0 D_G(\theta_{ML}\|\theta_0).$$

On the other side we have:

$$-\sum_{t=1}^{T}E[\log p(X,Y|\theta_{ML})|X=x_t,\theta_{t-1}]+\tau_0 D_G(\theta_{ML}||\theta_0)$$
$$=\sum_{t=1}^{T}(G(\theta_{ML})-\theta_{ML}^{\mathsf{T}}\bar{S}(x_t,\theta_{t-1}))+\tau_0 D_G(\theta_{ML}\|\theta_0)-\sum_{t=1}^{T}E[h(X,Y)|X=x_t,\theta_{t-1}],$$

where

$$\sum_{t=1}^{T}(G(\theta_{ML})-\theta_{ML}^{\mathsf{T}}\bar{S}(x_t,\theta_{t-1}))+\tau_0 D_G(\theta_{ML}\|\theta_0)$$
$$=\sum_{t=1}^{T}(G(\theta_{ML})-\theta_{ML}^{\mathsf{T}}\bar{S}(x_t,\theta_{t-1}))+\tau_0 D_F(\eta_0\|\eta_{ML})$$
$$=\sum_{t=1}^{T}(\theta_{ML}^{\mathsf{T}}\eta_{ML}-F(\eta_{ML})-\theta_{ML}^{\mathsf{T}}\bar{S}(x_t,\theta_{t-1}))+\tau_0(F(\eta_0)-F(\eta_{ML})-(\eta_0-\eta_{ML})^{\mathsf{T}}\theta_{ML})$$
$$=\tau_0 F(\eta_0)-\tau_T F(\eta_{ML})+\tau_T\theta_{ML}^{\mathsf{T}}\eta_{ML}-\theta_{ML}^{\mathsf{T}}\Big(\tau_0\eta_0+\sum_{t=1}^{T}\bar{S}(x_t,\theta_{t-1})\Big)$$
$$=\tau_0 F(\eta_0)-\tau_T F(\eta_{ML})+\tau_T\theta_{ML}^{\mathsf{T}}(\eta_{ML}-\eta_T)$$
$$=\tau_0 F(\eta_0)-\tau_T F(\eta_T)+\tau_T D_F(\eta_T\|\eta_{ML}).$$

Here we use the fact that η_T satisfies:

$$\eta_T=\tau_T^{-1}\Big(\tau_0\eta_0+\sum_{t=1}^{T}\bar{S}(x_t,\theta_{t-1})\Big).$$

As a result the second part of the regret is given by:

$$\min_{\theta}\Big(-\sum_{t=1}^{T}\log p(x_t|\theta)+\tau_0 D_G(\theta\|\theta_0)\Big)=\tau_0 F(\eta_0)-\tau_T F(\eta_T)+\tau_T D_F(\eta_T\|\eta_{ML})$$
$$-\sum_{t=1}^{T}E[h(X,Y)|X=x_t,\theta_{t-1}]+\sum_{t=1}^{T}E[\log p(Y|X,\theta_{ML})|X=x_t,\theta_{t-1}].$$

Finally by combining the two parts we obtain (26). □

3.5 On-line EM algorithm for the Generalized Hyperbolic Distributions

Now we consider the joint GH distribution (10) which is an exponential family. Suppose that we have a sequence of observations $x_1, x_2, \ldots, x_T$, it is not difficult to see that the on-line EM algorithm for the GH updates the sufficient statistics as follows:

$$\begin{aligned}
s_1^{(t+1)} &= s_1^{(t)} + \tau_{t+1}^{-1}(E[Y^{-1}|X=x_{t+1},\theta_t]-s_1^{(t)}),\\
s_2^{(t+1)} &= s_2^{(t)} + \tau_{t+1}^{-1}(E[Y|X=x_{t+1},\theta_t]-s_2^{(t)}),\\
s_3^{(t+1)} &= s_3^{(t)} + \tau_{t+1}^{-1}(E[\log Y|X=x_{t+1},\theta_t]-s_3^{(t)}),\\
s_4^{(t+1)} &= s_4^{(t)} + \tau_{t+1}^{-1}(x_j-s_4^{(t)})\\
s_5^{(t+1)} &= s_5^{(t)} + \tau_{t+1}^{-1}(x_j E[Y^{-1}X=x_{t+1},\theta_t]-s_5^{(t)}),\\
s_6^{(t+1)} &= s_6^{(t)} + \tau_{t+1}^{-1}(x_j x_j^{\mathsf{T}} E[Y^{-1}|X=x_{t+1},\theta_t]-s_6^{(t)}),
\end{aligned}$$

where the conditional expectations are given by (15). And the parameters $\theta_t=(\mu_t,\gamma_t,\Sigma_t,\lambda_t,\chi_t,\psi_t)$ are given by:

$$\begin{aligned}
\mu_t &= \frac{s_4^{(t)}-s_2^{(t)}s_5^{(t)}}{1-s_1^{(t)}s_2^{(t)}},\\
\gamma_t &= \frac{s_5^{(t)}-s_1^{(t)}s_4^{(t)}}{1-s_1^{(t)}s_2^{(t)}},\\
\Sigma_t &= s_6^{(t)}-s_5^{(t)}\mu_t^{\mathsf{T}}-\mu_t(s_5^{(t)})^{\mathsf{T}}+s_1^{(t)}\mu_t\mu_t^{\mathsf{T}}-s_2^{(t)}\gamma_t\gamma_t^{\mathsf{T}},
\end{aligned}$$

and

$$(\lambda_t, \chi_t, \psi_t) = \arg \min_{\lambda \in \mathbb{R}, \chi, \psi > 0} -\frac{\chi}{2} s_1^{(t)} - \frac{\psi}{2} s_2^{(t)} + \lambda s_3^{(t)} - \frac{\lambda}{2} \log \chi + \frac{\lambda}{2} \log \psi - \log K_\lambda(\sqrt{\chi\psi}).$$

Here is an numerical example for the on-line EM algorithm. Similar as before we first fit standardized daily returns of Dow Jones 30 stocks from 2008 to 2015. This time we will not fix λ for simplicity. Then we generate 1000 i.i.d samples from the estimated parameters. We divide the samples into two parts: the first 500 samples are fitted by the EM algorithm to get a set of initial parameters. Then we update these parameters via the on-line algorithm using the rest 500 samples. For each step we will get a new set of parameters. For comparison we also use the original EM algorithm to fit the sample with increasing size, i.e, at step t we will fit first $500 + t$ samples. The coefficient τ_t is also set to be $500 + t$. At each step we compute the squared Hellinger distances between our estimated parameters and the true distribution. It is clear that the Hellinger distance would decrease as new samples come in. Figure 4 compares the squared Hellinger distances of the original EM and the on-line EM algorithm. Although the Hellinger distances of the EM algorithm with increasing sample size are overall smaller than the ones of the on-line EM, both of them exhibit a similar decreasing rate. The differences between them are negligible.

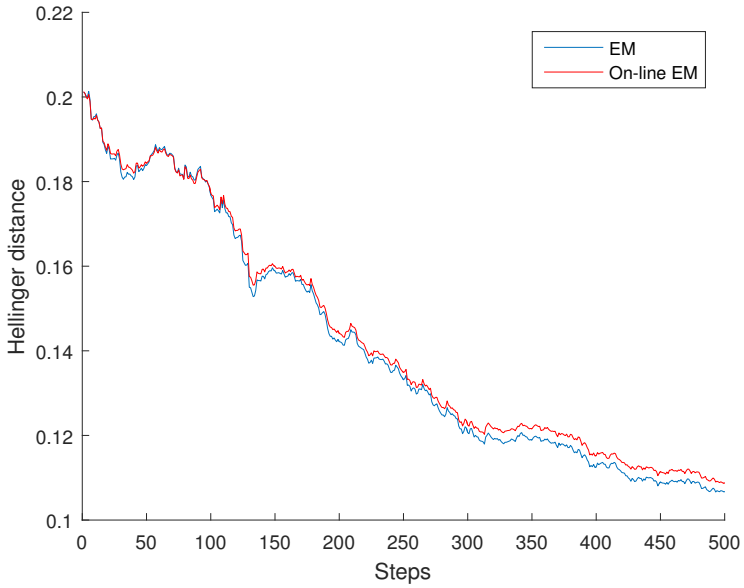


Figure 4: Squared Hellinger distance of the on-line EM algorithm

We also compute the cumulative loss of the two algorithms. The loss of the algorithm at step t is given by the minus log-likelihood of the estimated parameters given the data point at $t + 1$. Thus the original EM algorithm is not guaranteed to have the smallest loss. Table 7 lists the cumulative loss together with the Hellinger distance of two algorithms on step 50 to 500. Interestingly the cumulative loss of the EM algorithm is slightly larger than the loss of the on-line EM algorithm. But the increasing rate is very much the same and the differences are also negligible. On the other side, the speed of the on-line EM is much faster than the EM algorithm. One would expect that the EM algorithm would be slower as the sample size increases; while the speed of stepwise on-line EM algorithm is not affected by the sample size at all.

	H^2		Cumulative loss	
Steps	EM	On-line EM	EM	On-line EM
50	0.2260	0.2294	1262.1595	1262.0578
100	0.2085	0.2129	2455.9476	2454.4570
150	0.1854	0.1900	3534.6732	3532.4671
200	0.1658	0.1707	4707.1562	4703.7384
250	0.1463	0.1517	5905.4282	5900.1418
300	0.1445	0.1490	7065.7341	7059.2635
350	0.1381	0.1426	8243.4580	8235.8861
400	0.1385	0.1428	9369.5228	9360.6765
450	0.1308	0.1348	10435.9610	10425.5031
500	0.1123	0.1168	11583.7467	11571.7180

Table 7: Squared Hellinger distance and cumulative loss of the on-line EM

One drawback of the on-line EM algorithm proposed by [9] is that it may not converge for some curved exponential families that have the form:

$$p(x, y|u) = h(x, y) \exp(\theta(u)^\mathsf{T} S(x, y) - G(\theta(u))),$$

where $\theta(u)$ is a non-linear function which projects the parameter u to a higher dimensional space. The EM algorithm for curved exponential families solves the equation of u:

$$\nabla\theta(u)^\mathsf{T} \left(\frac{1}{n} \sum_{j=1}^{n} E[S(X, Y)|x_j, \theta(u)] - \nabla G(\theta(u)) \right) = 0,$$

while the on-line EM algorithm tends to solve:

$$\frac{1}{n} \sum_{j=1}^{n} E[S(X, Y)|x_j, \theta(u)] - \nabla G(\theta(u)) = 0,$$

which may not have a solution when the dimension of u is lower than θ. Thus unfortunately we are not be able to apply the on-line EM algorithm introduced in this section to the factor analysis.

3.6 Empirical Studies of the Generalized Hyperbolic Distribution

In this section we test the GH distribution with the estimation algorithms introduced in the previous sections in the real market. The data is the daily adjusted prices of S&P 500 stocks from 2010 to 2015 downloaded from Yahoo! Finance. We compute the standardized log-returns from the price data. Some of the equities that contain missing values are removed. So there are overall 453 equities left. The data is divided into two parts. The first part includes all the data from 2010 to 2013 which will be used for our estimation. The second part includes the data from 2014 to 2015 which will be used for out-of-sample backtesting.

For comparison we also fit the sample data to the NIG, VG and Gaussian distributions. The NIG and VG distributions are the special cases of the GH distribution so we can apply the same EM algorithm to them by introducing additional constraints. Then we generate 10^6 i.i.d random vectors for each distribution. In the meantime we also generate 100 random portfolio weights using uniform random numbers. We use these random weights to project the multivariate GH distributions to the univariate ones. So we are able to compare these univariate distributions to the realized portfolio returns via the two-sample Kolmogorov-Smirnov (KS) and Anderson-Darling (AD) tests. The reason for which we use the two-sample tests instead of one-sample tests is that generating random numbers from the GH distribution is relatively easy and efficient comparing to computing the CDF of GH distribution directly. For the KS test we use the Matlab function `kstest2` while for the AD test we refer [15] for details.

Table 8 shows the results of the KS tests for the original EM algorithm. The second and the fifth column lists the total number of rejections out of 100 in-sample tests and out-of-sample tests respectively. The third and the sixth columns are the average p-values; and the forth and the seventh columns are the average test statistics. Table 9 shows the results of the AD test. We can see that the GH, NIG and VG distributions are better than the Gaussian in both in-sample and out-sample tests. The performances of the three skewed and heavy-tailed distributions are very similar. The GH distribution is overall a bit better than the other two. One might worry that the GH distribution which has more parameters might over-fit the sample, but the its performance in out-of-sample tests are still the best over all four distributions. Figure 5 plots the estimated density function and sample histogram of one random

portfolio; and figure 5 polts its left tail. The differences between the GH, NIG and VG distributions are too small to be observed from the figure for that portfolio, which corresponds to the results of the KS tests.

	In-sample 2010-2013			Out-of-sample 2014-2015		
	Rejections	p-value	Statistics	Rejections	p-value	Statistics
GH	6	0.621031	0.024523	28	0.290718	0.050514
NIG	6	0.620764	0.024534	28	0.290330	0.050555
VG	6	0.620138	0.024564	28	0.289382	0.050574
Gaussian	16	0.395765	0.031707	41	0.218632	0.057003

Table 8: Kolmogorov-Smirnov test of the EM algorithm

	In-sample 2010-2013		Out-of-sample 2014-2015	
	Rejections	Statistics	Rejections	Statistics
GH	6	0.775848619	32	2.418293289
NIG	7	0.777107138	32	2.423416348
VG	7	0.790174126	32	2.432192316
Gaussian	19	1.763408361	43	3.304100218

Table 9: Anderson-Darling test of the EM algorithm

	λ	χ	ψ
GH	-1.8719	1.8396	11.0056
NIG	-0.5	1.4760	15.2859
VG	5.5186	0	35.6627

Table 10: Tail parameters from the EM algorithm

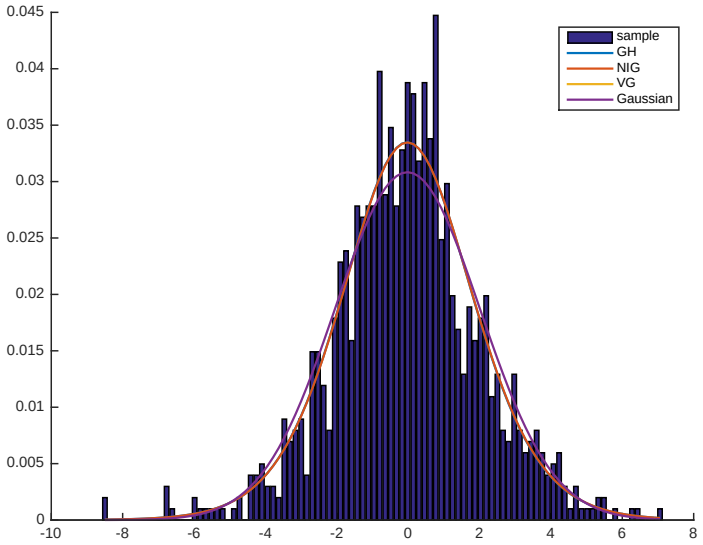


Figure 5: Sample histogram with fitted distributions

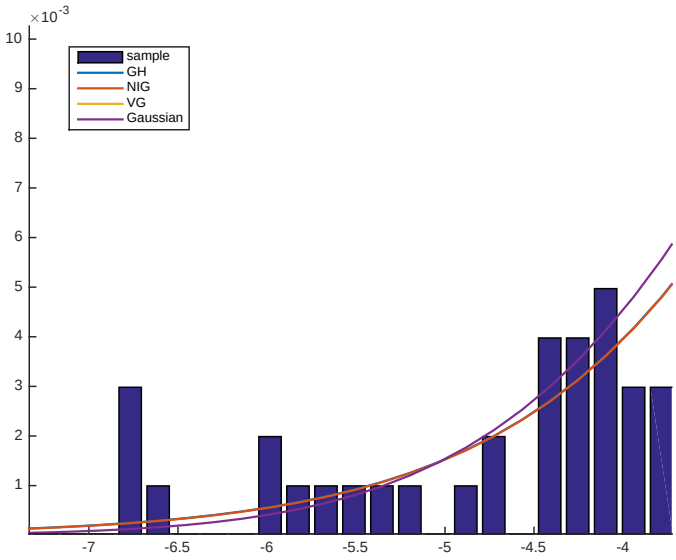


Figure 6: Comparison of the left tails

We also test the factor analysis using the same data and the same portfolio weights. The dimension of the factor is 200. The results are given by table 11 and 12. We find that the factor analysis actually does not improve the performance in both in-sample test and out-of-sample test.

	In sample 2010-2013			Out sample 2014-2015		
	Rejections	p-value	Statistics	Rejections	p-value	Statistics
GH	12	0.586917	0.026097	33	0.278299	0.052536
NIG	12	0.586117	0.026123	33	0.277507	0.052617
VG	12	0.585154	0.026165	33	0.276345	0.052631
Gaussian	16	0.405141	0.031402	40	0.221002	0.056863

Table 11: Kolmogorov-Smirnov test of the factor analysis

	In sample 2010-2013		Out sample 2014-2015	
	Rejections	Statistics	Rejections	Statistics
GH	12	1.098944595	34	2.769573648
NIG	12	1.112162981	34	2.781897982
VG	12	1.130512659	34	2.792557894
Gaussian	19	1.738648795	42	3.294081305

Table 12: Anderson-Darling test of the factor analysis

	λ	χ	ψ
GH	-1.2297	1.7876	11.8080
NIG	-0.5000	1.5797	13.9212
VG	5.3051	0.0000	31.5123

Table 13: Tail parameters from the factor analysis

However the above tests may not tell the whole story. The underlying assumption is that our prediction, i.e. the estimated distribution does not change over next two years. Thus these test are designed for long term predictions. In practice people usually update their prediction in a daily,

weekly or monthly basis. Therefore we are more interested in the accuracy of our prediction in a short term. In that case, the loss function defined in section 2.4 and 2.5 is a good way to evaluate the prediction. Recall that the loss function is defined as:

$$l(x_t, \theta_{t-1}) := -\log p(x_t|\theta_{t-1}),$$

where x_t is the observation on day t, p is the density function of a certain distribution and θ_{t-1} is the parameter estimated using the data before t. p and θ_{t-1} forms our prediction on $t-1$. Under this framework we test the following five strategies using the daily returns from 2014 to 2015:

1. Original EM algorithm with 1000 days' moving window;
2. Original EM algorithm with increasing size window starting with past 1000 days' data;
3. On-line EM algorithm whose initial parameters are given by the EM algorithm using past 1000 days' data;
4. Factor analysis with 1000 days' moving window;
5. Factor analysis with increasing size window starting with past 1000 days' data.

The moving window strategies are the most common approach in fitting financial data. The performance of the moving window strategies are consistent over time so it is easy to evaluate them. The increasing size window strategies are applied to compare with the on-line EM algorithm, as we have seen in the last section. Figure 7 plots the cumulative losses of the five strategies from 2014 to 2015. Note that we ignore the $\log(2\pi)$ constant in computing the log likelihood, so the loss function may not be positive. If we do not ignore that constant then the cumulative loss would be very close to a straight line and the differences between the strategies might not be visually clear.

Figure 7 plots the cumulative losses of the five strategies from 2014 to 2015. Table 14 lists the cumulative losses at the end of each month and the average computation time of each strategy. First of all, the algorithms with increasing window have smaller loss than the ones with moving or fixed size window. This indicates that 1000 days might not be the optimal window size in this case. The on-line EM algorithm's cumulative losses are

overall larger than the EM algorithm with increasing window; but the the difference is relatively small comparing to the EM algorithm with moving window. Surprisingly, the factor analysis which was not doing well in the first test outperforms the other strategies. This indicates that the factor analysis might be a better choice for short-term predictions. On the other side, the factor analysis's convergence speed is slow comparing to the original EM algorithm. The on-line EM algorithm is clearly the fastest one. Interestingly the increasing window strategies are faster then the moving window ones. The reason might be that we use the warm-start approach, i.e. using pervious estimated parameters as the initial value of the current algorithm. The changes of the parameters of the increasing window strategies are smaller than the fixing window ones. This makes the warm-start to be more efficient for the increasing window strategies.

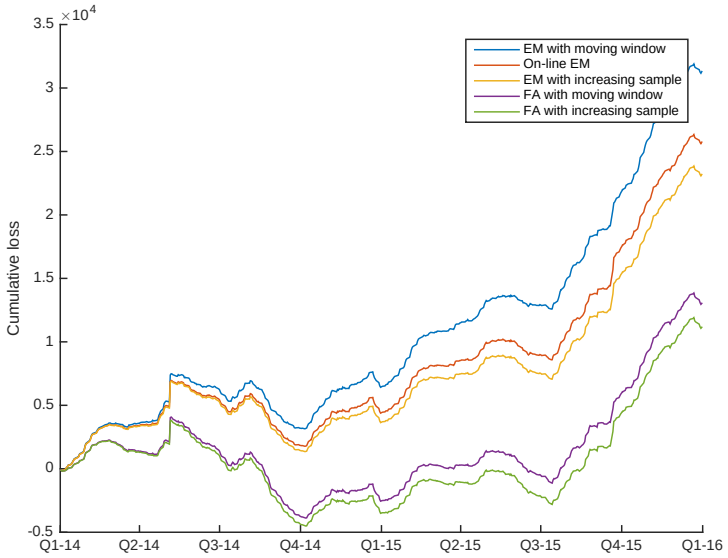


Figure 7: Cumulative losses

Year-Month	Moving EM	On-line EM	Increasing EM	Moving FA	Increasing FA
2014-1	2045.20	1992.91	1990.68	1306.11	1287.60
2014-2	3550.87	3437.04	3420.11	2148.18	2099.02
2014-3	3585.03	3407.00	3349.78	1405.82	1327.23
2014-4	5251.25	4915.39	4807.09	2230.11	2059.65
2014-5	6842.66	6208.52	6101.49	2785.56	2506.86
2014-6	6298.43	5503.32	5339.25	1475.94	1114.91
2014-7	6739.50	5734.93	5510.33	1193.97	740.16
2014-8	4630.87	3487.10	3168.90	-1572.30	-2072.81
2014-9	3201.66	1851.99	1454.69	-3713.38	-4319.83
2014-10	5503.15	3909.38	3396.40	-2159.57	-2876.82
2014-11	6633.18	4789.58	4192.89	-1710.60	-2549.19
2014-12	6417.91	4382.22	3636.45	-2553.23	-3510.56
2015-1	8676.51	6306.37	5467.32	-1042.13	-2152.82
2015-2	10737.50	8132.02	7194.13	326.54	-875.65
2015-3	11484.88	8528.51	7476.15	302.28	-1030.98
2015-4	13256.00	9957.27	8778.69	1413.33	-106.07
2015-5	13578.59	10037.39	8732.91	994.59	-538.08
2015-6	12934.59	9000.23	7548.03	-459.13	-2112.99
2015-7	15007.99	10769.27	9176.84	788.36	-964.79
2015-8	18407.57	13815.71	12038.33	3394.66	1540.98
2015-9	21751.93	17398.74	15293.74	5887.14	4341.62
2015-10	25870.80	21011.15	18758.47	9199.42	7473.96
2015-11	29302.42	23914.83	21585.85	11813.77	9910.38
2015-12	31293.90	25735.18	23186.19	13035.16	11150.92
Average time (sec)	1.65	0.23	1.31	2.44	1.64

Table 14: Cumulative losses per month

4 Portfolio Optimization and Risk Allocation with the Generalized Hyperbolic Distribution

4.1 Mean-Risk Optimization for the Normal Mixture Distributions

The mean-risk portfolio optimization problem is a generalization of Markowitz's mean-variance framework by replacing variance by other risk measures. In general, a risk measure is said to be coherent if it satisfies the axioms proposed by [3]:

Definition 3. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be the probability space and $\mathcal{L}(\Omega, \mathcal{F})$ be the set of one dimensional random variables on the space. The coherent risk measure is a function $\rho : \mathcal{L}(\Omega, \mathcal{F}) \to \mathbb{R}$ which satisfies the following axioms for all $X, Y \in \mathcal{L}(\Omega, \mathcal{F})$:*

1. *(Monotonicity) If $X \leq Y$, then $\rho(X) \geq \rho(Y)$;*
2. *(Translation invariance) For all $x \in \mathbb{R}$, $\rho(X + x) = \rho(X) - x$;*
3. *(Positive homogeneity) For all $\lambda \geq 0$, $\rho(\lambda X) = \lambda \rho(X)$;*
4. *(Subadditivity) $\rho(X + Y) \leq \rho(X) + \rho(Y)$.*

One of the classical examples of coherent risk measure is the conditional value at risk or average value at risk:

Definition 4. *Let $X \in \mathcal{L}(\Omega, \mathcal{F})$ whose distribution is continuous and $\alpha \in (0, 1)$, then the value at risk (VaR) is*

$$VaR_\alpha(X) := -\inf\{x \in \mathbb{R} : \mathbb{P}(X \leq x) > \alpha\};$$

and the conditional value at risk (CVaR) is

$$CVaR_\alpha(X) := -E[X | X \leq -VaR_\alpha(X)].$$

VaR is widely used in financial industry as an alternative risk measure to standard deviation. But it is well-know that VaR is not coherent since

it does not have subadditivity. Recall that a normal mixture random vector can be written as:

$$X \overset{d}{=} \mu + \gamma Y + \sqrt{Y} Z, \tag{27}$$

where $\mu, \gamma \in \mathbb{R}^d$, $d \in \mathbb{N}^+$, Y is a univariate non-negative random variable and Z is a random vector which follows multivariate normal distribution with zero mean and covariance $\Sigma \in \mathbb{R}^{d \times d}$. Y and Z are independent. Recall that when Y follows the GIG distribution then the above equation defines the GH distribution. In this section however we will not impose any specific distribution to Y.

Now let us consider the univariate case when the dimension $d = 1$ and $Z \sim N(0, \sigma^2)$ where $\sigma > 0$. The following theorem is simple but it builds an important connection between the Markowitz's portfolio theory to the normal mixture distribution.

Theorem 2. *Let ρ be a coherent risk measure which depends only on the distribution, i.e. if X_1 and X_2 have the same distribution then $\rho(X_1) = \rho(X_2)$, and X follows the normal mixture distribution with parameters μ,γ and σ defined by (27), then*

1. *$\mu \mapsto \rho(X)$ is decreasing in $\mathbb{R}$;*
2. *$\gamma \mapsto \rho(X)$ is non-increasing in $\mathbb{R}$;*
3. *$\sigma \mapsto \rho(X)$ is non-decreasing in $\mathbb{R}^+$.*

Proof. (i) is obvious since $\rho(X) = \rho(\mu + \gamma Y + \sqrt{Y}\sigma Z) = \rho(\gamma Y + \sqrt{Y}\sigma Z) - \mu$. For any $\Delta\gamma \geq 0$ we have

$$\begin{aligned}
&\rho((\gamma + \Delta\gamma)Y + \sqrt{Y}\sigma Z) \\
&\leq \rho(\gamma Y + \sqrt{Y}\sigma Z) + \rho(\Delta\gamma Y) \\
&\leq \rho(\gamma Y + \sqrt{Y}\sigma Z) + \rho(0) = \rho(\gamma Y + \sqrt{Y}\sigma Z),
\end{aligned}$$

which proves (ii).
For (iii) note that $\sigma \mapsto \rho(\gamma Y + \sigma\sqrt{Y} Z)$ is convex on the whole real line:

$$\begin{aligned}
&\rho(\gamma Y + (a\sigma_1 + (1-a)\sigma_2)\sqrt{Y} Z) \\
&= \rho((a + 1 - a)\gamma Y + (a\sigma_1 + (1-a)\sigma_2)\sqrt{Y} Z) \\
&\leq a\rho(\gamma Y + \sigma_1\sqrt{Y} Z) + (1-a)\rho(\gamma Y + \sigma_2\sqrt{Y} Z),
\end{aligned}$$

for any $a \in [0,1]$; and that $\sigma \mapsto \rho(\gamma Y + \sigma\sqrt{Y}Z)$ is symmetric about zero since $\gamma Y + \sigma\sqrt{Y}Z \overset{d}{=} \gamma Y - \sigma\sqrt{Y}Z$. Therefore it must be non-decreasing on $\mathbb{R}^+$. Otherwise suppose that there exits $0 < \sigma_1 < \sigma_2$ such that $\rho(\gamma Y + \sigma_1\sqrt{Y}Z) > \rho(\gamma Y + \sigma_2\sqrt{Y}Z)$. Then find $a := \frac{\sigma_1+\sigma_2}{2\sigma_2} \in (0,1)$ and therefore

$$\begin{aligned}
&\rho(a(\gamma Y + \sigma_2\sqrt{Y}Z) + (1-a)(\gamma Y - \sigma_2\sqrt{Y}Z)) \\
&= \rho(\gamma Y + \sigma_1\sqrt{Y}Z) \\
&> a\rho(\gamma Y + \sigma_2\sqrt{Y}Z) + (1-a)\rho(\gamma Y + \sigma_2\sqrt{Y}Z) \\
&= a\rho(\gamma Y + \sigma_2\sqrt{Y}Z) + (1-a)\rho(\gamma Y - \sigma_2\sqrt{Y}Z),
\end{aligned}$$

which draws the contradiction. □

Intuitively one can think (27) as an portfolio with a risk-free asset whose weight is μ, an asset with non-negative return whose weight is γ and a risky asset whose weight is σ. Any reasonable risk measure or risk-averse utility function should prefer large μ and γ and small σ.

Now consider again the high dimensional case where there are d assets whose returns are modeled by a normal mixture random variable X. A portfolio on these assets is given by a vector $w \in \mathbb{R}^d$ which denotes the weight on each asset and satisfies $w^\mathsf{T}\mathbf{e} = 1$ where $\mathbf{e} = (1,\ldots,1)^\mathsf{T}$. It is easy to check that the portfolio return also follows the normal mixture distribution:

$$w^\mathsf{T}X \overset{d}{=} w^\mathsf{T}\mu + w^\mathsf{T}\gamma Y + \sqrt{w^\mathsf{T}\Sigma w Y}Z, \tag{28}$$

and the expectation of the return is given by:

$$E[w^\mathsf{T}X] = w^\mathsf{T}\mu + w^\mathsf{T}\gamma E[Y].$$

Followed by the classical Markowitz portfolio optimization problem, the generalized mean-risk portfolio optimization problem is formulated as follows:

$$\begin{aligned}
&\min_w \rho(w^\mathsf{T}X) \\
&\text{s.t.} \begin{cases} w^\mathsf{T}\mathbf{e} = 1 \\ E[w^\mathsf{T}X] \geq m \end{cases},
\end{aligned} \tag{29}$$

where $m \in \mathbb{R}$ and ρ is a coherent risk measure. [27] introduce the mean-skewness frontier in a specific optimization problem. Here we can use this concept to reduce the dimension of the problem from d to 2:

Proposition 4. *The solution of the optimization problem (29) is given by*

$$w^* = \Sigma^{-1}[\mu\ \gamma\ \mathbf{e}]A^{-1}[\tilde{\mu}^*\ \tilde{\gamma}^*\ 1]^\mathsf{T},$$

where $[\mu\ \gamma\ \mathbf{e}] \in \mathbb{R}^{d\times 3}$ *are the matrix composed by* μ, γ *and* $\mathbf{e}$*; and* $\tilde{\mu}^*, \tilde{\gamma}^* \in \mathbb{R}$ *are the solutions of:*

$$\begin{aligned} &\min_{\tilde{\mu},\tilde{\gamma}} \rho(\tilde{\mu} + \tilde{\gamma}Y + \sqrt{g(\tilde{\mu},\tilde{\gamma})Y}Z) \\ &s.t.\ \tilde{\mu} + \tilde{\gamma}E[Y] \geq m, \end{aligned}$$

where

$$g(\tilde{\mu},\tilde{\gamma}) = [\tilde{\mu},\tilde{\gamma},1]A^{-1}[\tilde{\mu},\tilde{\gamma},1]^\mathsf{T},$$

and

$$A = \begin{bmatrix} \mu^\mathsf{T}\Sigma^{-1}\mu & \gamma^\mathsf{T}\Sigma^{-1}\mu & \mathbf{e}^\mathsf{T}\Sigma^{-1}\mu \\ \mu^\mathsf{T}\Sigma^{-1}\gamma & \gamma^\mathsf{T}\Sigma^{-1}\gamma & \mathbf{e}^\mathsf{T}\Sigma^{-1}\gamma \\ \gamma^\mathsf{T}\Sigma^{-1}\mathbf{e} & \gamma^\mathsf{T}\Sigma^{-1}\mathbf{e} & \mathbf{e}^\mathsf{T}\Sigma^{-1}\mathbf{e} \end{bmatrix}.$$

Proof. Define the function $w^*(\tilde{\mu},\tilde{\gamma}) : \mathbb{R}\times\mathbb{R} \to \mathbb{R}^d$ as the solution of the optimization problem:

$$\begin{aligned} w^*(\tilde{\mu},\tilde{\gamma}) &:= \arg\min_w \rho(w^\mathsf{T}X) \\ \text{s.t. } &\begin{cases} w^\mathsf{T}\mathbf{e} = 1 \\ w^\mathsf{T}\mu = \tilde{\mu} \\ w^\mathsf{T}\gamma = \tilde{\gamma} \end{cases} \end{aligned} \tag{30}$$

Then (30) is equivalent to

$$\begin{aligned} &\min_{\tilde{\mu},\tilde{\gamma}} \rho(w^*(\tilde{\mu},\tilde{\gamma})^\mathsf{T}X) \\ &\text{s.t. } \tilde{\mu} + \tilde{\gamma}E[T] \geq m. \end{aligned}$$

Theorem 1 and equation (28) implies that if $w^\mathsf{T}\gamma$ and $w^\mathsf{T}\gamma$ are fixed then $\rho(w^\mathsf{T}X)$ is non-decreasing with respect to $w^\mathsf{T}\Sigma w$. Therefore (30) is equivalent to

$$w^*(\tilde{\mu},\tilde{\gamma}) = \arg\min_w (w^\mathsf{T}\Sigma w)$$

with the same set of constraints. This problem can be easily solved by Lagrange multiplier. One can shown that the solution is given by

$$w^*(\tilde{\mu},\tilde{\gamma}) = \Sigma^{-1}[\mu\ \gamma\ \mathbf{e}]A^{-1}[\tilde{\mu}\ \tilde{\gamma}\ 1]^\mathsf{T},$$

Applying the solution to (28), we prove the theorem. □

Note that we are able to extend the efficient frontier to three dimensional space using equation (30). In this case, the three dimensions are given by the location parameter, the skewness parameter and the risk respectively. Here we call the 3 dimensional "efficient frontier" as the efficient surface in order to distinguish it with the traditional definition of the efficient frontier. Figure 8 plots the efficient surface of CVaR of a GH distribution. Here we assume that x-axis is for the location parameter, y-axis is for the skewness parameter and z-axis is for the risk. Each point on the surface is the solution of the optimization problem (30) given $\tilde{\mu}$ and $\tilde{\gamma}$. From figure 8 one can observe that surface is convex with respect to the location parameter $\tilde{\mu}$ and the skewness parameter $\tilde{\gamma}$.

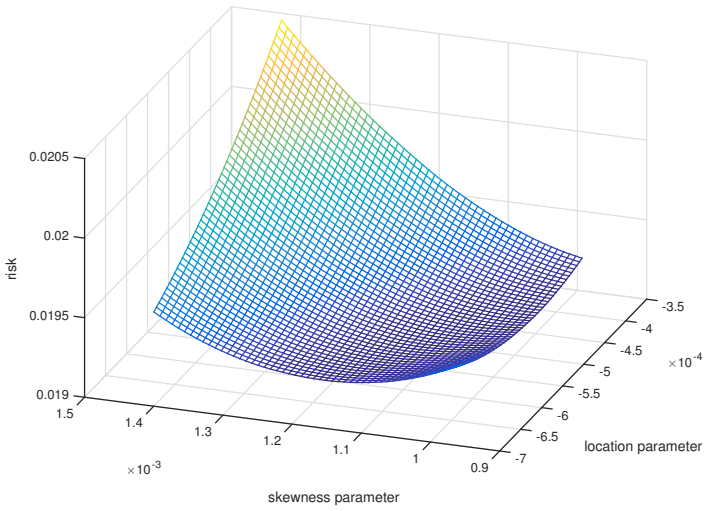
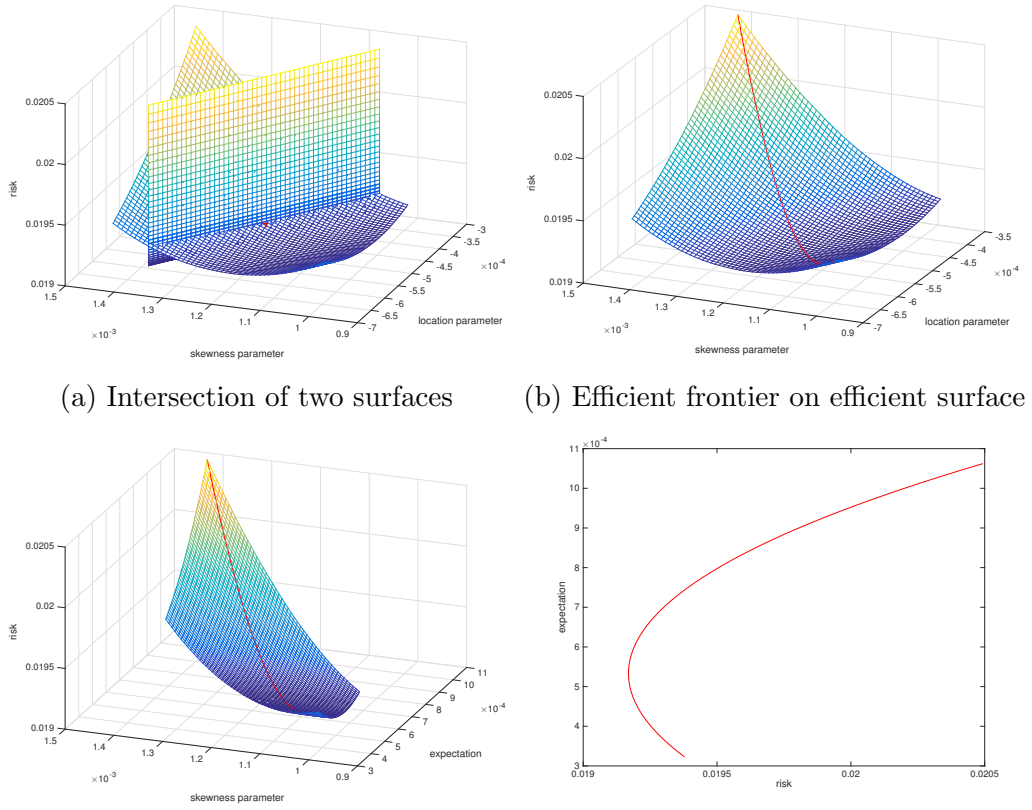


Figure 8: Efficient Surface

To visualize the proof of the proposition, observe that once the objective expectation m is fixed, the space of $\tilde{\mu}$ and $\tilde{\gamma}$ is the straight line $\tilde{\mu}+\tilde{\gamma}E[Y] = m$. In the 3 dimensional space, it is a plane orthogonal to the x-y plane, as shown by figure 9a. The intersection of the plane and the efficient surface is a convex curve from which we find the minimal of point of the risk. The minimal risk together with the target expectation m forms the efficient frontier, which is the red curve in figure 9b. To make it more clear, we are able to transform the coordinate system by replacing $\tilde{\mu}$ by $m = \tilde{\mu} + \tilde{\gamma}E[T]$. In this case we have figure 9c. By projecting the red curve onto the y-z plane we obtain the

efficient frontier in figure 9d.



(a) Intersection of two surfaces

(b) Efficient frontier on efficient surface

(c) Change of the coordinate system

(d) Efficient frontier

Figure 9: Geometry of the proposition

Here we given another example of the application of theorem 2 to finance. [17] shows that the worst-case conditional value at risk (WCVaR) of generalized hyperbolic distribution under the box uncertainty can be solved explicitly. By theorem 1 we can extend this result to the general case for the coherent risk measure and the normal mixture distribution.

Definition 5. *Let X be a random vector whose density is given by f and w be the portfolio weights, the worst-case coherent risk measure is defined as a function of w:*

$$\rho^*(w) := \sup_{f \in \mathcal{P}} \{\rho(w^\mathsf{T} X)\},$$

where $\mathcal{P}$ is some sets of probability density function.

There are several specification of the set $\mathcal{P}$, see [48]. Here we only consider the box uncertainty set for normal mixture models defined by

$$\mathcal{P} = \{p_X(\cdot|\mu, \gamma, \Sigma) \text{ such that } \underline{\mu} \preceq \mu \preceq \overline{\mu}, \underline{\gamma} \preceq \gamma \preceq \overline{\gamma}, \underline{\Sigma} \preceq \Sigma \preceq \overline{\Sigma}, \ p_Y \text{ is fixed.}\},$$

where p_X and p_Y are the marginal densities of X and Y defined by (27):

$$p_X(x|\mu, \gamma, \Sigma) = \int_0^\infty \frac{1}{\sqrt{(2\pi)^d \det(\Sigma) y}} \exp\left(-\frac{1}{2y}(x - \mu - \gamma y)^\mathsf{T} \Sigma^{-1} (x - \mu - \gamma y)\right) p_Y(y) dy,$$

and p_Y is defined on the interval $(0, \infty)$. By $\preceq$ we mean the each elements of a vector are smaller than the elements of another one. Then it is straight forward to get the following results based on theorem 2:

Proposition 5. *For any $w \in \mathbb{R}^d_+$,*

$$\rho^*(w) = \rho(w^\mathsf{T}\underline{\mu} + w^\mathsf{T}\underline{\gamma} + \sqrt{w^\mathsf{T}\overline{\Sigma} w} Y Z).$$

4.2 CVaR Derivatives of the Normal Mixture Distributions

In this section we compute the derivatives of CVaR for the normal mixture distributions. For simplicity let us define $r_{VaR_\alpha}(w) := VaR_\alpha(w^\mathsf{T}X)$ and $r_{CVaR_\alpha}(w) := CVaR_\alpha(w^\mathsf{T}X)$ as the VaR and CVaR of a portfolio where $X \in \mathbb{R}^d$ is the vector of returns and $w \in \mathbb{R}^d$ is the portfolio weights. First we review the general results of the first and second derivatives of r_{CVaR_α} given by [39] and [43].

Assumption 1. *Let $p(x_1|x_2, \ldots, x_n)$ be the conditional density function of X_1 given $X_2, \ldots, X_n$. p satisfies:*

1. *$y \mapsto p(y|x_2, \ldots, x_n)$ is continuous for fixed $x_2, \ldots, x_n$.*
2. *The mapping*

$$(y, w) \mapsto E[p(w_1^{-1}(y - \sum_{l=2}^n w_l X_l)|X_2, \ldots, X_n)]$$

is finite valued and continuous on $\mathbb{R} \times \mathbb{R}/0 \times \mathbb{R}^{n-1}$.

3. For each $w \in \mathbb{R}/0 \times \mathbb{R}^{n-1}$

$$E[p(w_1^{-1}(-r_{VaR_\alpha}(w) - \sum_{l=2}^{n} w_l X_l)|X_2, \ldots, X_n)] > 0.$$

4. The mapping

$$(y, w) \mapsto E[X_j p(w_1^{-1}(y - \sum_{l=2}^{n} w_l X_l)|X_2, \ldots, X_n)]$$

is finite valued and continuous on $\mathbb{R} \times \mathbb{R}/0 \times \mathbb{R}^{n-1}$ *for each* $j = 2, \ldots, n$.

5. The mapping

$$(y, w) \mapsto E[X_j X_k p(w_1^{-1}(y - \sum_{l=2}^{n} w_l X_l)|X_2, \ldots, X_n)]$$

is finite valued and continuous on $\mathbb{R} \times \mathbb{R}/0 \times \mathbb{R}^{n-1}$ *for each* $j, k = 2, \ldots, n$.

Proposition 6. *If* $(X_1, \ldots, X_n)$ *satisfies Assumption 1 1-4 and* $w \in \mathbb{R}/0 \times \mathbb{R}^{n-1}$, *then* $r_{VaR_\alpha}(w)$ *and* $r_{CVaR_\alpha}(w)$ *are partially differentiable with continuous derivatives:*

$$\frac{\partial r_{VaR_\alpha}}{\partial w_j}(w) = -\frac{E\Big[X_j p\big(w_1^{-1}(-r_{VaR_\alpha}(w) - \sum_{j=2}^{n} w_j X_j)|X_2, \ldots, X_d\big)\Big]}{E\Big[p\big(w_1^{-1}(-r_{VaR_\alpha}(w) - \sum_{j=2}^{n} w_j X_j)|X_2, \ldots, X_d\big)\Big]},$$

$$\frac{\partial r_{VaR_\alpha}}{\partial w_1}(w) = w_1^{-1}\left(r_{VaR_\alpha}(w) - \sum_{j=2}^{n} w_j \frac{\partial r_{VaR_\alpha}}{\partial w_j}(w)\right),$$

and

$$\begin{aligned}\frac{\partial r_{CVaR_\alpha}}{\partial w_j}(w) &= -E[X_j|w^\mathsf{T} X \leq -r_{VaR_\alpha}(w)] \\ &= -\alpha^{-1} E[X_j I_{\{w^\mathsf{T} X \leq -r_{VaR_\alpha}(w)\}}],\end{aligned}$$

for $j = 1, \ldots, n$.

The second derivatives are just the results of direct computation under Assumption 1.

Proposition 7. *If $(X_1, \ldots, X_n)$ satisfies Assumption 1 and $w \in \mathbb{R}/0 \times \mathbb{R}^{n-1}$, $r_{CVaR_\alpha}(w)$ are second order differentiable:*

$$\frac{\partial^2 r_{CVaR_\alpha}}{\partial w_j \partial w_k}(w) = (\alpha|w_1|)^{-1} E\bigg[X_k\Big(\frac{\partial r_{VaR_\alpha}}{\partial w_j}(w) + X_j\Big) \\ p\Big(w_1^{-1}\Big(-r_{VaR_\alpha}(w) - \sum_{l=2}^{n} w_l X_l\Big)\Big| X_2, \ldots, X_n\Big)\bigg],$$

where $j, k = 2, \ldots, n$.

$$\frac{\partial^2 r_{CVaR_\alpha}}{\partial w_j \partial w_1}(w) = -w_1^{-1} \sum_{k=2}^{n} w_k \frac{\partial^2 r_{CVaR_\alpha}}{\partial w_j \partial w_k}(w),$$

where $j = 1, \ldots, n$.

Proof. Let $j, k = 2, \ldots, n$ and $w_1 > 0$,

$$\frac{\partial^2 r_{CVaR_\alpha}}{\partial w_j \partial w_k}(w) = -\alpha^{-1} \frac{\partial}{\partial w_j} E[X_k I_{\{w^\intercal X \le -r_{VaR_\alpha}(w)\}}]$$
$$= -\alpha^{-1} \frac{\partial}{\partial w_j} E\big[X_k E[I_{\{w^\intercal X \le -r_{VaR_\alpha}(w)\}} | X_2, \ldots, X_n]\big]$$
$$= -\alpha^{-1} E\bigg[X_k \frac{\partial}{\partial w_j} P\bigg(X_1 \le w_1^{-1}\Big(-r_{VaR_\alpha}(w) - \sum_{l=2}^{n} w_l X_l\Big)\Big| X_2, \ldots, X_n\bigg)\bigg]$$
$$= (\alpha w_1)^{-1} E\bigg[X_k\Big(\frac{\partial r_{VaR_\alpha}(w)}{\partial w_j}(w) + X_j\Big)$$
$$p\Big(w_1^{-1}\Big(-r_{VaR_\alpha}(w) - \sum_{l=2}^{n} w_l X_l\Big)\Big| X_2, \ldots, X_n\Big)\bigg].$$

If $w_1 < 0$,

$$\frac{\partial^2 r_{CVaR_\alpha}}{\partial w_j \partial w_k}(w)$$
$$= -\alpha^{-1} E\bigg[X_k \frac{\partial}{\partial w_j} P\bigg(X_1 \ge w_1^{-1}\Big(-r_{VaR_\alpha}(w) - \sum_{l=2}^{n} w_l X_l\Big)\Big| X_2, \ldots, X_n\bigg)\bigg]$$
$$= -(\alpha w_1)^{-1} E\bigg[X_k\Big(\frac{\partial r_{VaR_\alpha}(w)}{\partial w_j}(w) + X_j\Big)$$
$$p\Big(w_1^{-1}\Big(-r_{VaR_\alpha}(w) - \sum_{l=2}^{n} w_l X_l\Big)\Big| X_2, \ldots, X_n\Big)\bigg].$$

Note that Assumption 1 ensures the continuity and differentiability of the conditional distribution function of X_1 given $X_2, \ldots, X_n$, and the finiteness of the results. Thus we are able to interchange the expectation and the differentiation. The second equation is simply the result of 1-homogeneity of r_{CVaR_α}:

$$\frac{\partial r_{CVaR_\alpha}}{\partial w_j}(w) = \frac{\partial}{\partial w_j}\sum_{k=1}^{n} w_k \frac{\partial r_{CVaR_\alpha}}{\partial w_k}(w) = \frac{\partial r_{CVaR_\alpha}}{\partial w_j}(w) + \sum_{k=1}^{n} w_k \frac{\partial^2 r_{CVaR_\alpha}}{\partial w_j \partial w_k}(w).$$

□

Recall that the representation of a univariate normal mixture distribution (27) can be viewed as a "portfolio" with two risky assets; and that the portfolio weight in propositions 6 and 7 is not necessarily normalized. Thus we can apply the propositions directly to (27) and get the CVaR derivatives of normal mixture distributions with respect to three parameters $\mu, \gamma, \sigma \in \mathbb{R}, \sigma > 0$. Let us define:

$$\begin{aligned} r_{VaR_\alpha}(\mu, \gamma, \sigma) &:= VaR_\alpha(\mu + \gamma Y + \sigma\sqrt{Y}Z), \\ r_{CVaR_\alpha}(\mu, \gamma, \sigma) &:= CVaR_\alpha(\mu + \gamma Y + \sigma\sqrt{Y}Z), \end{aligned}$$

to be VaR and CVaR of univariate normal mixture distributions for simplicity. Then by applying proposition 6 we have

$$\begin{aligned} &\frac{\partial r_{VaR_\alpha}}{\partial \mu}(\mu, \gamma, \sigma) = -1, \\ &\frac{\partial r_{VaR_\alpha}}{\partial \gamma}(\mu, \gamma, \sigma) = -\frac{E\big[\sqrt{Y} p_{norm}\big(\frac{-r_{VaR_\alpha}(\mu,\gamma,\sigma)-\mu-\gamma Y}{\sigma\sqrt{Y}}\big)\big]}{E\big[p_{norm}\big(\frac{-r_{VaR_\alpha}(\mu,\gamma,\sigma)-\mu-\gamma Y}{\sigma\sqrt{Y}}\big)/\sqrt{Y}\big]}, \\ &\frac{\partial r_{VaR_\alpha}}{\partial \sigma}(\mu, \gamma, \sigma) = \sigma^{-1}\left(VaR_\alpha(X) + \mu - \gamma\frac{\partial}{\partial \gamma} r_{VaR_\alpha}(\mu, \gamma, \sigma)\right), \end{aligned}$$

where p_{norm} denotes the density function of standard normal distribution:

$$p_{norm}(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2);$$

and

$$\begin{aligned}
&\frac{\partial r_{CVaR_\alpha}}{\partial \mu}(\mu,\gamma,\sigma) = -1,\\
&\frac{\partial r_{CVaR_\alpha}}{\partial \gamma}(\mu,\gamma,\sigma) = -\alpha^{-1}E\left[YF_{norm}\left(\frac{-r_{VaR_\alpha}(\mu,\gamma,\sigma)-\mu-\gamma Y}{\sigma\sqrt{Y}}\right)\right],\\
&\frac{\partial r_{CVaR_\alpha}}{\partial \sigma}(\mu,\gamma,\sigma) = \sigma^{-1}\left(r_{CVaR_\alpha}(\mu,\gamma,\sigma)+\mu-\gamma\frac{\partial r_{CVaR_\alpha}}{\partial \gamma}(\mu,\gamma,\sigma)\right),
\end{aligned}$$

where F_{norm} denotes the cumulative distribution function of the standard normal distribution. Here we assume that the distribution of Y satisfies Assumption 1. Then the second derivatives are just the results of direct computations:

$$\begin{aligned}
\frac{\partial^2 r_{CVaR_\alpha}}{\partial \mu^2}(\mu,\gamma,\sigma) =& \frac{\partial^2 r_{CVaR_\alpha}}{\partial \mu\partial \gamma}(\mu,\gamma,\sigma) = \frac{\partial^2 r_{CVaR_\alpha}}{\partial \mu\partial \sigma}(\mu,\gamma,\sigma) = 0,\\
\frac{\partial^2 r_{CVaR_\alpha}}{\partial \gamma^2}(\mu,\gamma,\sigma) =& \frac{1}{\alpha\sigma}E\left[\sqrt{Y}p_{norm}\left(\frac{-r_{VaR_\alpha}(\mu,\gamma,\sigma)-\mu-\gamma Y}{\sigma\sqrt{Y}}\right)\right.\\
&\left.\left(\frac{\partial r_{VaR_\alpha}}{\partial \gamma}(\mu,\gamma,\sigma)+Y\right)\right],\\
\frac{\partial^2 r_{CVaR_\alpha}}{\partial \gamma\sigma}(\mu,\gamma,\sigma) =& -\frac{\gamma}{\sigma}\frac{\partial^2 r_{CVaR_\alpha}}{\partial \gamma^2}(\mu,\gamma,\sigma),\\
\frac{\partial^2 r_{CVaR_\alpha}}{\partial \sigma^2}(\mu,\gamma,\sigma) =& -\frac{\gamma}{\sigma}\frac{\partial^2 r_{CVaR_\alpha}}{\partial \gamma\sigma}(\mu,\gamma,\sigma).
\end{aligned}$$

The above derivatives can all be computed efficiently via Monte Carlo by generating i.i.d samples of the subordinator Y.

Now let us consider the derivatives of $r_{CVaR_\alpha}(w)$ for the normal mixture distributions. We are able to write $r_{CVaR_\alpha}(w) = r_{CVaR_\alpha}(w^\mathsf{T}\mu, w^\mathsf{T}\gamma, \sqrt{w^\mathsf{T}\Sigma w})$ without loss of generality. Thus the derivatives of $r_{CVaR_\alpha}(w)$ can be repre-

sented by the derivatives of $r_{CVaR_\alpha}(\mu, \gamma, \sigma)$:

$$\begin{aligned}
\frac{\partial r_{CVaR_\alpha}}{\partial w_j}(w) &= -\mu_j + \gamma_j \frac{\partial r_{CVaR_\alpha}}{\partial \gamma}(w^\mathsf{T}\mu, w^\mathsf{T}\gamma, \sqrt{w^\mathsf{T}\Sigma w}) \\
&\quad + \frac{(\Sigma w)_j}{\sqrt{w^\mathsf{T}\Sigma w}} \frac{\partial r_{CVaR_\alpha}}{\partial \sigma}(w^\mathsf{T}\mu, w^\mathsf{T}\gamma, \sqrt{w^\mathsf{T}\Sigma w}), \\
\frac{\partial^2 r_{CVaR_\alpha}}{\partial w_j \partial w_k}(w) &= \gamma_j \gamma_k \frac{\partial^2 r_{CVaR_\alpha}}{\partial \gamma^2}(w^\mathsf{T}\mu, w^\mathsf{T}\gamma, \sqrt{w^\mathsf{T}\Sigma w}) \\
&\quad + \left(\gamma_j \frac{(\Sigma w)_k}{\sqrt{w^\mathsf{T}\Sigma w}} + \gamma_k \frac{(\Sigma w)_j}{\sqrt{w^\mathsf{T}\Sigma w}} \right) \frac{\partial^2 r_{CVaR_\alpha}}{\partial \gamma \partial \sigma}(w^\mathsf{T}\mu, w^\mathsf{T}\gamma, \sqrt{w^\mathsf{T}\Sigma w}) \\
&\quad + \frac{(\Sigma w)_j (\Sigma w)_k}{w^\mathsf{T}\Sigma w} \frac{\partial^2 r_{CVaR_\alpha}}{\partial \sigma^2}(w^\mathsf{T}\mu, w^\mathsf{T}\gamma, \sqrt{w^\mathsf{T}\Sigma w}) \\
&\quad + \frac{1}{\sqrt{w^\mathsf{T}\Sigma w}} \left(\sigma_{jk} - \frac{(\Sigma w)_j (\Sigma w)_k}{w^\mathsf{T}\Sigma w} \right) \frac{\partial r_{CVaR_\alpha}}{\partial \sigma}(w^\mathsf{T}\mu, w^\mathsf{T}\gamma, \sqrt{w^\mathsf{T}\Sigma w}),
\end{aligned}$$

where $(\cdot)_j$ denotes the j-th element of an vector. The matrix representation of the above equation is given by:

$$\begin{aligned}
H_{r_{CVaR_\alpha}}(w) &= \gamma\gamma^\mathsf{T} \frac{\partial^2 r_{CVaR_\alpha}}{\partial \gamma^2}(w^\mathsf{T}\mu, w^\mathsf{T}\gamma, \sqrt{w^\mathsf{T}\Sigma w}) \\
&\quad + (w^\mathsf{T}\Sigma w)^{-\frac{1}{2}} (\gamma w^\mathsf{T}\Sigma + \Sigma w \gamma^\mathsf{T}) \frac{\partial^2 r_{CVaR_\alpha}}{\partial \gamma \sigma}(w^\mathsf{T}\mu, w^\mathsf{T}\gamma, \sqrt{w^\mathsf{T}\Sigma w}) \\
&\quad + (w^\mathsf{T}\Sigma w)^{-1} \Sigma w w^\mathsf{T}\Sigma \frac{\partial^2 r_{CVaR_\alpha}}{\partial \sigma^2}(w^\mathsf{T}\mu, w^\mathsf{T}\gamma, \sqrt{w^\mathsf{T}\Sigma w}) \\
&\quad + (w^\mathsf{T}\Sigma w)^{-\frac{3}{2}} (\Sigma w^\mathsf{T}\Sigma w - \Sigma w w^\mathsf{T}\Sigma) \frac{\partial r_{CVaR_\alpha}}{\partial \sigma}(w^\mathsf{T}\mu, w^\mathsf{T}\gamma, \sqrt{w^\mathsf{T}\Sigma w}),
\end{aligned}$$

where $H_{r_{CVaR_\alpha}}$ denote the Hessian matrix of the function r_{CVaR_α}.

4.3 Portfolio Optimization with Transaction Costs

In this section we consider the following d dimensional portfolio optimization problem:

$$\begin{aligned}
&\max_{w \in \mathbb{R}^d} w^\mathsf{T} m - c_1 r(w) - c_2 \|w - w_0\|_1 \\
&\text{s.t.} \begin{cases} w^\mathsf{T}\mathbf{e} = 1 \\ Aw \leq b \end{cases},
\end{aligned} \tag{31}$$

where $w \in \mathbb{R}^d$ is the portfolio weight to be optimized, $m \in \mathbb{R}^d$ is the expected return, $r : \mathbb{R}^d \to \mathbb{R}$ is some convex non-negative risk function, $w_0 \in \mathbb{R}^d$ is the current portfolio weight that satisfies the constraints, and c_1, c_2 are some positive constants. Thus the first term in the objective function is the expected return of the portfolio, the second term measures the risk, and the third term $c_2\|w - w_0\|_1$ measures the portfolio turnovers or the transaction costs.

Because of the existence of transaction costs we are not able to use the same approach as we discussed in the section 3.1 when $r(w) := \rho(w^\mathsf{T} X)$ for some coherent risk measure ρ and the asset returns are modeled by a normal mixture random vector X. On the other side, the solution of the optimization problem cannot be very different with w_0 if there is a relatively strong constraint on the transaction costs in practise. Thus we are able to approximate the convex function $r(w)$ by its Taylor expansion:

$$r(w) \approx r(w_0) + (w - w_0)^\mathsf{T} \nabla r(w_0) + \frac{1}{2}(w - w_0)^\mathsf{T} H_r(w_0)(w - w_0).$$

Thus the mean-risk portfolio optimization problem can be approximated by

$$\begin{aligned} &\max_{w \in \mathbb{R}^d} w^\mathsf{T}(m - c_1 \nabla r(w_0)) - \frac{c_1}{2}(w - w_0)^\mathsf{T} H_r(w_0)(w - w_0) - c_2\|w - w_0\|_1 \\ &\text{s.t.} \begin{cases} w^\mathsf{T}\mathbf{e} = 1 \\ Aw \leq b \end{cases}, \end{aligned} \tag{32}$$

where ∇r is the gradient and H_r is the Hessian matrix of r. This problem is still a general convex optimization problem that is not smooth at w_0. But it can be transferred to a quadratic programming problem:

$$\begin{aligned} &\max_{w \in \mathbb{R}^d} v^\mathsf{T}\tilde{m} - \frac{1}{2}c_1 v^\mathsf{T}\tilde{H}v \\ &\text{s.t.} \begin{cases} v^\mathsf{T}\tilde{\mathbf{e}} = 0 \\ \tilde{A}v \leq \tilde{b} \\ v \geq 0 \end{cases}, \end{aligned} \tag{33}$$

where

$$\tilde{m} = \begin{pmatrix} m - c_1 \nabla r(w_0) - c_2 \mathbf{e} \\ -m + c_1 \nabla r(w_0) - c_2 \mathbf{e} \end{pmatrix},$$
$$\tilde{H} = \begin{pmatrix} H_r(w_0) & -H_r(w_0) \\ -H_r(w_0) & H_r(w_0) \end{pmatrix},$$
$$\tilde{\mathbf{e}} = \begin{pmatrix} \mathbf{e} \\ -\mathbf{e} \end{pmatrix},$$
$$\tilde{A} = (A \ -A),$$
$$\tilde{b} = b - Aw_0.$$

The dimension of the above problem is $2d$. If v^* is the solution of problem (33), then the optimal solution of problem (32) is given by:

$$w^* = w_0 + (I \ -I)v^*,$$

where I is the d dimensional identity matrix. Note that $\tilde{H}$ is not of full rank, thus the quadratic programming problem (33) is not strictly convex. For some algorithms one may shrink the zero eigenvalues of the semi-positive definite matrix $\tilde{H}$ a bit to make the problem strictly convex. In that case we can only get a suboptimal solution w^*. In practice we can check compare objective function's value at w^* with the one at w_0 in each step. If it is not greater than the later then we may just hold the current position.

To test the approximated optimal portfolio, we use the daily returns of Dow Jones 30 companies from 2010 to 2014. We fit 2010 to 2013's returns by GH distribution and get a set of initial parameters. Than we apply the on-line EM algorithm to update GH parameters each day from Jan 2014 to Jan 2015. The initial portfolio weights are set to be equally weighted. And we consider long only strategies, i.e the inequality constraints are set such that $w \geq 0$. Parameter c_1 and c_2 are set to be 0.1 and 0.01 respectively. First we solve the general mean-risk convex optimization problem (31), where ρ is given by CVaR, using Matlab function `fmincon` with the interior-point algorithm. The CVaR is computed by Monte Carlo with 10^5 scenarios. Since the original problem is not smooth, `fmincon` may not always return the optimal solution. So at each step we check whether the objective function value is improved by rebalancing the portfolio. If not then we will keep the position unchanged. Then we compute the gradient and Hessian of CVaR at current position, using the formulas derived in the last section. The quadratic

programming problem (33) is solved by Matlab function `quadprog` with the interior-point algorithm. In order to compare the two portfolios, we compute not only the true object function value given by (31), but also their expected turns, CVaR and turnovers. The benchmark we used is the equally weighted portfolio. Note that the equally weighted portfolio is not free of transaction costs; one need to constantly rebalance it in order to keep the weights stay the same. But here we just ignore their turnovers for connivence. If our approximated optimal portfolio is correct, then its expected returns, CVaR and objective function values should be similar as the true optimal one, and be superior than the equally weighted one. The comparison among three portfolios are shown by figure 10.

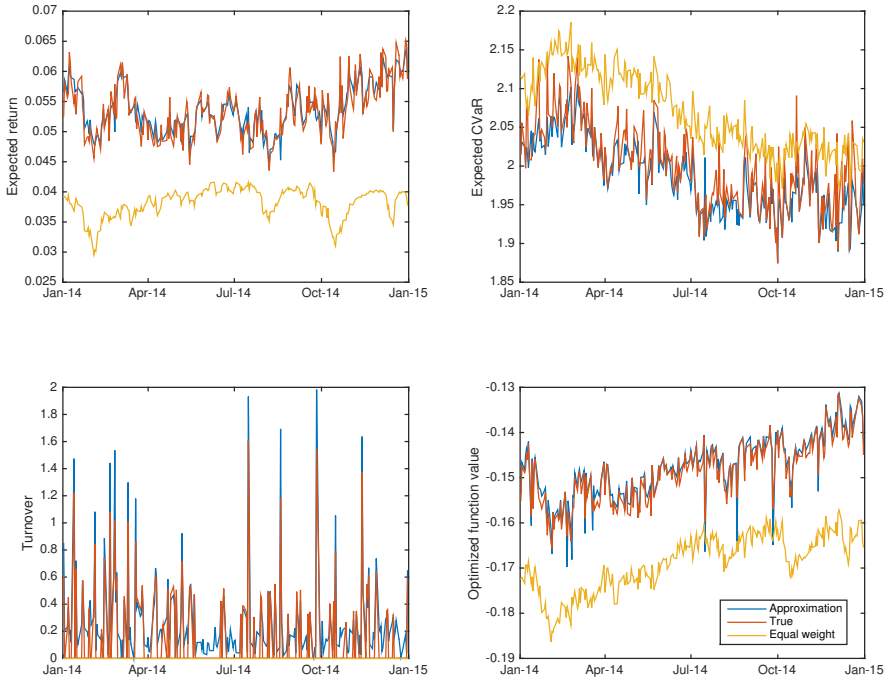


Figure 10: Optimization Results

The four subfigures are the expected return, CVaR, turnover and value of (31) over 252 days in 2014. The red, blue and yellow lines represent the original mean-CVaR portfolio, the Taylor approximation portfolio and

the benchmark equally weighted portfolio. All four statistics of first two portfolios are very close comparing to the ones of the benchmark. One may expect that the Taylor approximation portfolio would stick to the true one for a while and diverges ultimately. But this is not true in our experiment. The portfolio weights of the 30 assets at the end of year is plotted in figure 11.

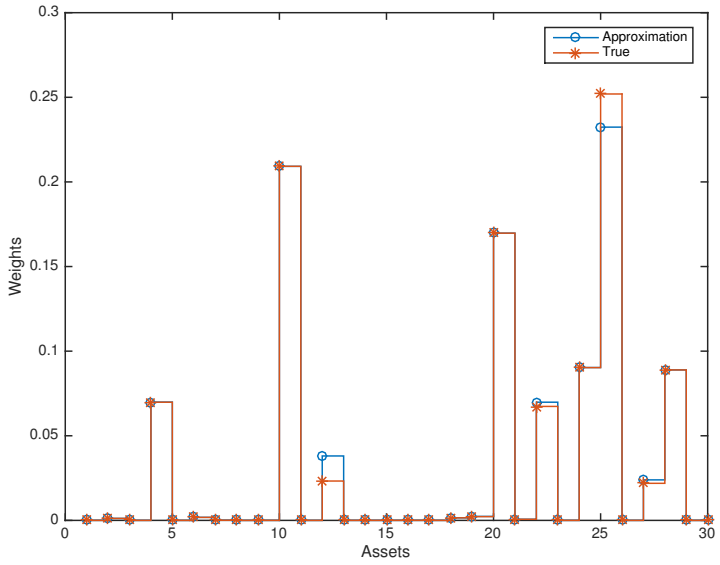


Figure 11: Terminal Weights

We can observe that the approximation is doing well even after 252 periods. There are a few exceptions when the mean-CVaR portfolio changes a lot because of strong signals. One can observe few downside blue spikes in the last subfigure. In these cases the Taylor approximation would be less accurate than the usual case. Despite of these differences, the Taylor approximation with quadratic programming is about 500 times faster than the general convex optimization problem on a laptop with 1.8 GHz Intel Core i7 processor and 4 GB 1333 MHz DDR3 memory. Finally, the realized returns of the mean-CVaR portfolio and its approximation does not outperform the equally weighted benchmark. The first reason is that our model just assumes that the log returns are i.i.d GH distributed without any cross-sectional dependency. A good time series model for alpha prediction is necessary for

constructing a portfolio that outperform the market. But this topic is out of scope of this paper. Secondly we did not add any additional constraints to make portfolio well diversified. From figure 11 we can see that the mean CVaR portfolio only concentrates on 9 out of 30 stocks. We will discuss how to measure portfolio diversification in the following sections.

4.4 Effective Number of Bets and Minimum Torsion

In this section we briefly review the effective number of bets (ENB) together with the minimum torsion approach for measuring portfolio diversification proposed by [30] and [31].

Similar as before, let $X \in \mathbb{R}^n$ be a random vector which denotes the return of $n > 0$ assets. Here we do not impose any distribution assumptions on X. The portfolio weights are given by the vector $w \in \mathbb{R}^n$ with $w^\mathsf{T}\mathbf{e} = 1$. Then the portfolio return is given by $w^\mathsf{T}X$. Let Σ denotes the covariance matrix of X, the portfolio variance is denoted by $r_{Var}(w) := w^\mathsf{T}\Sigma w$.

The gradient of the portfolio variance is given by $\nabla r_{Var}(w) = 2\Sigma w$. In [43] the marginal contribution of the k-th asset to the variance is defined as $w_k(\Sigma w)_k$, where $(\cdot)_k$ denotes the k-th element of the vector. So the risk contributions are clearly not independent. To solve this problem, notice that there exists a square matrix $T \in \mathbb{R}^{n\times n}$ such that $T\Sigma T^\mathsf{T} = D$ is diagonal. For example, T can be the unitary matrix U^T in the SVD of the covariance: $\Sigma = USU^\mathsf{T}$. And it is obvious that T is not unique, for example, you can also set T to be the inverse of the Cholesky factor of Σ.

Let $Y = TX$ be the linear transformation of all asset returns, and $v = (T^\mathsf{T})^{-1}w$. It is clear that the portfolio return remain unchanged, i.e. $w^\mathsf{T}X = v^\mathsf{T}Y$ and $w^\mathsf{T}\Sigma w = v^\mathsf{T}Dv$. The random vector Y can be viewed as uncorrelated risky factors of the portfolio. Now the first derivative w.r.t v is $Dv = (d_1v_1, \ldots, d_nv_n)^\mathsf{T}$, where $d_1, \ldots, d_n$ are the diagonal elements of D. Then the risk contributions are given by $d_1v_1^2, \ldots, d_nv_n^2$, and the sum of which is equal to the portfolio variance. We can normalize these numbers by dividing them by $w^\mathsf{T}\Sigma w$.

Let p_k, $k = 1, \ldots, n$ be the normalized risk contributions:

$$p_k = \frac{d_kv_k^2}{w^\mathsf{T}\Sigma w},\ k = 1, \ldots, n. \tag{34}$$

Note that $\{p_k\}$ forms a discrete probability distributions since for each k $0 \le p_k \le 1$ and $\sum_{k=1}^n p_k = 1$. [30] defines the ENB as the exponential

entropy of the distribution $\{p_k\}$:

$$N = \exp\Big(-\sum_{k=1}^{n} p_k \log p_k \Big).$$

In fact $-\log N = -\sum_{k=1}^{n} p_k \log p_k$ is proportional to the Kullback-Leibler divergence between $\{p_k\}$ and the uniform distribution $q_k = 1/n$, $k = 1, \ldots, n$. Therefore N reaches to its maximum which is exactly n if $p_k = 1/n$ for each k. On the other hand, when one element of $\{p_k\}$ is 1 and the rest are 0, the ENB is equal to 1. Intuitively, N measures the "real" numbers of uncorrelated risky factors in the portfolio.

Now the problem is how to construct a transformation T which diagonalize the covariance Σ. Note that diagnoalizing Σ is equivalent to diagnoalizing the correlation matrix C: $\Sigma = diag(\Sigma)^{\frac{1}{2}} C diag(\Sigma)^{\frac{1}{2}}$ where $diag(\cdot)$ denotes the diagonal part of a square matrix and the square root of a diagonal matrix means the square root of its diagonal elements. Let us consider the SVD of C: $C = USU^{\mathsf{T}}$ where S is diagonal and U is unitary. Using the uniqueness of SVD it is easy to show that:

Proposition 8. *Let Σ be a positive definite covariance matrix and T is a square invertible matrix and D is diagonal. Then $T\Sigma T^{\mathsf{T}} = D$ if only if there is a unitary matrix V such that*

$$T = D^{\frac{1}{2}} V S^{-\frac{1}{2}} U^{\mathsf{T}} diag(\Sigma)^{-\frac{1}{2}}, \tag{35}$$

where S and U are the SVD of the correlation matrix of Σ: $C = USU^{\mathsf{T}}$.

Proof. The "if" part is obvious; the "only if" part is the result of:

$$USU^{\mathsf{T}} = \big(diag(\Sigma)^{-\frac{1}{2}} T^{-1} D^{\frac{1}{2}}\big)\big(diag(\Sigma)^{-\frac{1}{2}} T^{-1} D^{\frac{1}{2}}\big)^{\mathsf{T}},$$

and the fact that the SVD of $diag(\Sigma)^{-\frac{1}{2}} T^{-1} D^{\frac{1}{2}}$ must have the form $US^{\frac{1}{2}} V^{\mathsf{T}}$. □

Lemma 1. *Let N be the ENB given covariance Σ and transformation T which has the expression (35), then N is independent with the choice of D.*

Proof. Let $u = VS^{-\frac{1}{2}} U^{\mathsf{T}} diag(\Sigma)^{\frac{1}{2}} w$ and $v = D^{-\frac{1}{2}} u$. It is clear that p_k given by (34) is independent with D since $d_k v_k^2 = u_k^2$. □

This implies that it is enough to consider $T = VS^{\frac{1}{2}}U^{\mathsf{T}}diag(\Sigma)^{-\frac{1}{2}}$ and pick $v = (T^{\mathsf{T}})^{-1}w$ as the adjusted portfolio weights. However the unitary matrix V is just a rotation of the vector $S^{-\frac{1}{2}}U^{\mathsf{T}}diag(\Sigma)^{\frac{1}{2}}w$. Since there is no restriction on V we can rotate the vector to any direction we want. We can always find a V such that $v_1 = \cdots = v_n = \|v\|/\sqrt{n}$ where $\|\cdot\|$ denotes the vector norm. In this case $N = n$ implies that the portfolio is completely diversified. We can also find another V such that $v_1 = \|v\|$ and $v_k = 0$, $k = 2, \ldots, n$ (using Householder reflection for example). In this case $N = 1$ implies that the portfolio is not diverse at all. In fact we can get any $1 \leq N \leq n$ we want by choosing V. As a result, we should find a reasonable way to determine the linear transformation T.

[31] proposed an approach called minimum torsion to find the proper transformation T. The rationale of the method is as follows. First note that in practice, we usually consider a portfolio which puts small weights on a large number of assets to be more diversified than a portfolio which puts large weights on a few assets. This implies that the linear transformation T should keep that property of the original weights, i.e. if w is close to equally-weighted, then $v = (T^{\mathsf{T}})^{-1}w$ should also be close to equally-weighted and vise versa. Thus we want the change of the original w to be as small as possible while making the covariance diagonalized. The degree of the change is measured by normalized tracking error (NTE) given by:

$$NTE(T) = \sqrt{\frac{1}{n}tr\big(diag(\Sigma)^{-\frac{1}{2}}(T-I)\Sigma(T-I)^{\mathsf{T}}diag(\Sigma)^{-\frac{1}{2}}\big)},$$

where $tr(\cdot)$ denote the trace of the matrix. This can be viewed as a distance between T and identity transformation with metric Σ. And our goal is to find T that minimize $NTE(T)$ while $T\Sigma T^{\mathsf{T}}$ is diagonal. Applying the representation (35) we can rewrite the minimization problem as:

$$\min_{D,V} tr(D - 2D^{\frac{1}{2}}VS^{\frac{1}{2}}U^{\mathsf{T}}),$$
$$\text{s.t. } D \text{ is diagonal, } V \text{ is unitary.}$$

Although we have shown that the ENB does not depend on the choice of D, but the above optimization problem clearly depends on D. If we fix D in the above equation to be identity matrix, then the optimal solution would simply be $V^* = U$. If there is no further restrictions on D then we can solve the minimization problem using an iterative algorithm which converges very fast, we refer [31] for details.

4.5 Generalized Effective Number of Bets

In this section we extend the ENB to general risk measures. First we present the major results in [43].

Definition 6. *A function $r : \mathbb{R}^n \to \mathbb{R}$ is τ-homogeneous if for each $w \in \mathbb{R}^n$ and $t > 0$ we have $t^\tau r(w) = r(tw)$.*

Proposition 9 (Tasche)**.** *Let $r : \mathbb{R}^n \to \mathbb{R}$ be a real-valued function and $\tau \in \mathbb{R}$ be fixed.*

1. *If r is τ-homogeneous and partially differentiable in w_k for some $k = 1, \ldots, n$, then the derivative $\partial r / \partial w_k$ is $(\tau - 1)$-homogeneous.*
2. *If r is totally differentiable then it is τ-homogeneous if only if for all $w \in \mathbb{R}^n$*

$$\tau r(w) = \sum_{k=1}^{n} w_k \frac{\partial r}{\partial w_k}(w) = w^\mathsf{T} \nabla r(w). \tag{36}$$

We can think r as the measure of risk and w as the weight of a portfolio. For example $r(w) = w^\mathsf{T} \Sigma w$ is a 2-homogeneous function. If $\tau \neq 0$ the above theorem tells us that the portfolio risk can be decomposed as the sum of $\frac{w_k}{\tau} \frac{\partial r}{\partial w_k}(w)$, $k = 1, \ldots, n$, which are the marginal contributions to risk of each asset. So similar as the covariance case we may define

$$p_k(v) = \frac{v_k}{\tau r(v)} \frac{\partial r}{\partial v_k}(v),$$

and compute the ENB.

However there are two problems need to be solved. First, similar as the covariance case which we have discussed before, $\frac{\partial r}{\partial v_k}(w), k = 1, \ldots, n$ are not independent. Secondly, $p_k(v)$ may be negative so its entropy is not well-defined.

To solve the first problem, assume that r is totally differentiable, then from Taylor expansion we have the following approximation

$$r(w + \Delta w) \approx r(w) + \Delta w^\mathsf{T} \nabla r(w) + \frac{1}{2} \Delta w^\mathsf{T} H_r(w) \Delta w,$$

where $H_r(\cdot)$ is the Hessian matrix of r and Δw is a small perturbation of w. If r is convex then H_r is semi-positive definite. This implies that $r(w)$ can

be approximated by a quadratic function which is similar as covariance in a small neighborhood of w. Therefore by diagonalizing $H_r(w)$ we can extract locally independent marginal contributions to $r(w)$ as what we discussed in the previous section. Let $T(w)H_r(w)T(w)^\intercal = D(w)$ where $D(w)$ is a diagonal square matrix and $T(w)$ is invertible. Then we may change the local coordinate system whose origin is w using linear transformation $T(w)$ so that

$$\begin{aligned}
&r(w+\Delta w)\\
&\approx r(w) + \Delta w^\intercal T(w)^{-1} T(w) \nabla r(w) + \frac{1}{2}\Delta w^\intercal T(w)^{-1} D(w) (T(w)^\intercal)^{-1} \Delta w\\
&= r(w) + \Delta v^\intercal T(w) \nabla r(w) + \frac{1}{2}\Delta v^\intercal D(w) \Delta v,
\end{aligned}$$

where $\Delta v = (T(w)^\intercal)^{-1}\Delta w$. An important result of proposition 9 is that $(\nabla r(w))_k$, $k = 1, \ldots, n$ is also $\tau - 1$ homogeneous:

$$(\tau - 1)(\nabla r(w))_k = \sum_{j=1}^{n} w_j \frac{\partial^2 r}{\partial w_k \partial w_j}(w) = (H_r(w)w)_k.$$

Therefore:

$$\begin{aligned}
(\tau - 1) T(w) \nabla r(w) &= T(w) H_r(w) w\\
&= T(w) H_r(w) T(w)^\intercal (T(w)^\intercal)^{-1} w = D(w) v,
\end{aligned}$$

where $v = (T(w)^\intercal)^{-1} w$. If $\tau > 1$

$$r(w) = \frac{1}{\tau} w^\intercal \nabla r(w) = \frac{1}{\tau} v^\intercal T(w) \nabla r(w) = \frac{1}{\tau(\tau-1)} v^\intercal D(w) v.$$

Thus we decompose $r(w)$ into n risky factors $\frac{1}{\tau(\tau-1)} d_k(w) v_k^2$, $k = 1, \ldots, n$ where $d_k(w)$ is the diagonal elements of $D(w)$ which is greater or equal to zero since $H_r(w)$ is semi-positive definite. So the second problem is solved by letting $\tau > 1$. To see why these factors are locally independent, we can

rewrite the Taylor expansion of r as:

$$\begin{aligned} r(w+\Delta w) &= \frac{1}{\tau}(w+\Delta w)^{\mathsf{T}}\nabla r(w+\Delta w) \\ &\approx \frac{1}{\tau}(w+\Delta w)^{\mathsf{T}}(\nabla r(w)+H_r(w)\Delta w) \\ &= \frac{1}{\tau}(v+\Delta v)^{\mathsf{T}}\bigg(\frac{1}{\tau-1}D(w)v+D(w)\Delta v\bigg) \\ &= r(w)+\frac{1}{\tau-1}\Delta v^{\mathsf{T}}D(w)v+\frac{1}{\tau}\Delta v^{\mathsf{T}}D(w)\Delta v, \end{aligned}$$

where each components of the small increment Δv has approximately independent contributions to the difference $r(w+\Delta w)-r(w)$. Finally we have the ENB:

$$\begin{aligned} p_k(w) &:= \frac{d_k(w)v_k^2}{\tau(\tau-1)r(w)},\ k=1,\dots,n \\ N(w) &:= \exp\Big(-\sum_{k=1}^{n}p_k(w)\log p_k(w)\Big). \end{aligned}$$

Unlike the covariance case, the Hessian matrix depends on the choice w, so each time the linear transformation T has to be recomputed. But proposition 8 and lemma 1 is still valid. That is, let

$$C(w) := diag(H_r(w))^{-\frac{1}{2}}H_r(w)diag(H_r(w))^{-\frac{1}{2}}$$

be the 'correlation' of the Hessian matrix and $U(w)$, $S(w)$ be the SVD of $C(w)$. Then $T(w)$ must have the representation:

$$T(w) = D^{\frac{1}{2}}VS(w)^{-\frac{1}{2}}U(w)^{\mathsf{T}}diag(H_r(w))^{-\frac{1}{2}},$$

where D is diagonal and V is unitary. And it is clear that the choice of D does not effect the ENB $N(w)$. So we define the constrained minimum torsion linear transformation as

$$T_{MT}(w) := U(w)S(w)^{-\frac{1}{2}}U(w)^{\mathsf{T}}diag(H_r(w))^{-\frac{1}{2}},$$

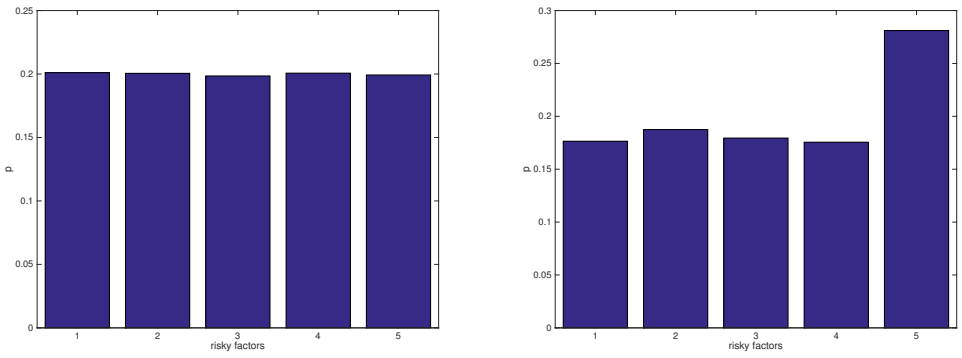
and the corresponding ENB as $N_{MT}(w)$.

Now given a coherent risk measure ρ and define $r_\rho(w) := \rho(w^{\mathsf{T}}X)$, then it is clear that $r_\rho(w)$ is a convex and 1-homogeneous function due to the

subadditivity and positive homogeneity of ρ. Thus we have to consider the power of r_ρ, or more specifically, the square of r_ρ in order to apply the ENB to the coherent risk measure. This is the same reason why we use variance instead of standard deviation to compute the ENB.

Here we give a simple numerical example to illustrate the advantage of the generalized ENB. Consider a unrealistic portfolio which includes five independent assets. The return of first four assets follow identically standard normal distribution. While the return of the fifth asset has a normalized student's t distribution with degree of freedom $\nu = 5$ and unit variance. So the covariance of these assets is just identity, it is obvious that the original ENB of the equally weighted portfolio is just 5 and the corresponding normalized marginal distribution p given by (34) is just $p_1 = \cdots = p_5 = 0.2$. This implies that the equally weighted portfolio is perfectly diversified.

This is clearly not true since the last asset has very heavy tails and is more risky than the others. To compute the CVaR-based ENB, we generate 200,000 i.i.d samples of the 5 assets' returns. Then we are able to compute the Hessian matrix of squared-CVaR with $\alpha = 0.01$ using proposition 6 and 7 with Monte Carlo. The conditional distribution $p(\cdot|X_2, X_3, X_4, X_5)$ is just the density of standard normal distribution because of the independency. Note that the Hessian of squared-CVaR is not necessarily diagonal even if the assets' returns are all independent. A simple calculation will show that the generalized ENB is 4.905. It would be much clear by looking at figure 1 which compares the normalized marginal contribution p of two ENB.



(a) Variance-based ENB (b) CVaR-based ENB

Figure 12: Comparison of two ENB

For comparison we also compute the variance-based ENB using the sample covariance matrix. It is clear that the CVaR-based ENB depends on the tail risk instead of variance.

One might ask whether the ENB would be smaller if we reduce the weight of the heavy-tailed asset or replaced it by an asset with thiner tails. The answer is given by figure 2.

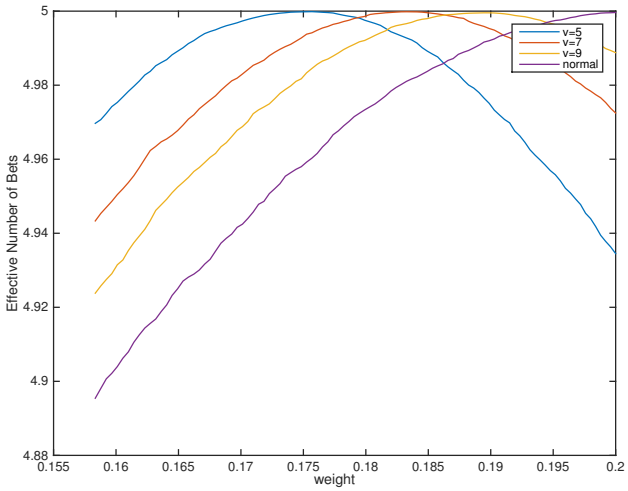


Figure 13: ENB with different weights and ν

Figure 2 plots the CVaR-based ENB against the weight of the fifth asset. Here we keep other assets' weights to be identical and the sum of weight to be 1. Each curve corresponds to different distributions of the last asset's return: student's t with $\nu = 5, 7, 9$ and standard normal distribution. Observe that the blue curve ($\nu = 5$) reaches it maximum around 0.175 instead of 0.2 of the equally weighted portfolio. As the degree of freedom increase, or the tail risk reduces, the optimal weight is closer to 0.2. Finally note that the ENB in this case would above 4 unless the last asset's weight is exactly zero.

Now we use the returns of Dow Jones 30 stocks from 2005 to 2013 to test the performance of the generalized ENB. The data is separated into 3 groups: 2005-2007 (prior-crisis), 2008-2010 (crisis), 2010-2013 (post-crisis). We first use 2005-2007 data to fit the GH distribution via EM algorithm. Based on our i.i.d GH model, we construct four constantly rebalanced and long only portfolios by (i) letting all weights to be equal, (ii) maximizing the

variance-based ENB, (iii) minimizing the CVaR, (iv) maximizing the CVaR-based ENB, respectively. The weights of the portfolios are given by figure 14.

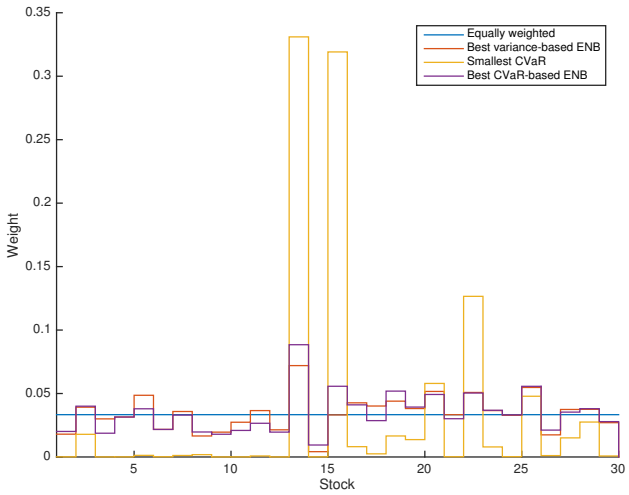


Figure 14: Portfolio weights

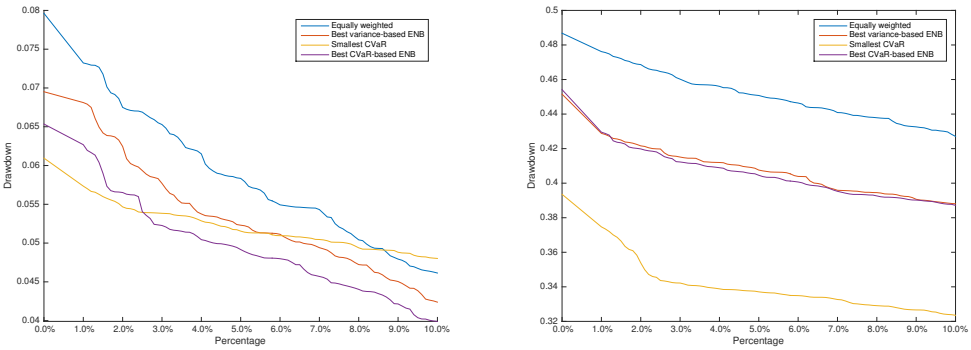
The equally-weighted portfolio is represented by the blue flat line. One can observe that the portfolio with the smallest CVaR concentrates only on a few stocks. This corresponds to our observation in the last section. The variance-based and CVaR-based ENB are very similar; even the Hessian of CVaR and the covariance matrix are quite different. Table 15 shows the variance-based ENBs, CVaR-based ENBs and CVaRs of four portfolios.

	Variance-ENB	CVaR-ENB	CVaR
Equally-weighted	28.7637	28.3638	0.0245
Best variance-ENB	29.4521	29.2177	0.0231
Smallest CVaR	14.9157	15.3699	0.0182
Best CVaR-ENB	29.2859	29.3985	0.0224

Table 15: Portfolio statistics

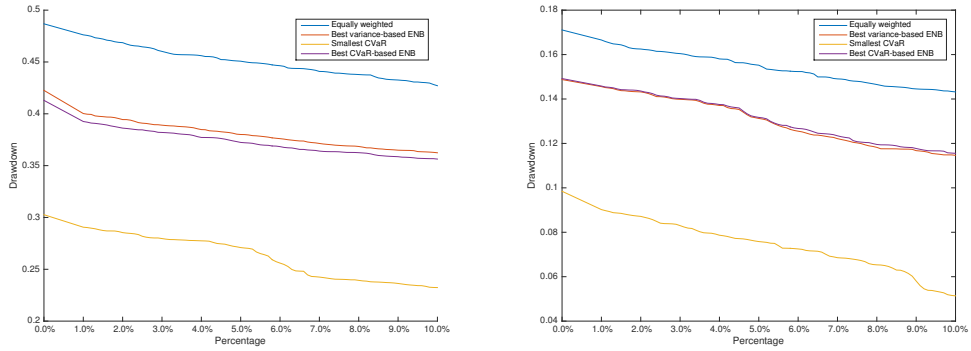
Then we compute the drawdowns of each portfolio during 2005-2007 (in-sample) and 2008-2010 (out-of-sample). The quantiles of the drawdowns of

each portfolio are shown by figure 16. The same test is applied to 2008-2010 (in-sample) and 2011-2013 (out-of-sample) data. The drawdown quantiles are given by figure



(a) In-sample 2005-2007 (b) Out-of-sample 2008-2010

Figure 15: Drawdown quantiles 2005-2010



(a) In-sample 2008-2010 (b) Out-of-sample 2011-2013

Figure 16: Drawdown quantiles 2008-2013

The values at 0% percentage is the well-known maximum drawdown. The smallest CVaR portfolio has the smallest maximum drawdowns in all four figures. The drawdowns of ENB portfolios are consistently better than equally-weighted benchmark. The CVaR-based ENB portfolio outperforms the variance-based one in both in-sample tests; but their differences become

very small in out-of-sample tests. One possible explanation for which the CVaR-based and the variance-based ENBs are so similar is that the tail of the GH distribution and other normal mixture distributions is controlled by a single random variable. Therefore all the components of a GH random vector have the same tails. The differences between the marginal contributions to CVaR are determined by the dispersion matrix Σ and the skewness vector γ, both of which are the components of the covariance matrix.

On the other side, the 30 stocks became strongly correlated during the financial crisis. Thus a portfolio that is well-diversified among these 30 stocks is still exposed to the downside risk. The CVaR portfolio, on the other side, only picks several stocks with the lowest historical risks. This might be the reason why the smallest CVaR portfolio outperforms the diversified ones in terms of controlling the drawdown risk. The advantage of portfolio diversification is not significant in a small pool of equities. However, maximizing the generalized ENB in high dimension is numerically difficult. A fast approach to compute the gradient and Hessian of the ENB is not known yet.

5 Conclusion and Future Work

This dissertation addresses the parameter estimation and portfolio allocation problems with the GH distribution. A lot of the algorithms and techniques introduced in this paper can also be applied to a general class of normal mixture distributions. The most attractive properties of the GH distribution are: (i) it has heavy tails; (ii) it introduces skewness; (iii) it has a natural multivariate extension; (iv) the multivariate GH distribution is closed under linear transformation; (v) the joint distribution of a GH variable and its GIG subordinator is an exponential family.

However there are also several drawbacks of the GH distribution and other normal mixture distributions. First of all, these distributions are nearly elliptical and therefore their components have the same tails. And these components can only be uncorrelated but never be independent because they share the same univariate subordinator. In financial markets however, different assets clearly have different tail risks; and there might be independent risky drivers behind them. Therefore a mixture of these assets such as financial index often has smaller tails than the individual assets, due to the central limit theorem. The multivariate GH distribution fails to capture these stylized facts. Variety of extensions of the GH distribution are constructed to solve

these problems; but efficient parameter estimation and portfolio optimization methods still need to be discovered.

The linear combination of independent GH random variables does not follow the GH distribution. This makes it very hard to apply the GH distribution to a majority of financial econometrics models such as ARMA-GARCH, in which the residuals are assumed to be i.i.d. A possible way to solve this problem is to use Gaussian copula together with GH marginal distributions to model the cross-sectional dependency of a financial time series. Therefore a lot of statistical properties of the Gaussian time series models can be preserved.

The on-line EM algorithm we introduced in this paper is just one of the many on-line density estimation algorithms. For parameter estimation, the on-line learning technique is not just a faster way to solve the convex optimization problem of maximum likelihood. It also naturally introduces some shrinkage estimators or Bayesian priors. For example, the Vovk-Azoury-Warmuth on-line forecaster proposed by [47] is closely related to the ridge regression, see [11]. The regret we derived in section 3.4 is also a simple example of the relationship between on-line learning and traditional information theory. Unfortunately we do not have an on-line EM algorithm for factor analysis yet. But we may solve the problem proposed at the end of section 3.5 by approximating the curved exponential family by a strict exponential family using the Talyor expansion of $\theta(u)$.

References

[1] Milton Abramowitz and Irene A Stegun. *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*, volume 55. Courier Corporation, 1964.

[2] Shun-Ichi Amari. Differential geometry of curved exponential families-curvatures and information loss. *The Annals of Statistics*, pages 357–385, 1982.

[3] Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. Coherent measures of risk. *Mathematical finance*, 9(3):203–228, 1999.

[4] Katy S Azoury and Manfred K Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Machine Learning*, 43(3):211–246, 2001.

[5] O Barndorff-Nielsen. Hyperbolic distributions and distributions on hyperbolae. *Scandinavian Journal of statistics*, pages 151–157, 1978.

[6] Ole E Barndorff-Nielsen. Processes of normal inverse gaussian type. *Finance and stochastics*, 2(1):41–68, 1997.

[7] Michele Leonardo Bianchi, Svetlozar T Rachev, Young Shin Kim, and Frank J Fabozzi. Tempered stable distributions and processes in finance: numerical analysis. In *Mathematical and Statistical Methods for Actuarial Sciences and Finance*, pages 33–42. Springer, 2010.

[8] Fischer Black and Myron Scholes. The pricing of options and corporate liabilities. *The journal of political economy*, pages 637–654, 1973.

[9] Olivier Cappé and Eric Moulines. On-line expectation–maximization algorithm for latent data models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):593–613, 2009.

[10] Peter Carr, Hélyette Geman, Dilip B Madan, and Marc Yor. The fine structure of asset returns: An empirical investigation*. *The Journal of Business*, 75(2):305–333, 2002.

[11] Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

[12] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

[13] Ernst Eberlein. Application of generalized hyperbolic lévy motions to finance. In *Lévy processes*, pages 319–336. Springer, 2001.

[14] Ernst Eberlein and Karsten Prause. The generalized hyperbolic model: financial derivatives and risk measures. In *Mathematical FinanceBachelier Congress 2000*, pages 245–267. Springer, 2002.

[15] Sonja Engmann and Denis Cousineau. Comparing distributions: the two-sample anderson-darling test as an alternative to the kolmogorov-smirnoff test. *Journal of Applied Quantitative Methods*, 6(3).

[16] Eugene F Fama. Mandelbrot and the stable paretian hypothesis. *The journal of business*, 36(4):420–429, 1963.

[17] Martin Hellmich and Stefan Kassberger. Efficient and robust portfolio optimization in the multivariate generalized hyperbolic framework. *Quantitative Finance*, 11(10):1503–1516, 2011.

[18] Wenbo Hu. Calibration of multivariate generalized hyperbolic distributions using the em algorithm, with applications in risk management, portfolio optimization and portfolio credit risk. 2005.

[19] William James and Charles Stein. Estimation with quadratic loss. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 361–379, 1961.

[20] Bent Jørgensen. *Statistical properties of the generalized inverse Gaussian distribution*, volume 9. Springer Science & Business Media, 2012.

[21] Young Shin Kim, Rosella Giacometti, Svetlozar T Rachev, Frank J Fabozzi, and Domenico Mignacca. Measuring financial risk and portfolio optimization with a non-gaussian multivariate model. *Annals of operations research*, 201(1):325–343, 2012.

[22] Young Shin Kim, Svetlozar T Rachev, Michele Leonardo Bianchi, and Frank J Fabozzi. Tempered stable and tempered infinitely divisible garch models. *Journal of Banking & Finance*, 34(9):2096–2109, 2010.

[23] Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2):365–411, 2004.

[24] Burton G Malkiel and Eugene F Fama. Efficient capital markets: A review of theory and empirical work. *The journal of Finance*, 25(2):383–417, 1970.

[25] Benoit Mandelbrot. The variation of certain speculative prices. *The Journal of Business*, 36(4):394–419, 1963.

[26] Alexander J McNeil, Rüdiger Frey, and Paul Embrechts. *Quantitative risk management: concepts, techniques, and tools*. Princeton university press, 2010.

[27] Javier Mencía and Enrique Sentana. Multivariate location–scale mixtures of normals and mean–variance–skewness portfolio allocation. *Journal of Econometrics*, 153(2):105–121, 2009.

[28] Christian Menn and Svetlozar T Rachev. Calibrated fft-based density approximations for α-stable distributions. *Computational statistics & data analysis*, 50(8):1891–1904, 2006.

[29] Christian Menn and Svetlozar T Rachev. Smoothly truncated stable distributions, garch-models, and option pricing. *Mathematical Methods of Operations Research*, 69(3):411–438, 2009.

[30] Attilio Meucci. Managing diversification. 2010.

[31] Attilio Meucci, Alberto Santangelo, and Romain Deguest. Measuring portfolio diversification based on optimized uncorrelated factors. *Available at SSRN 2276632*, 2014.

[32] Stefan Mittnik, Toker Doganoglu, and David Chenyao. Computing the probability density function of the stable paretian distribution. *Mathematical and Computer Modelling*, 29(10):235–240, 1999.

[33] Stefan Mittnik, Svetlozar T Rachev, and Marc S Paolella. Stable paretian modeling in finance: Some empirical and theoretical aspects. *A Practical Guide to Heavy Tails*, pages 79–110, 1998.

[34] Rostislav S Protassov. Em-based maximum likelihood parameter estimation for multivariate generalized hyperbolic distributions with fixed λ. *Statistics and Computing*, 14(1):67–77, 2004.

[35] Svetlozar Rachev and Stefan Mittnik. *Stable Paretian models in finance*. John Willey\ & Sons, New York, 2000.

[36] Svetlozar T Rachev, Young Shin Kim, Michele L Bianchi, and Frank J Fabozzi. *Financial models with Lévy processes and volatility clustering*, volume 187. John Wiley & Sons, 2011.

[37] Svetlozar T Rachev, Christian Menn, and Frank J Fabozzi. *Fat-tailed and skewed asset return distributions: implications for risk management, portfolio selection, and option pricing*, volume 139. John Wiley & Sons, 2005.

[38] Svetlozar Todorov Rachev. *Handbook of Heavy Tailed Distributions in Finance: Handbooks in Finance*, volume 1. Elsevier, 2003.

[39] Hans Rau-Bredow. Value at risk, expected shortfall, and marginal risk contribution.

[40] Xiang Shi. Marginal contribution to risk and generalized effective number of bets. *Available at SSRN 2642408*, 2015.

[41] Xiang Shi, Lihua Zhang, and Young Shin Aaron Kim. A markov chain approximation for american option pricing in tempered stable-garch models. *Frontiers in Applied Mathematics and Statistics*, 1:13, 2015.

[42] Peter Tankov. *Financial modelling with jump processes*, volume 2. CRC press, 2004.

[43] Dirk Tasche. Risk contributions and performance measurement. *Report of the Lehrstuhl für mathematische Statistik, TU München*, 1999.

[44] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[45] Cristina Tortora, Paul D McNicholas, and Ryan P Browne. A mixture of generalized hyperbolic factor analyzers. *Advances in Data Analysis and Classification*, pages 1–18, 2013.

[46] Vladimir V Uchaikin and Vladimir M Zolotarev. *Chance and stability: stable distributions and their applications*. Walter de Gruyter, 1999.

[47] Volodya Vovk. Competitive on-line statistics. *International Statistical Review/Revue Internationale de Statistique*, pages 213–248, 2001.

[48] Shushang Zhu and Masao Fukushima. Worst-case conditional value-at-risk with application to robust portfolio management. *Operations research*, 57(5):1155–1168, 2009.

A Asymptotic Approximation of Modified Bessel function of the second kind

First we review the definition and some basic properties of modified Bessel functions.

Definition 7. *The modified Bessel function of the first kind $I_\nu(z)$ is defined as:*

$$I_\nu(z) := \sum_{k=0}^{\infty} \frac{(-1)^k}{\Gamma(k+\nu+1)k!}\left(\frac{z}{2}\right)^{2k+\nu},$$

where $\nu, z \in \mathbb{R}$ and $\Gamma(\cdot)$ is the gamma function. The modified Bessel function of the second kind $K_\nu(z)$ is defined as:

$$K_\nu(z) := \frac{\pi \csc(\pi\nu)}{2}(I_{-\nu}(z) - I_\nu(z)).$$

It is obvious from the definition that $K_\nu(z) = K_{-\nu}(z)$. There first derivative of $K_\nu(z)$ is given by:

$$\frac{d}{dz}K_\nu(z) = -K_{\nu-1}(z) - \frac{\nu}{z}K_\nu(z) = \frac{\nu}{z}K_\nu(z) - K_{\nu+1}(z).$$

On the interval $(0, \infty)$ $K_\nu(z)$ is a positive function that diverges as $z \to 0$ and decays exponentially as $z \to \infty$. The following equations show the asymptotic properties of $K_\nu(z)$;

$$K_\nu(z) \propto \sqrt{\frac{\pi}{2z}}e^{-z}, \text{ as } |z| \to \infty, \tag{37}$$

$$K_\nu(z) \propto \begin{cases} -\log\frac{z}{2} - \gamma_{em} & \text{if } \nu = 0 \\ \frac{\Gamma(\nu)}{2}\left(\frac{2}{z}\right)^\nu & \text{if } \nu > 0 \end{cases}, \text{ as } |z| \to 0, \tag{38}$$

where γ_{em} is the Euler-Mascheroni constant and $\Gamma(\cdot)$ is the gamma function. We refer [1] for more properties of the modified Bessel functions.

The asymptotic properties are very useful in the computation of the log of GH density (11) which contains $\log K_\nu(z)$. In practice $K_\nu(z)$ may exceed the largest floating-point number $\bar{N}$ in IEEE double precision when z is large or close to zero. In Matlab for example, the function `log(besselk(v,z))` may just return Inf or -Inf when its actual value is far smaller than $\bar{N}$.

There are many ways to compute $\log K_\nu(z)$, in this paper we just take the logarithm of (37) and (38) and adjust it by a constant. The constant is to fill the gap between the asymptotic values and the largest or smallest possible value from direct computation. Figure 17 and 18 plot the $\log K_\nu(z)$ with asymptotic tails when $\nu = 250, 500$. The blue line is the direct computation of `log(besselk(v,z))`. The red line represents the asymptotic approximation. One can observe that as ν increases the left red dash tail turns to be longer.

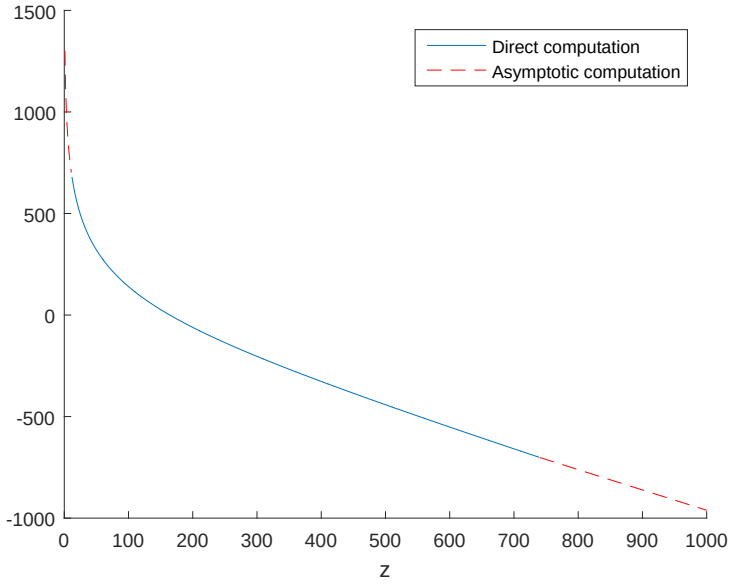


Figure 17: Asymptotic approximation of $\log K_{250}(z)$

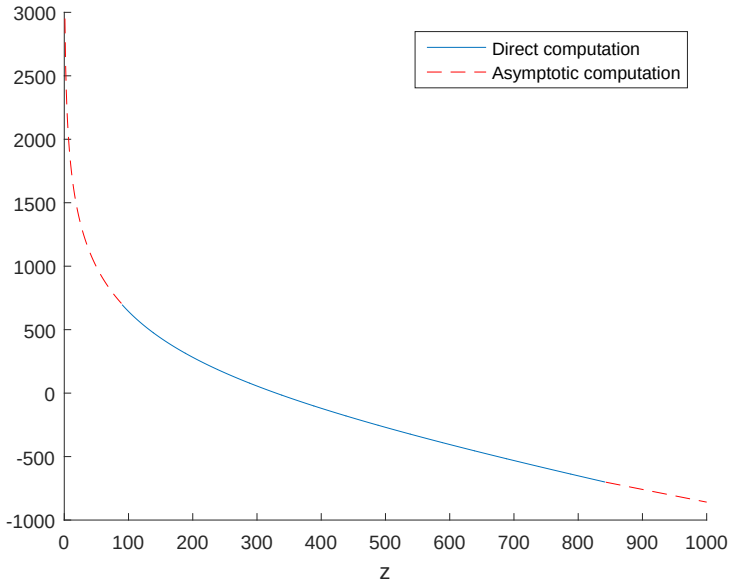


Figure 18: Asymptotic approximation of $\log K_{500}(z)$

The asymptotic approximation of course is not a perfect approach for the computation of $\log K_\nu(z)$. The error of the right side asymptotic approximation, for example, would increase as λ grows large. To our experience the approximation is good enough for about 500 dimension. A better numerical computation of $\log K_\nu(z)$ would be very helpful if we want to calibrate the GH distribution under higher dimensions.

The computation of the ratio $K_{\nu_1}(z)/K_{\nu_2}(z)$ is even more crucial than $\log K_\nu(z)$ since it is the key step (15) in the EM algorithm. The idea of approximation is basically the same. Figure 19 and 20 compare the asymptotic approximation and the direct computation. The vertical axis is measured by logarithmic scale.

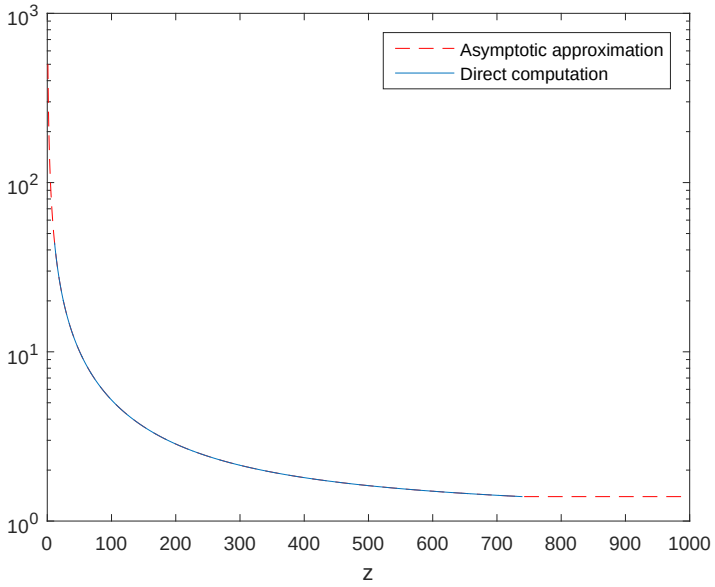


Figure 19: Asymptotic approximation of $\frac{K_{-251}(z)}{K_{-250}(z)}$

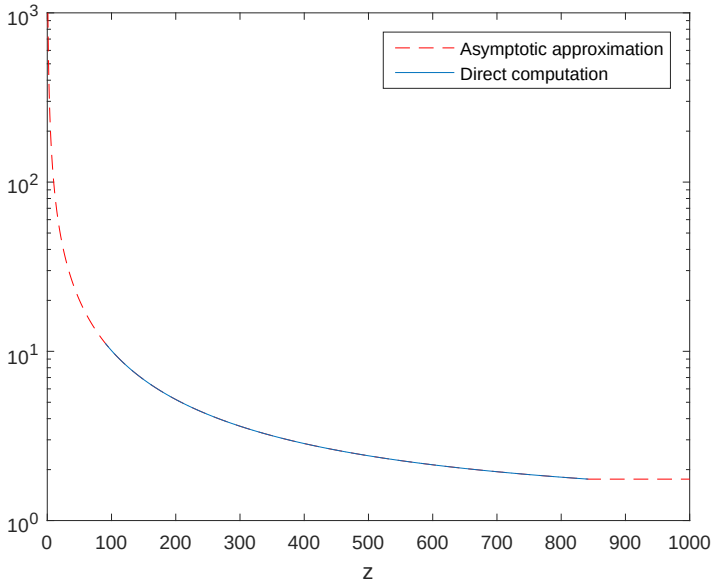


Figure 20: Asymptotic approximation of $\frac{K_{-501}(z)}{K_{-500}(z)}$