Stony Brook University

ERROR REDUCTION IN PARAMETER ESTIMATION WITH CASE STUDIES IN RISK MEASUREMENT AND PORTFOLIO OPTIMIZATION

A Dissertation Presented

by

Xiaoping Zhou

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 2013

Stony Brook University

The Graduate School

Xiaoping Zhou

We, the dissertation committee for the above candidate for the Doctor of Philosophy degree, hereby recommend acceptance of this dissertation.

Svetlozar Rachev – Dissertation Advisor
Director of Quantitative Finance Program
Department of Applied Mathematics and Statistics, Stony Brook University
Professor of Finance
College of Business, Stony Brook University

James Glimm – Chairperson of Defense
Distinguished Professor
Department of Applied Mathematics and Statistics, Stony Brook University

Haipeng Xing – Committee Member
Associate Professor
Department of Applied Mathematics and Statistics, Stony Brook University

Noah Smith – External Committee Member
Assistant Professor
College of Business, Stony Brook University

This dissertation is accepted by the Graduate School

Charles Taber
Interim Dean of the Graduate School

Abstract of the Dissertation

ERROR REDUCTION IN PARAMETER ESTIMATION WITH CASE STUDIES IN RISK MEASUREMENT AND PORTFOLIO OPTIMIZATION

by

Xiaoping Zhou

Doctor of Philosophy

in

Applied Mathematics and Statistics

Quantitative Finance

Stony Brook University

2013

Parameter estimation is an important step in probabilistic and statistical modeling. Maximum likelihood estimators (MLE) are the classical parameter estimators. However, in the real world, for small samples or ill-posed problems maximum likelihood estimates may be difficult to find through direct numerical optimization, and are unstable and sensitive to outliers.

In this dissertation, we study two popular problems in financial risk management and portfolio selection. The first problem is operational risk modeling. Data insufficiency and reporting threshold are two main difficulties in estimating the loss severity distributions in operational risk modeling. We investigate four methods including MLE, expectation-maximization (EM), penalized likelihood estimator (PLE), and Bayesian method. Without expert information, Jeffreys' priors for truncated distributions are used for the Bayesian method. Using the popular lognormal distribution as an example, we provide an extensive simulation study to demonstrate the superiority of Bayesian method in reducing parameter estimation errors for truncated distributions with small sample size. In addition, we apply the methods to actual operational loss data from a European bank using the log-normal and log-gamma distributions. The stability and credibility of the parameters are improved compared to MLE, and the granularity of units of measures for operational risk modeling are improved.

The second problem is a classical high dimensional problem—covariance estimation. Sample covariance is known to be a poor input when the sample size is relative small compared to the dimension. There is a vast literature that suggests factor models, shrinkage methods, random matrix theory (RMT) approaches, Bayesian methods, and regularization methods for dealing this problem. We consider an interesting case where there is a natural ordering among the random variables. We study a smooth monotone regularization approach, which is often useful for fixed-income instruments and options when the instruments have a natural ordering (e.g., by maturities, strikes, or ratings). We analyze the performance of smooth monotone covariance in reducing various statistical distances and improving optimal portfolio selection. We also extend its use in non-Gaussian cases by incorporating various robust covariance estimates for elliptical distributions. Finally, we provide two empirical examples where the smooth monotone covariance improves the out-of-sample

covariance prediction and portfolio optimization.

Contents

Acknowledgements

In the period of my doctoral study, I have received considerable help and support from numerous people.

I need to first thank Prof. Xiaolin Li, who was the graduate program director when I was applying for admission to Stony Brook's AMS PhD program. Without the Full Tuition Scholarship, I would not have the opportunity to come to study at Stony Brook.

I am also thankful to Prof. Ann Tucker, who was a visiting associate professor at AMS department and the executive director of Stony Brook's Quantitative Finance (SBQF) program in the early stage. As the main contributor in working towards the official startup of the SBQF program, she brought us fresh knowledge about quantitative finance from her industry experience, and provided us a fabulous platform as in the QF program to learn about the mysterious "Wall Street". Luckily, I was one of the few students at the start of the QF program. She was kind and friendly to every student, and we missed her so much when she left the department.

Prof. Svetlozar Rachev, my academic advisor, undoubtedly provided me huge guidance on how to conduct research and how to present research to the academic world. He was always available and patient for any discussion not only limited to academic research and he was a beloved professor by all students at Stony Brook. He also brought his wide and strong research nexus and provided us numerous opportunities to study the latest and most challenging topics in quantitative finance, especially about risk management. My first research topic — operational risk modelling — actually dates back to a visiting talk of Prof. Rachev in 2010 at Stony Brook. After Prof. Ann Tucker's interim guidance, I became a PhD student of Prof. Rachev later, with operational risk modelling as my first PhD topic. In this research, I received the most important resource for operational risk research — real operational risk loss data — from Prof. Rosella Giacometti, one of Prof. Rachev's collaborators. I need to thank Prof. Giacometti for her kind and careful review with constructive discussion in our communication. I am also grateful to Prof. Frank Fabozzi, another collaborator of Prof. Rachev, for his earnest and conscientious review and corrections for the manuscripts, which is decisive to the advent of the first journal paper in my PhD study. I have also learned a lot from him about academic writing skills and experience, which will benefit me for a life-long time.

Prof. Andrew Mullhaupt, who had over twenty years experience in high-frequency trading, brought us an insider vision to the state-of-the-art science and technology in quantitative finance. His lectures and ideas were always insightful and inspiring. He was always generous and kind in sharing his opinions and providing his guidance. I am thankful to all his help in answering my questions and feel fortunate that I had the opportunity to learn from him.

Dr. Dmitry Malioutov, who was actually the initial author of the smooth monotone covariance from which the second topic of this dissertation originates, provided considerable help and beneficial guidance in my research. We had very active communication through email and he always replied to me timely with very detailed answer and constructive opinion. I greatly appreciate his highly responsible attitude and multi-task spirit as he was always busy at his work in the meantime.

I also want to thank Prof. Robert Frey for his enormous effort in starting such a wonderful program to attract so many prominent professors, and his friendliness and patience in communicating with us on line. Other professors such as Prof. Haipeng Xing and Prof. Jiaqiao Hu also provided useful comments in an earlier stage of the first topic. I also want to express my appreciation to many colleagues including (but not limited to) Xu Dong, Rong

Lin, Naoshi Tsuchida and Ke Zhang, with whom I had chance to work together or discuss about problems and ideas during my research.

Since the last semester of my PhD study, I have been working as an operational risk quantitative analyst at RBS citizens bank. In addition to academic research on campus, this industry experience has provided me deeper understanding about practical operational risk modelling. I am grateful to this opportunity and I want to express my special thanks to my supervisor Antonina Durfee. And, I am thankful to the beneficial discussions with my colleagues Dane Johnson and Qiang Gao.

Last but most, I would like to thank the support from my family members during all my study life, without whom I would not succeed in any accomplishments.

Vita, Publications and/or Fields of Study

I was born in Jiujiang, Jiangxi, P.R. China in 1987. I received my B.S in Mathematics and Applied Mathematics and M.S in Mathematics at Beijing Institute of Technology in 2007 and 2009 respectively. My research interests include applied probability and statistics, quantitative risk management and quantitative trading.

I was admitted to the PhD program in Applied Mathematics and Statistics at Stony Brook University with Full Tuition Scholarship in 2009. My concentration is quantitative finance and my academic advisor is Professor Svetlozar Rachev. My research topics in PhD include operational risk modelling, portfolio optimization and covariance estimation. The related research work include two published papers and one working paper:

1. Xiaoping Zhou, Rosella Giacometti, Frank J. Fabozzi, Ann H. Tucker. Bayesian estimation of truncated data with applications to operational risk measurement. Quantitative Finance. 1-26. (2013)

2. Xiaoping Zhou, Dmitry Malioutov, Frank J. Fabozzi and Svetlozar T. Rachev. Smooth monotone covariance for elliptical distributions and applications in finance. Working paper. (2013)

3. Naoshi Tsuchida, Xiaoping Zhou, Svetlozar T. Rachev. Mean-ETL Portfolio Selection under Maximum Weight and Turnover Constraints Based on Fundamental Security Factors. The Journal of Investing, 21(1), 14-24. (2012)

I am currently working as an operational risk quantitative analyst in the Advanced Analytics group at RBS citizens bank, Boston, MA.

1 Introduction

Probabilistic and statistical models are important tools to extract information and uncover the hidden laws behind observable phenomena in financial modeling and risk management. How to build a model and how to estimate the model also vary under different goals of modelling. There are usually two branches in financial modelling. First is large data analysis. Large sample theory and high dimensional models have found important applications in this area. Time series analysis in the econometrics and finance world and signal processing in the engineering world are popular tools in developing models to analyze the historical data of observed variables and forecast their future movements. In econometrics world, the research direction usually lies in constructing heavy-tailed models and estimating them using maximum likelihood method (or, in most cases, quasi maximum likelihood method). The abundant research include ARMA-GARCH type time series models (Engle, 1982; Bollerslev, 1986; Tsay, 2002) and heavy-tailed distributions (e.g., normal mixture models (Kelker, 1970; Cambanis et al , 1981), generalized hyperbolic distributions (Barndorff-Nielsen, 1978), stable and tempered stable distributions (Rachev, Menn and Fabozzi, 2005; Rachev et al, 2011)). On the other hand, in engineering world, the research applied to finance usually focus more on state-space models and the estimation methods are mostly least-squares methods with robustness adjustment. The econometric and stochastic process models tend to mimic the actual price/return process by capturing the characteristics reflected in real process, while the engineering models tend to capture the driven trend and filter out the noise effect. Both models are better suitable when a large sample is available for modelling.

Another branch of financial modelling is small-sample problems when there are not enough data or ill-posed problems when the classical estimation methods could not generate stable and reliable estimates. This is the common feature of the two practical problems that are studied in this dissertation. The first problem is operational risk modelling. Operational risk is defined as, according to Basel II Accord (BIS, 2004), the risk of direct or indirect loss resulting from inadequate or failed internal processes, people and systems or from external events. Due to data availability, operational risk, though greatly important to banking management and survival, were modeled by simple qualitative approach in the past. Since 2002, US and European banks have started collecting operational losses and only in recent years, the large banks in US are required to calculate the operational risk capital in a more quantitative way.

Loss distribution approach (LDA) is one of the advanced measurement approaches (AMA) in quantifying operational risk. It assumes the losses follow a compound distribution of loss frequency distribution and loss severity distribution. Operational risk capital is then calculated as the 99.9% quantile of the annual aggregated loss distribution. Given the limited loss data, an extrapolation beyond the empirical losses is often required thus a parametric form is usually assumed for the loss distribution. Compared to the loss frequency distribution modelling, there are much more difficulty in modelling the loss severities. First of all, there are often very few losses for those event types where the loss frequency is low. Second, the loss distribution is highly right-skewed and has an exceptionally heavier tail compared to market or credit loss distribution. Fitting a heavy-tailed distribution to a small sample may result in large instability of the parameters. What exacerbates the issue is that the losses are often only reported if they are above a threshold by the bank. The losses below the threshold is not collected. Neglecting the reporting threshold and fitting a full distribution to the losses above the threshold clearly misspecifies the problem. Chernobai et al (2007) discussed the bias in estimated parameters and capital charges when

fitting a full distribution to a truncated sample from lognormal distribution. Instead, a truncated distribution is often suggested to fit the data. However, although the specification of the problem is easy, the maximum likelihood estimates of the parameters are much more difficult to find through numerical optimization and are often highly unstable leading to unreliable capital charges. To avoid dealing with truncated distributions, expectation-maximization (EM) algorithms are suggested. The EM algorithm is an iterative method for computing maximum likelihood estimates in the presence of missing data (Dempster et al, 1977). Bee (2005) developed the explicit EM iteration formulas for estimating parameters for truncated and censored samples from a lognormal distribution, and shows that it converges better than some gradient-based algorithms. However, when the expectation of complete-data sufficient statistics is hard to obtain in an explicit formula, the EM algorithm may not be simpler than the direct MLE. Thanks to the development of more advanced optimization toolboxes, the difficulty in numerical optimization has been relieved and the convergence of optimization to the maximum can be improved by carefully specifying the optimization algorithms, starting points and termination rules. However, the instability of the maximum likelihood estimates for truncated distribution could not be circumvented. This issue has not been stressed in academic research of operational risk until recent years. Cope (2011) proposed a penalized likelihood estimator for truncated data and discussed its performance in decreasing estimation error and increasing robustness than MLE. It is pointed out that the likelihood surface of a truncated distribution is greatly distorted by the denominator introduced into the density function due to data truncation, which leads to large variability and instability of the estimates. By adding a linear term related to direction with large variance in the parameter estimates, the penalized likelihood surface becomes more well-shaped and the optimum is easier to locate at the price of a small bias. However, this approach requires a good guess about a suitable penalty parameter which requires numerically calculating the Hessian of the likelihood function at true parameters. Without knowledge about true parameters, the optimal penalty parameter may be hard to find. In addition, when the Hessian is ill-conditioned, the direction and size of the penalty term may be impossible to obtain. Inspired by this paper, we considered using Bayesian methods for estimating parameters of truncated distributions to reduce the estimation error. In literature about operational risk, Bayesian methods are usually suggested as a way to combine expert opinions using elicit priors (see Cruz (2002) and Shevchenko (2011)), or only applied to full samples and found little difference from MLE (see Peters and Sisson (2006), Dalla Valle and Giudici (2008)). The reduction of estimation errors by Bayesian methods for truncated distributions especially for operational risk has never been studied. Without expert opinion information, non-informative priors are used in our study. We derived the Jeffreys' (1967) priors for truncated normal and gamma distribution since we found that lognormal and loggamma distributions are suitable for the loss data we have and also widely used as a standard example and a heavier-tailed extension respectively. As one of the non-informative priors, Jeffreys' prior is interesting for us because of its invariant form under different parametrization and simple representation as the square root of the determinant of Fisher information. We provide an extensive simulation study to compare the Bayesian method with the usual direct MLE, EM and PLE methods in reducing estimation error. In addition, we apply Bayesian methods in estimating loss severity distributions for an actual loss data set from a European large bank and provide an empirical study of operational loss modelling by calculating the capital charge for each intersection of event type and business line with available data.

The second problem studied in this dissertation is a classical high dimensional problem

— covariance estimation. Covariance estimation is a very important problem in statistics, engineering and finance. It is an important input for linear regression, signal processing and optimal portfolio selection. However, sample covariance is known to be a poor input when the number of samples n is about the dimension of random variables p and leads to unrobust or biased estimates for prediction or asset allocation. When the dimension exceeds the number of samples, the sample covariance is not positive definite and this leads to the multicollinearity problem in regression. Therefore, improved covariance estimates are greatly important for many applications.

There is a vast literature regarding improved covariance estimation, and we try to classify them into shrinkage estimators, random matrix theory (RMT) approaches, structured models and Bayesian estimators. They are not isolated from each other but connected to each other on many aspects.

In statistical inference, there is always a trade-off between bias and variance of an estimator. Shrinkage estimators are constructed in a way to "shrink" the plain-vanilla estimators toward some targets or priors so as to reduce the mean-squared error (MSE) by seeking a balance between bias and variance. The squared error's counterpart in high dimensions is the Frobenius norm of the difference between two covariance matrix. This metric is the simplest and does not require the positive-definiteness of the covariance matrix. It has been used as a criterion for deriving a wide variety of shrinkage covariance estimates such as Ledoit and Wolf (2003, 2004) Chen et al (2009, 2011) by weighting the sample covariance with an identity matrix. These estimators have been successfully used in obtaining well-conditioned covariance matrices regardless of whether n is greater than p and reducing MSE of covariance estimates. The shrinkage estimators can also be understood in an empirical Bayesian framework where the shrinkage target acts as the center of the prior distribution of covariance (Ledoit and Wolf, 2003). There are a series of Bayesian or empirical Bayesian estimators of covariance such as Efron and Morris (1976), Haff (1980), Daniels and Kass (1999, 2001) and Yang and Berger (1994).

Classical results from random matrix theory suggest that the eigenvalue spectrum of a sample covariance matrix is not consistent to the true spectrum. Marcenko and Pastur (1967) show that, when n and p both increase while the ratio n/p being a constant, the asymptotic spectrum of sample covariance matrix of Gaussian random variables has a known specific shape. Ledoit and Wolf (2003) also show that the eigenvalues of the sample covariance tend to be more dispersed than those of the true covariance. Therefore, there have been many attempts to adjust the value of the sample eigenvalues, called random matrix theory (RMT) approaches. These research, to name a few, include Stein (1975), Plerou et al (1999), Laloux et al (1999), Kwapien et al (2006), Conlon et al (2007), Park and O'Leary (2010). Karoui (2008) and Ledoit and Wolf (2012) instead tried to estimate the distribution of the eigenvalues by inverting the Marcenko-Pastur equation. All these estimators based on RMT tend to adjust the eigenvalues in a "nonlinear" way and are often called "nonlinear" shrinkage estimators, as opposed to the usual shrinkage estimators which shrink all the eigenvalues with a same rate.

Another type of estimation approaches come from the dimension reduction point of view. The number of parameters to be estimated in a covariance matrix increases as the squared dimension does. This is called "curse of dimensionality" and it leads to much noise or instability in parameter estimates due to large degrees of freedom. Therefore, there have been many researchers trying to impose a structure on the covariance matrix to reduce the degrees of freedom in the estimation. The first class is low-rank models, which correspond to a broad branch in multivariate analysis such as principal component analysis (PCA) and

factor analysis (see Rencher and Christensen, 2012). The application of PCA and factor models for covariance estimation can be seen in Elton and Gruber (1973), Bai and Shi (2011), and Fan et al (2008). The second class is imposing a sparse structure on the covariance by setting some elements to be zero, called thresholding method (Bickel and Levina, 2008a), or imposing a sparse structure on the inverse covariance, which is closely related to an assumption of conditional independence. There is a flourish research on sparse inverse covariance in both statistics and engineering world, called "covariance selection" models in the former and Gaussian Markov random field (GMRF) models in the latter. Some of contributions include, for instance, d'Aspremont et al (2008), Rothman et al (2008), Yuan and Lin (2007), and Friedman et al (2008).

When there is a natural ordering among the random variables, a particular structure can be imposed by utilizing this information. For example, by assuming variables far apart in the ordering have smaller correlation, Furrer and Bengtsson (2007) considered tapering the covariance matrix by gradually shrinking the off-diagonal entries within a band to zero. Bickel and Levina (2008b) considered a broad class of regularized covariance including banding the covariance matrix, banding the inverse covariance matrix, and tapering the covariance matrix. Related approaches can be found in Wu and Pourahmadi (2003, 2009) and Huang et al (2006). Malioutov (2011) proposed a nonparametric approach to regularize the covariance by imposing a monotone and smooth structure, which is where the second topic in this dissertation origins from and is based on. The monotone and smooth structure can be observed in many examples in finance such as autocorrelated time series, bonds with different maturities or ratings, options with respect to different strikes or maturities.

We study the error reduction of smooth and monotone covariance than sample covariance by analyzing a variety of other error metrics in addition to the Frobenius norm error. Fan et al (2008) showed that Frobenius norm error may not be a good measure to tell the difference between two covariance estimates. In literature, there are several other typical distances including the quadratic loss (or Frobenius norm error of inverse covariance to measure distances, see Haff (1980)), entropy loss (or, Stein's loss (James and Stein, 1961), which is closely related to one type of Kullback-Leibler divergence) and Hellinger distance (Lehmann and Romano, 2005). We analyze the relation between them and the performance of smooth monotone covariance under these distances. We also incorporate the smooth monotone covariance with various robust covariance estimates for elliptical distributions, which capture the heavy-tailedness better than Gaussian distribution.

The rest of this dissertation is structured as follows. Section 2 reviews usual parameter estimation methods and connections between them in light of error reduction. The first topic — operational risk modelling and improved parameter estimation for truncated data inference — is presented in Section 3 and Section 4. Section 3 introduces the background knowledge and practical issues in operational risk modelling. Section 4 focuses on the loss severity distribution modelling by comparing the proposed Bayesian method for truncated data with several other estimators, and includes an empirical study of operational risk measurement using real loss data from a European bank. The second topic — covariance estimation — is presented in Section 5 and Section 6, as a high dimensional case study of error reduction. Section 5 introduces the deficiency of sample covariance and reviews the literature regarding improved covariance estimates. Section 6 focuses on the smooth monotone covariance in comparison to the sample covariance in error reduction under various distance measures and practical applications in optimal portfolio allocation and out-of-sample covariance prediction.

2 Parameter Estimation Methods Review

In finance, probabilistic and statistical models are important for quantifying the uncertainty and extracting information from observed random samples. After a model is constructed, the most important problem is parameter estimation. In classical statistics, there are many methods for parameter estimation, each having its own advantages and disadvantages and being suitable for different problems. The oldest and simplest method dates back to the method of moments. However, for parametric models, with an assumed parametric family, the "likelihood" introduced by Fisher plays a more central role in modern parameter estimation methods. Maximum likelihood method has been one the most favorable and widely used approaches due to its asymptotic efficiency and pure objective property based on data. However, for small samples, it may lead to large variability in the estimated parameters and result in a large estimation error. Bayesian approaches instead treat the parameters as random variables whose distribution are updated by the available data. There have been long-time debates between "frequentist" and "Bayesians" for parameter estimation. Bayesian approaches have been successfully used as a way to incorporate prior information with sample information and also a way to reduce estimation error when maximum likelihood estimates are highly volatile. However, it has been criticized by the frequentists due to its subjectivity in choosing a prior. Therefore, there have been many researchers striving to find non-informative priors for Bayesian inference such as Jeffreys' prior (1967), and reference priors by a series work of Berger and Bernado and their collaborators (1979, 1991, 2009). Alternatively, empirical Bayesian methods take a trade-off between the "frequentist" and "Bayesians" by estimating the prior utilizing the information in the data. This approach has found many successful use in reducing estimation error with less subjectivity in specifying the prior.

On the other hand, there is another line of thought from robust estimation methods by reducing the impact of outliers. There are very common everyday examples such as calculating the average grades by removing the top and bottom scores. This is the trimmed mean from robust estimation. Another example is that the median is often more robust than the mean. In robust estimation, there are quantitative measures defined to evaluate the robustness of an estimator such as the breakdown point and the influence function. Penalized likelihood estimators are one of the robust estimators that have close relationship to maximum likelihood estimators. There are also estimators designed particularly to reduce the mean-squared error, from which James-Stein estimator in estimating mean or covariance might be the most well-known estimators. The estimators usually have a form of shrinking the sample estimators to certain prior and thus are also called shrinkage estimators. Interestingly, the shrinkage estimators can also be understood from an empirical Bayesian point of view, and sometimes may related to penalized estimators. In this section, we will provide an extensive review of main estimation methods for parametric models and their relation with each other.

2.1 Method of moments

The method of moments is a method of finding estimators by matching the population moments with sample moments, and then solving the equations for the quantities to be estimated.

Suppose that in order to estimate $\theta = (\theta_1, ..., \theta_k)$ of a distribution $f_X(x; \theta)$, and the first k moments of the distribution are: $\mu_1, ..., \mu_k$ as functions of θ. Given a sample $X =$

$(X_1, ..., X_n)$, assume that the observed sample moments are $m_i = \Sigma_{j=1}^{n} (X_j)^i, i = 1, ..., k$. Then the moment estimates of θ are found by solving:

$$\begin{aligned} m_1 &= \mu_1 = g_1(\theta_1, ..., \theta_k) \\ &... \\ m_k &= \mu_k = g_k(\theta_1, ..., \theta_k) \end{aligned}$$

This method of moments is often used in obtaining approximations to the distributions of statistics. It is called "moment matching" sometimes. The problem of this estimator is that it may result in estimates outside the parameter space (e.g., for Binomial distributions). In addition, it relies on the existence of moments with order as high as the number of parameters. In some respects, this method has been mostly superseded by Fisher's maximum likelihood method. However, in some cases, when the likelihood function is intractable, moment estimators are preferred.

2.2 Maximum likelihood estimators

Maximum likelihood estimators (MLE) are the most popular estimators in classical statistics. Intuitively, it finds the parameter point for which the observed sample is most likely to appear. Suppose $X = (X_1, ..., X_n)$ is an i.i.d. sample from a distribution with pdf or pmf $f(x; \theta)$, the likelihood function is defined as:

$$L(\theta|X) = \prod_{i=1}^{n} f(X_i; \theta)$$

In practice, it is often more convenient to work on the log-likelihood:

$$l(\theta|X) = \log L(\theta|X) = \Sigma_{i=1}^{n} \log f(X_i|\theta)$$

Then, the maximum likelihood estimate is obtained as:

$$\widehat{\theta} = \arg\max_{\theta \in \Theta} l(\theta|X)$$

The maximum likelihood estimator has some nice properties:

(1) Invariance property: If $\hat{\theta}$ is the MLE of θ, then for any function $g(\theta)$, the MLE of $g(\theta)$ is $g(\hat{\theta})$.

(2) Consistency: $\hat{\theta}_{MLE} \xrightarrow{p} \theta_0$.

(3) Asymptotic normality: Under certain regularity conditions, the maximum likelihood estimator has an asymptotically normal distribution

$$\sqrt{n}(\hat{\theta}_{MLE} - \theta_0) \xrightarrow{d} N(0, I(\theta)^{-1})$$

This property has also been often used as an approximation method to estimate a lower bound of MLE.

Theorem 2.1 *(Cramér-Rao Inequality)* The variance of any unbiased estimator $\hat{\theta}$ is greater than or equal to the inverse of Fisher information:

$$var(\hat{\theta}) \geq \frac{1}{I(\theta)}$$

where $I(\theta)$ is defined as:

$$I(\theta) = E\left[\left(\frac{\partial l(x;\theta)}{\partial \theta}\right)^2\right] = -E\left[\frac{\partial^2 l(x;\theta)}{\partial \theta^2}\right]$$

and $l(x;\theta) = \log f(x;\theta)$.
In higher dimension, for $\theta = (\theta_1, ..., \theta_k)$, then $I(\theta)$ is a $k \times k$ matrix with element:

$$I_{ij} = E\left[\frac{\partial l(x;\theta)}{\partial \theta_i}\frac{\partial l(x;\theta)}{\partial \theta_j}\right]$$

then the covariance matrix

$$cov(\hat{\theta}) \geq I(\theta)^{-1}$$

2.3 Bayes estimators

Bayesian approach and frequentist approach are two classical branches of probability. In frequentist approach, the source of uncertainty is the randomness inherent in realizations of a random variable. The probability of an event is the limit of its long-run relative frequency. Once the realizations are known and the distribution function is chosen, the probability distribution (e.g., the parameters of the distribution) is decided without any uncertainty. In contrast, the Bayesian approach treats probability distributions as uncertain (e.g., treats the parameters as random variables). It first assigns a distribution for the parameters, called prior distribution. When sample data become available, the distribution of the parameters is updated as the conditional distribution of the parameters given the information from sample data, called posterior distribution.

Suppose X is sample data from a probability distribution with density $f(x;\theta)$, where θ is the parameter. Assume the prior distribution of θ is $\pi(\theta)$, then the posterior distribution of θ is

$$\begin{aligned}\pi(\theta|x) &= \frac{p(\theta, x)}{p(x)} \\ &= \frac{p(x|\theta)\pi(\theta)}{\int p(x|\theta)\pi(\theta)d\theta}\end{aligned}$$

Bayesian estimators are obtained by minimizing the posterior expected value of a loss function (i.e., posterior expected loss). When the loss function is defined as the mean squared error

$$MSE = E[(\hat{\theta} - \theta)^2],$$

the realized estimator is the minimum mean squared error (MMSE)

$$\hat{\theta} = E[\theta|x] = \int \theta\pi(\theta|x)d\theta,$$

which is simply the mean of posterior distribution.

An alternative estimate within Bayesian statistics is the so-called maximum a posteriori (MAP) estimate, which is a mode of the posterior distribution:

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \pi(\theta|x)$$

There have been two major reasons that the Bayesian approach has not been widely acceptable as the classical MLE. First of all is the choice of prior distribution, and second is that the posterior expected loss is generally not easy to be calculated and one has to use numerical simulation methods.

2.3.1 Prior Selection

The selection of prior distribution is the most important problem in Bayesian estimation. The priors can be divided into three groups: elicit priors, vague (or flat) priors, and non-informative priors. Elicit priors can be used when strong prior information about the distribution is known, or when reliable and trustable expert opinion is incorporated into the estimation. The Bayesian approach has been widely used in practice such as in the Black-Litterman portfolio selection model to incorporate the purely objective mean-variance model with the subjective opinion from experienced traders or extra information that is not reflected in the market.

However, the Bayesian approach has also often been criticized due to its subjectivity of selecting prior distribution when no reliable expert information is available. Therefore, to alleviate the impact of prior on the parameter estimation, vague priors are often used. Vague priors are also called flat priors because the density of the prior distribution is almost flat over the parameter space. Vague (or flat) priors are often used when very weak information is known about the true distribution. For example, a normal distribution with a zero mean but a very large variance can be used for the prior distribution for the expected return when no information is available.

Nevertheless, the specification of a vague or flat prior distribution relies on the parametrization of the distribution. In other words, a vague prior in one parametrization may imply strong prior information under another parametrization. For example, when estimating the loss severity distribution of operational losses, a lognormal distribution $LN(\mu, \sigma^2)$ is usually chosen. The mean of a random variable from $LN(\mu, \sigma^2)$ is $e^{\mu+\sigma^2/2}$. If we want to select a prior distribution for μ and assume we know the empirical average loss amount of a single event is μ_0 around 10^5, a normal distribution $N(\mu_0, 10^4)$ or one may choose a normal distribution $N(\log_{10} \mu_0, 4)$. The extent of informativeness in the two distributions above are completely different. The first distribution contains much stronger information than the second. This leads to the advent of one of the most famous non-informative priors - Jeffreys' prior.

Jeffreys (1946) described a method of constructing a prior distribution such that the posterior is invariant under reparametrization. It is defined as a prior distribution that is proportional to the square root of the determinant of the Fisher information matrix:

$$p(\theta) \propto \sqrt{\det I(\theta)}.$$

This representation does not depend on the parametrization. Assume that there are two parametrization θ and φ, using the change of variables theorem, the Jeffreys priors $p(\theta)$ and $p(\varphi)$ are related to each other as follows:

$$
\begin{aligned}
p(\varphi) &= p(\theta)\left|\det\frac{\partial\theta_i}{\partial\varphi_j}\right| \\
&\propto \sqrt{\det I(\theta)\left(\det\frac{\partial\theta_i}{\partial\varphi_j}\right)^2} \\
&= \sqrt{\left(\det\frac{\partial\theta_i}{\partial\varphi_j}\right)\det I(\theta)\left(\det\frac{\partial\theta_i}{\partial\varphi_j}\right)} \\
&= \sqrt{\left(\det\frac{\partial\theta k}{\partial\varphi_i}\right)\det E\left[\frac{\partial\ln L}{\partial\theta_k}\frac{\partial\ln L}{\partial\theta_l}\right]\left(\det\frac{\partial\theta_l}{\partial\varphi_j}\right)} \\
&= \sqrt{\det E\left[\sum_{k,l}\frac{\partial\theta_k}{\partial\varphi_i}\frac{\partial\ln L}{\partial\theta_k}\frac{\partial\ln L}{\partial\theta_l}\frac{\partial\theta_l}{\partial\varphi_j}\right]} \\
&= \sqrt{\det E\left[\frac{\partial\ln L}{\partial\varphi_i}\frac{\partial\ln L}{\partial\varphi_j}\right]} = \sqrt{\det I(\varphi)}
\end{aligned}
$$

This implies that the Jeffreys' prior is invariant under any parametrization.

Example 2.1 *(Jeffreys' prior for normal distribution)* For a normal distribution with mean μ and standard deviation σ, the Jeffreys' prior for the mean is the flat prior $\pi(\mu) = 1$, and for the standard deviation is the inverse prior $\pi(\sigma) = 1/\sigma$,which means the $\log(\sigma)$ is flat. The Jeffreys' priors are usually improper priors because the integral of the prior distribution is infinite instead of one. When these improper priors are used, an estimator that minimizes the posterior expected loss is called a generalized Bayes estimator.

The Jeffreys' prior has been widely used in one-dimensional cases. However, there are difficulties with the method when there are multiple parameters, and there might be ad-hoc modifications to the prior needed. Bernardo (1979) introduced the reference prior method for developing noninformative priors by dividing the parameters into "parameter of interest" and "nuisance parameters". In usual one-dimensional problems, the approach yielded the Jeffreys' prior. There have been a series of work by Berger and Bernardo and their collaborators in defining and developing the reference priors such as Bernardo (1979), Berger and Bernardo (1989), and Berger and Bernardo (1991). Berger, Bernardo and Sun (2009) provides a formal definition of reference priors and a constructing formula for reference prior in one-dimensional cases. However, in general, the computation of reference prior is very difficult to use.

There is also another direction in constructing Bayesian predictive distributions using shrinkage priors. Pioneering work can be found in Komaki (2006), where it is shown that there exist shrinkage priors asymptotically dominating the Jeffreys' prior or other vague priors if the model manifold satisfies some differential geometric conditions under the Kullback-Leibler divergence loss function. Tanaka and Komaki (2006) proposed the superharmonic prior on the spectral density for autoregressive process.

2.3.2 Simulation Algorithms

The simulation algorithms of drawing samples from posterior distribution includes independence sampling and dependence sampling (Rachev et al, 2008). There are two typical

representatives in he first category: rejection sampling and importance sampling. In the second category, all algorithms are based on generation of a Markov chain, called Markov chain Monte Carlo methods.

- Rejection Sampling

Rejection sampling applies to the case when sampling from distribution $f(x)$ itself is difficult, but sampling from an envelope distribution of it $Mg(x)$, where $M > 1$ such that $f(x) < Mg(x)$, is easier. The algorithm is is following:

1. Sample x from $g(x)$;

2. Generate u from $U(0,1)$. If $u < f(x)/Mg(x)$, accept x; otherwise, reject x and re-sample.

The algorithm is also called acceptance-rejection algorithm.

- Importance Sampling

Importance sampling is a variance reduction by assigning more weights to the frequency of "important" samples. Suppose $h(X)$ is a function of random variable X and we are interested in calculating the expectation $E[h(X)]$. If X is difficult to simulate from $f(x)$, but easier to sample from $g(x)$, then according to

$$\begin{aligned} E_f[h(X)] &= \int h(x)f(x)dx \\ &= \int h(x)\frac{f(x)}{g(x)}g(x)dx \\ &= E_g[h(x)\frac{f(x)}{g(x)}] \end{aligned}$$

Therefore the original sampling algorithm by simulating $\{X_i\}_{i=1}^n$ from $f(x)$, and calculating $\frac{1}{n}\sum_{i=1}^n h(X_i)$ can be changed to:

1. Simulating $\{X_i\}_{i=1}^n$ from $g(x)$

2. Calculating $\frac{1}{n}\sum_{i=1}^n h(X_i)\frac{f(X_i)}{g(X_i)}$.

A variance reduction can be achieved by choosing an appropriate $g(x)$ to allow more weights on the more important simulations so that the variance of calculating the expectation is reduced.

$$Var(h(X_i)\frac{f(X_i)}{g(X_i)}) \leq Var(h(X))$$

The above expression has a smaller variance if the weight $\frac{f(x)}{g(x)}$is small, so a better approximation of $f(x)$ by $g(x)$ in the important region is essential.

- Metropolis-Hasting (MH) Algorithm

In general, it may be difficult to select a selection of envelope distribution or an approximate distribution. Markov chain Monte Carlo method provides a general algorithm to simulate random variables from an arbitrary posterior distribution. Suppose we are interested in sampling θ from a posterior distribution $p(\theta|x)$, the M-H algorithm is following:

1. Initiate the algorithm by setting the $\theta^{(0)}$;

2. At iteration t, choose a proposal density $q(\theta^*|\theta^{(t-1)})$,where parameters θ^* is a proposed parameter drawn from q based on the current $\theta^{(t-1)}$;

3. Compute the acceptance probability:

$$\alpha = \min\{1, \frac{p(\theta^*)/q(\theta^*|\theta^{(t-1)})}{p(\theta^{(t-1)})/q(\theta^{(t-1)}|\theta^*)}\}$$

4, Generate u from $U(0,1)$. If $u \leq \alpha$, accept θ^* as $\theta^{(t)}$; otherwise, reject θ^*,set $\theta^{(t)} = \theta^{(t-1)}$.Go to step 2 until the pre-specified total iterations T.

When $q(\theta^*|\theta^{(t-1)}) = q(\theta^{(t-1)}|\theta^*)$, i.e. the proposal distribution is symmetric, the calculation of acceptance probability is simplified as:

$$\alpha = \min\{1, \frac{p(\theta^*)}{p(\theta^{(t-1)})}\}$$

In this case, it is called Metropolis algorithm.

- Gibbs Sampler

When there are multiple parameters, it may not be possible to find an appropriate proposal density, and it may be very slow to reject or update the whole parameters vector at one time. Alternatively, the parameters vector can be grouped into several blocks to facilitate the sampling, which is called block structure M-H algorithm. Gibbs sampler is a special case of block structure M-H algorithm, where each group includes only one parameter. Assume the parameter vector $\theta = (\theta_1, \theta_2, ..., \theta_q)$, and it is easy to sample from the conditional distributions $p(\theta_i|\theta_{-i})$, then the Gibbs sampler is described in following steps:

1. Initialize $\theta^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, ..., \theta_q^{(0)})$.

2. At iteration t, the draw of $\theta^{(t)}$ is obtained by:

Draw $\theta_1^{(t)}$ from $p(\theta_1|\theta_2^{(t-1)}, \theta_3^{(t-1)}, ..., \theta_q^{(t-1)})$;

Draw $\theta_2^{(t)}$ from $p(\theta_2|\theta_1^{(t)}, \theta_3^{(t-1)}, ..., \theta_q^{(t-1)})$;

...

Draw $\theta_q^{(t)}$ from $p(\theta_q|\theta_1^{(t)}, \theta_2^{(t)}, ..., \theta_{q-1}^{(t)})$;

3. Repeat step 2 until the chain converges.

In practice, if it is difficult to sample directly from the conditional density $p(\theta_i|\theta_{-i})$, other algorithms such as rejection sampling and M-H algorithm can be incorporated with Gibbs sampler together. A typical application can be found for Bayesian estimation of stochastic volatility models (see Jacquier et al, 1994).

2.4 Empirical Bayes methods

Given that frequentist approaches are often too sensitive to changes in the data, and full Bayesian approaches are criticized for its substantial flexibility in specifying prior distributions, there is another branch of estimation methods which take a trade-off between the two approaches, called empirical Bayes methods. In a full Bayesian approach, the prior distribution is prespecified without looking into the data. However, for empirical Bayes method, the prior distribution is estimated from the data.

Suppose we have observations $y = (y_1, ..., y_n)$ from a distribution with density $f(y|\theta)$, where $\theta = (\theta_1, ..., \theta_k)$ is a vector of unknown parameters. In Bayesian statistics, a prior

distribution for θ is chosen as $\pi(\theta|\eta)$, where η is a vector of known hyperparameters. The posterior distribution of θ is calculated by:

$$\begin{aligned} p(\theta|y) &= \frac{f(y|\theta)\pi(\theta|\eta)}{\int f(y|\theta)\pi(\theta|\eta)d\theta} \\ &= \frac{f(y|\theta)\pi(\theta|\eta)}{m(y|\eta)} \end{aligned}$$

where $m(y|\eta)$ denotes the marginal distribution of y.

When f and π form a conjugate pair of of distributions, the marginal distribution $m(y|\eta)$ is available in closed form. Then, an empirical Bayes approach finds an estimate of η using the marginal distribution $m(y|\eta)$ such as using maximum likelihood estimate $\hat{\eta} = \arg\max m(y|\eta)$, and then plugs in $\hat{\eta}$ to the prior to compute the posterior distribution:

$$p(\theta|y) = \frac{f(y|\theta)\pi(\theta|\hat{\eta})}{m(y|\hat{\eta})}.$$

Example 2.2 *(Empirical Bayes estimators of mean parameter for normal distribution)*: Suppose we need to estimate the mean parameter for a sample from a normal distribution with unknown mean θ, but known variance σ^2.

$$f(y|\theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\theta)^2}{2\sigma^2}}$$

If a normal distribution with zero mean and variance as τ^2 are assigned to the prior distribution of θ :

$$\pi(\theta|\tau) = \frac{1}{\sqrt{2\pi}\tau} e^{-\frac{\theta^2}{2\tau^2}}$$

The posterior distribution of θ :

$$p(\theta|y) \propto f(y|\theta)\pi(\theta|\tau)$$

$$\theta|y \sim N(\frac{\tau^2}{\tau^2+\sigma^2}y, \frac{\sigma^2\tau^2}{\tau^2+\sigma^2})$$

and the marginal distribution $m(y|\tau) = \int f(y|\theta)\pi(\theta|\tau)d\theta$ can be obtained in closed-form:

$$\begin{aligned} m(y|\tau) &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\theta)^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi}\tau} e^{-\frac{\theta^2}{2\tau^2}} d\theta \\ &= \frac{1}{\sqrt{2\pi(\tau^2+\sigma^2)}} e^{-\frac{y^2}{2(\tau^2+\sigma^2)}} \end{aligned}$$

Given a sample $y = (y_1, ..., y_n)$, then the log-likelihood of the marginal distribution $m(y|\tau)$ is:

$$l(y_1, ..y_n; \tau) = -\sum_{i=1}^{n} \frac{y_i^2}{2(\tau^2+\sigma^2)} - \frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\tau^2+\sigma^2)$$

Taking derivative w.r.t. τ, and set it as 0, we have an MLE for $\hat{\tau}$:

$$\frac{1}{(\tau^2+\sigma^2)} = \frac{n}{\sum_{i=1}^{n} y_i^2}$$

In this case the posterior estimates of $\hat{\theta}$ will be $\left(1 - \frac{n\sigma^2}{\sum_{i=1}^n y_i^2}\right) y$.

Let's consider a method-of-moments estimator for $\hat{\tau}$. Since $\sum_{i=1}^n y_i^2/\sigma^2 \sim \chi^2(n)$,

$$E[\frac{1}{\sum_{i=1}^n y_i^2}] = \frac{\sigma^2}{n-2}$$

then a method-of-moments estimator for τ is:

$$\frac{1}{(\tau^2+\sigma^2)}\frac{\sigma^2}{n-2} = \frac{1}{\sum_{i=1}^n y_i^2}$$

Then the posterior mean of θ given y is:

$$\hat{\theta} = (1 - \frac{(n-2)}{\sum_{i=1}^n y_i^2})y$$

This is right the well known James-Stein (JS) estimator. However, this estimator has non-Bayesian interpretation by explicitly shrinking the MLE to certain estimate and reaches lower mean squared error. It was shown that the JS estimator has uniformly smaller mean squared error than the MLE (James and Stein, 1961). The shrinkage estimators have a broad usage in linear regression, prediction and covariance estimation. They are closely related to Bayesian estimators and penalized likelihood estimators.

Other typical examples of EB can also be found in linear regression (van Houwelingen, 2011), analysis of variance (ANOVA) and covariance estimation (Casella, 1992).

However, when the likelihood for the marginal distribution $m(y|\eta)$ is not sharp, it might not be appropriate using one proxy $\hat{\eta}$ for the prior distribution. A full Bayesian analysis (called Bayes empirical Bayes) augments the posterior distribution using a hyperprior distribution $h(\eta|\lambda)$:

$$\begin{aligned} p(\theta|y,\lambda) &= \frac{\int f(y|\theta)\pi(\theta|\eta)h(\eta|\lambda)d\eta}{\int\int f(y|\theta)\pi(\theta|\eta)h(\eta|\lambda)d\theta d\eta} \\ &= \int p(\theta|y,\eta)h(\eta|y,\lambda)d\eta \end{aligned}$$

The posterior is a mixture of conditional posterior $p(\theta|y,\eta)$ in Bayesian approach and the hyperprior $h(\eta|y,\lambda)$ updated by the data y.In contrast to the Bayesian approach, both EB and BEB use the observed data to estimate the prior information on η, and can be considered as a hybrid of frequentist and Bayesian inference. The EB approach has found successful applications in a broad area and has close ties to shrinkage estimators. However, the EB approach replaces the prior with a point estimate $\hat{\eta}$ and may ignore the uncertainty in estimating η, while the performance of BEB approach depends on the choice of hyperprior h. A review of the development of EB and BEB can be found in Carlin and Louis (2000).

2.5 Robust statistics methods

2.5.1 Evaluation of estimators

To evaluate the performance of an estimator or compare the performance of different estimators, it is necessary to define an evaluation criteria. In statistics, mean-squared error

(MSE), bias and variance are probably the most important quantities to be computed in evaluating an estimator. Each definition of these statistics are shown below:

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$$

$$Bias(\hat{\theta}) = E\hat{\theta} - \theta$$

$$Var(\hat{\theta}) = E[(\hat{\theta} - E\hat{\theta})^2]$$

They are related together in the equation below

$$MSE(\hat{\theta}) = Bias^2(\hat{\theta}) + Var(\hat{\theta}).$$

The equation can be easily proved since

$$E[(\hat{\theta} - \theta)^2] = E[(\hat{\theta} - E\hat{\theta} + E\hat{\theta} - \theta)^2] = E[(\hat{\theta} - E\hat{\theta})^2] + (E\hat{\theta} - \theta)^2.$$

2.5.2 Robust statistics indicators

In real world problems, the data may subject to collection errors and include outliers, for which the maximum likelihood estimate may become very sensitive and unstable. Therefore, robust estimation methods are very important. The typical example is that when estimating the location parameter for a sample coming from normal distribution but with outliers include, the mean is insensitive to outliers while using median is much better. There are also other robust estimators such as trimmed means or winsorized means.

In general, robust estimators include M-estimator (maximum likelihood type estimator), L-estimator (linear combinations of order statistics) and R-estimator (estimator based on rank transformation) (see Huber 1981), RM estimator (repeated median) (Siegel 1982), and LMS estimator (estimator using the least median of squares) (Rousseeuw, 1984).

In robust statistics world, an estimator is usually evaluated by the breakdown point and the influence function.

- Breakdown point

The breakdown point of an estimator is defined as the proportion of incorrect observations (i.e., arbitrarily large observations) an estimator can handle before making the estimator arbitrarily bad. For example, for a sample $\{X_i\}_{i=1}^n$ from $N(\mu, \sigma^2)$, the sample mean $\overline{X_n} = 1/n \sum_{i=1}^n X_i$ has a breakdown point of 0 because any arbitrarily large X_i can lead to arbitrarily large mean $\overline{X_n}$.However, the median has a breakdown point of 0.5, and the $x\%$ trimmed mean has breakdown point of $x\%$.

- Influence function

The influence function $IF(x; T, F)$ describes how an estimator $T(F)$ for distribution $F(y; \theta)$ behaves when the data follows another unspecified distribution at a particular value x. Mathematically, it is defined as following:

$$IF(x; T, F) = \lim_{\varepsilon \to 0+} \frac{T(t\Delta_x + (1-t)F) - T(F)}{t}$$

where Δ_x, a probability measure putting mass 1 to $\{x\}$:

$$\Delta_x(y) = \begin{cases} 0, & y < x \\ 1, & y \geq x \end{cases}$$

When F is the empirical distribution $\widehat{F}_n$, the IF is called empirical influence function (EIF). IF describes the marginal impact by arbitrarily distributed data at the specific value of x on the parameter estimates. In other words, IF is the functional derivative of the estimator with respect to the assumed distribution. IF is very useful in analyzing the behavior of the impact of data contamination on the estimator.

Hampel et al (1986) showed that IFs for M estimators with objective function $\rho(x,\theta)$ and distribution density function $f(x;\theta)$ has the following form:

$$IF(x;\theta) = \frac{\varphi_\theta}{\int \varphi_{\theta\theta} f(x;\theta)dx}$$

where $\varphi_\theta = \frac{\partial\rho(x,\theta)}{\partial\theta}$, and $\varphi_{\theta\theta} = \frac{\partial^2\rho(x,\theta)}{\partial\theta^2}$.

For MLE, a special class of M estimators,

$$\rho(x,\theta) = -\ln f(x;\theta)$$

$$\varphi_\theta = -\frac{1}{f(x;\theta)}\frac{\partial f(x;\theta)}{\partial\theta}$$

$$\varphi_{\theta\theta} = -\frac{1}{f(x;\theta)}\frac{\partial^2 f(x;\theta)}{\partial\theta^2} + \frac{1}{f(x;\theta)^2}\left(\frac{\partial f(x;\theta)}{\partial\theta}\right)^2$$

then, the IF can be written as:

$$IF(x;\theta) = \frac{\frac{1}{f(x;\theta)}\frac{\partial f(x;\theta)}{\partial\theta}}{\int \frac{1}{f(x;\theta)}\left(\left(\frac{\partial f(x;\theta)}{\partial\theta}\right)^2 - \frac{\partial^2 f(x;\theta)}{\partial\theta^2}f(x;\theta)\right)dx}$$

For multiple-parameter cases, the first-order derivatives are vectors, and the second-order derivatives form a Hessian matrix. The consequent matrix form of IF is:

$$IF(x;\theta) = A(\theta)^{-1}\varphi_\theta$$

where $A(\theta)_{ij} = -\int \frac{\partial\varphi_{\theta_i}}{\partial\theta_j} f(y;\theta)dy$, and $\varphi_\theta = (\varphi_{\theta_1},...,\varphi_{\theta_k})^T$.(Stefanski and Boos (2002)).

Opdyke and Cavallo (2012) derived the analytical form of IFs for truncated distributions for operational loss severities, and used them to demonstrate why the MLE often produces unrealistic estimates when applied to fitting the operational losses with truncation and possible data contamination issues.

2.6 Penalized likelihood estimators

Penalized likelihood approach is another class of error reduction methods by adding a penalty term(e.g., roughness of estimates) onto the likelihood to reduce the variability of estimators at the price of a small bias. A penalized likelihood estimator is an M-estimator, which is also consistent and asymptotically normal distributed under certain regularity conditions (van der Varart (1998)). It has been widely used as a robust method in linear regression (Tibshirani, 1996; Breiman, 1995). It has also been used as a nonparametric method in density estimation by imposing smoothness of the density or other function (Eggermont and LaRiccia, 2001). Penalized likelihood estimators also have close relation to shrinkage estimators and Bayesian estimators (Good and Gaskins, 1971; van Houwelingen, 2001).

Example 2.3 *(Linear regression)* Consider the linear regression problem:

$$y = X\beta + \varepsilon$$

where y is a $n \times 1$ observed vector of dependent variable, $X = (X_1, ..., X_k)$ is a $(n \times k)$ matrix are the n observations of k factors, and we wish to estimate the factor loadings $\beta_{k\times 1}$. Without loss of generality, the mean of y and X_i are assumed to be zero, so that ε has a zero mean.

When ε is assumed to be i.i.d. Gaussian noise, the maximum likelihood estimate of β is equivalent to the ordinary least-squares estimate:

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T y$$

Maximizing likelihood is equivalent to minimizing the sum of squares estimators. However, when there is multicollinearity (e.g., $X^T X$ close to singular), the estimate $\hat{\beta}$ may behave wildly. A natural idea is penalizing the variability of β. Let P be a non-negative definite matrix, $\beta^T P\beta$ could be a reasonable penalty term. The penalized likelihood problem is:

$$\min_{\beta} \ ||y - X\beta||^2 + \lambda\beta^T P\beta$$

Thus, the penalized likelihood estimate can be obtained as:

$$\hat{\beta}_P = (X^T X + \lambda P)^{-1} X^T y$$

When P is identity, this yields the ridge regression, introduced by Hoerl and Kennard (1970).

This can also be viewed as a result from using shrinkage covariance estimates instead of sample covariance.

$$\hat{\beta}_P = (nS + \lambda P)^{-1} X^T y$$

where $S = X^T X/n$ is the sample covariance.

Example 2.4 *(Mean-variance portfolio selection)* Consider the mean-variance portfolio selection problem:

$$\max_{w} w^T\mu - \frac{\lambda}{2} w^T\Sigma w$$

where $w_{n\times 1}$ are the weight vector, $\Sigma_{n\times n}$ is the covariance matrix, $\mu_{n\times 1}$ is the expected return vector, λ is a Lagrangian multiplier that adjusts the return by risk.

Consider the simplest case without any constraints on w, the optimal allocation is

$$\hat{w} = \frac{1}{\lambda}\Sigma^{-1}\mu$$

If one wishes to penalize the size of weight in each instrument, a similar penalty function as in the linear regression example can be defined as $\frac{\gamma}{2} w^T P w$. Then the penalized mean-variance portfolio selection problem is:

$$\max_{w} w^T\mu - \frac{\lambda}{2} w^T\Sigma w - \frac{\gamma}{2} w^T P w$$

It is easy to see that the optimal weight is

$$\hat{w}_P = (\lambda\Sigma + \gamma P)^{-1}\mu$$

Again, this is closely related to use a shrinkage covariance estimate. It also has another interpretation from Bayesian or robust portfolio optimization point of view. The penalty term $\frac{\gamma}{2}w^T Pw$ can be viewed as the uncertainty in expected return μ (see Fabozzi et al, 2007).

From these examples, we see that in many cases penalized likelihood estimates are closely related to Bayesian inference and other types of error-reduction techniques. In Bayesian framework, penalized likelihood function corresponds to a posterior density while the penalized maximum likelihood estimate is a maximum a posteriori estimate. For example, the MAP estimate using Jeffreys' prior is obtained by penalizing the likelihood function with the square root of determinant of Fisher information matrix. Cope (2011) developed a penalized likelihood estimator for truncated distribution estimation, where the likelihood function was penalized by a linear term that tilts the likelihood surface to reduce the variability of estimates.

2.7 Model Mis-specification

In practice, the true model is often unknown. The assumed model may not be the same as the true model. In this case, the model is called mis-specified.

Suppose $x_1, ..., x_n$ are an i.i.d sample from an unknown distribution $g(x)$, but our assumed distribution for x_i is $f(x;\theta)$. Then maximizing the likelihood is equivalent to

$$\max_{\theta} \frac{1}{n}\sum_{i=1}^{n} \log f(x_i;\theta)$$

which in large samples is equivalent to

$$\max_{\theta} E_g \left[\log f(X;\theta)\right]$$

This is equivalent to minimizing the Kullback-Leibler divergence $D(g||f)$:

$$\min_{\theta} E_g \left[\log \frac{g(X;\theta)}{f(X;\theta)}\right]$$

Definition 2.1 *(Kullback-Leibler divergence)* Suppose $f(x)$ is an assumed density, and $g(x)$ is the true density, then Kullback-Leibler (KL) divergence $D(g||f)$ is defined as

$$\begin{aligned} D(g||f) &= E_g \left[\log \frac{g(X)}{f(X)}\right] \\ &= \int g(x) \log \frac{g(x)}{f(x)} dx \end{aligned}$$

An important property of KL divergence is that it is always nonnegative.

Proposition 2.1 *(Information Inequality)* If $f(x)$ and $g(x)$ are two densities, then

$$E_g \left[\log \frac{g(X)}{f(X)}\right] \geq 0$$

The proof is easily obtained by using Jensen's inequality:

$$E(-\log Y) \geq -\log EY$$

$$\begin{aligned} E_g\left[-\log\frac{f(X)}{g(X)}\right] &\geq -\log E_g\left[\frac{f(X)}{g(X)}\right] \\ &= -\log\int f(x)dx = 0 \end{aligned}$$

The information inequality implies that the log-likelihood of the "true" model tends to be larger than the log-likelihood of a "wrong" model:

$$E_g\left[\log g(X)\right] \geq E_g\left[\log f(X)\right].$$

This in fact provides an intuition for the consistency of MLE.

In practice, the true model may even not exist. Even if it exists, the observed sample is just a random sample drawn from the true model. By assuming a true model and applying MLE, what we are actually doing is minimizing the KL divergence from the assumed model f to the empirical model $\hat{f}_n = \frac{1}{n}\sum_{i=1}^n I_{\{x=x_i\}}$:

$$\min D(\hat{f}_n||f) = \int \frac{1}{n}\sum_{i=1}^n I_{\{x=x_i\}} \log\frac{\frac{1}{n}\sum_{i=1}^n I_{\{x=x_i\}}}{f(x)}dx$$

$$\Leftrightarrow \max \frac{1}{n}\sum_{i=1}^n \log f(x_i)$$

This concept should be always kept in mind that maximizing likelihood is in fact minimizing the KL divergence from assumed model to the empirical data. If the assumed model family could not approximate the true model well, maximum likelihood method will result in biased and inconsistent estimates. Given that in practice the true model is often unknown, there is always some trade-off between the model-misspecification and over-fitting. A too complex model faces difficulty in parameter estimation and over-fits the in-sample data but has poor performance in out-of-sample forecasting. A restricted model may face risk of model misspecification and be too narrow or unrealistic to be useful in practice. What we seek is a flexible enough model to approximate the truth while in the mean time we need to consider the brevity of the model.

3 Operational Risk Modelling

3.1 Operational Risk

Operational risk is defined as the risk of direct or indirect loss resulting from inadequate or failed internal processes, people and systems or from external events, according to the Basel II Accord (2006). It includes legal risk but excludes strategic and reputational risk. Along with market risk and credit risk, it has become a main type of risk for financial institutions. Historically, the quantification of operational risk were not possible due to lack of data. However, in recent years, major banks have started quantifying the operational risk under the Basel II's guidance and regulations because there are more loss data available since the initiation of collecting losses from 2002 or 2003.

According to Basel II accord, the losses are categorized into seven event types: (1) Internal Fraud; (2) External Fraud; (3) Employment Practices and Workplace Safety; (4) Clients, Products, & Business Practice; (5) Damage to Physical Assets; (6) Business Disruption & Systems Failures; (7) Execution, Delivery, & Process Management. The definition of each loss event type and detailed loss event type classification including level 2 categories can be found in BIS (2006) page 305.

Historically, under the so-called standardized approach which calculate the capital charge based on a weighted average of gross income, the banks' activities are divided into eight business lines: (1) Corporate Finance; (2) Trading & Sales; (3) Retail Banking; (4) Commercial Banking; (5) Payment & Settlement; (6) Agency Services; (7) Asset Management; (8) Retail Brokerage. Each business line was assigned a beta factor as the weight of the gross income in total capital charge (see BIS (2006) page 147). Therefore, the standard Basel loss data categories include seven event types and eight business lines. The banks are required to follow the standard in categorizing data or be able to map their losses into these categories.

Unlike market and credit risk modelling which have been well developed for decades, the quantification of operational risk has a relative short history and there has been much more flexibility in modelling since it is still underdeveloped and actively growing. Since the Basel II Accord, major banks have started to develop their operational risk modelling system following the regulations and the guidance by Basel II. A widely adopted framework by many banks in the world is the so-called advanced measurement approaches (AMA), as more sophisticated when compared to the historical simple basic indicators approach and standardized approach. Specifically, banks using the AMA approach have to: (1) follow the scope of operational risk and the loss event types defined by the Basel Committee (BIS, 2006); (2) calculate regulatory capital requirement as the sum of expected loss (EL) and unexpected loss (UL); (3) have sufficient granularity to capture the major drivers of operational risk affecting the shape of the tail of the loss estimates; (4) aggregate the operational risk estimates across unit of measures using summation or correlation based aggregation techniques with a sound determination of correlation; (5) use elements including internal data, relevant external data, scenario analysis and factors reflecting the business environment and internal control systems; (6) have a credible, transparent, well-documented and verifiable approach for weighting the different elements in the overall operational risk measurement system. Detailed description about each data element is in BIS (2006) page 152 - 154.

In practical operational risk management, all the four data elements need to be considered. However, in our research, we mainly consider the internal loss data and external loss data only, which are the most important elements. The research in scenario analysis

(SA) for operational risk is a new topic as well and some recently proposed methodologies that have been used in practice are Dutta ad Babbel (2013) and Cope (2012). The last data component — business environment and internal control factors (BEICF) — usually include data about audit ratings, near miss event counts, and opening & closed issues. These factors reflect the management effort and performance, and are used to adjust the final capital estimates. We do not discuss the topics related to SA and BEICF in this dissertation but focus on internal loss data and external loss data modelling.

3.2 Loss distribution approach

Among the advanced measurement approaches, Loss distribution approach (LDA) is the most standard approach for estimating the distribution of operational losses. It is a statistical method popular in actuarial sciences for computing the aggregated loss distributions, and was mentioned in Annex 6 of an early-stage supporting document for operational risk management by Basel II (BIS, 2001). Under the LDA, the bank estimates the distribution of losses over the next one year period for each unit of measure (UoM) by aggregating the loss frequency distribution and the loss severity distribution. The risk measure used for capital charge is obtained by calculating the 99.9% quantile of the aggregated loss distributions. At last, the total capital charge is the sum of the capital charge for each UoM. During the past over 10 years development of both the data collection and modelling research, the LDA has become a prevalent approach in major banks in Europe and North America.

The assumptions of LDA include:

(1) The inter-arrival time of the loss events in the same cell are independent;

(2) The loss severity of each loss event are independent and identically distributed;

(3) The aggregated loss distribution for different cells are perfectly dependent to each other.

Suppose the loss frequency n during a one-year period follows a distribution $P(n = N) = p(n)$, and the loss severity distribution of each single event is $F(X)$, then the aggregated loss distribution is

$$\begin{aligned} F_L(l) &= P(L < l) \\ &= \sum_{n=0}^{\infty} p(n) P(\sum_{i=0}^{n} X_i < l) \\ &= \sum_{n=0}^{\infty} p(n) F^{n*}(l) \end{aligned}$$

where $F^{n*}(\cdot)$ is the n-fold convolution of F with itself.

Then the capital charge is obtained by calculating the α- quantile of the aggregated loss distribution:

$$C_\alpha = \inf\{l | F_L(l) \geq \alpha\}$$

where α is usually specified as 99.9% to safeguard the unexpected tail losses.

Suppose there are in total K UoMs (or cells, e.g., if there are eight business lines and seven event types, then there are 56 cells), according to the standard LDA, the total capital charge for the bank is the sum of the capital charge of each cell:

$$C_\alpha = \sum_{i=1}^{K} C_\alpha^{(i)}$$

where $C_\alpha^{(i)}$ is the capital charge for the i-th cell.

Theoretically, this in fact assumes that the losses from different cells are perfectly dependent, which would greatly overestimate the total capital charge for the whole bank. Basel II also allows bank to use correlation-based techniques to aggregate the capital charges across UoMs. In later sections, we will discuss modified LDA framework for operational risk measurement by considering the inter-cell dependence with copula models.

3.3 Loss frequency modelling

The most popular way of modeling the loss frequency is by using Poisson distribution due to its broad use in counting processes. The Poisson distribution can be derived in the following definition.

Definition 3.1 *(Poisson Process)* A counting process $\{N(t), t \geq 0\}$ is a Poisson process with rate λ if:

(1) $N(0) = 0$;

(2) (Stationary independent increments) $N(t+\Delta t) - N(t)$ is stationary and independent with t, but only depends on Δt;

(3) $P(N(\Delta t) = 1) = \lambda \Delta t + o(\Delta t)$;

(4) $P(N(\Delta t) \geq 2) = o(\Delta t)$.

Given the above definition, the probability mass function of $N(t)$ can be derived as:

$$P\{N(t) - N(s) = k\} = \frac{(\lambda(t-s))^k}{k!} e^{-\lambda(t-s)}, k = 0, 1, 2, ...$$

The Poisson distribution has following properties: if $X \sim P(\lambda), E(X) = Var(X) = \lambda$. Because of this property, the Poisson distribution has been criticized in practice due to the observed over-dispersion effect of the loss frequency data.

Negative binomial distribution is a generalization of the Poisson distribution, in which the intensity rate λ is not a constant but follows a gamma distribution $\Gamma(r, p/(1-p))$. The negative binomial distribution $NB(r, p)$ is usually used to describe the distribution of the number of successes in a sequence of Bernoulli trials with successful probability p before a specified number of failures (denoted r) occur. The probability mass function of a negative binomial distributed random variable $X \sim NB(r, p)$

$$P(X = k) = \begin{pmatrix} r+k-1 \\ k \end{pmatrix} p^k (1-p)^r, k = 0, 1, ...$$

The mean and variance of a negative binomial distributed random variable $X \sim NB(r, p)$ are $\frac{pr}{(1-p)}$ and $\frac{pr}{(1-p)^2}$.With $p > 0$,the variance is always greater than the mean. Therefore, in modelling the operational losses data, some researchers found it superior to the Poisson distribution on goodness-of-fit tests. Moscadelli (2004) examines the data collected by the Risk Management Group of the Basel Committee in its June 2002 Operational Risk Loss Data Collection Exercise (LDCE), with over 47,000 observations from 89 participated banks. They found that the negative binomial distribution provides a better fit than the Poisson distribution.

Since the collection of operational losses have been developing over time since the Basel II, the intensity rate of losses may exhibit a time-varying effect. Therefore, a nonhomogeneous Poisson process might be better.

Definition 3.2 *(Nonhomogeneous Poisson process)* A counting process $\{N(t), t \geq 0\}$ is called a nonhomogeneous Poisson process with rate $\lambda(t)$ if

(1) $N(0) = 0$;
(2) $N(t)$ has independent increments;
(3) $N(t) - N(s)$ is a Poisson process with rate $\int_s^t \lambda(\tau)d\tau$.

Chernobai and Rachev (2004) applied a nonhomogeneous Poisson process with two cdf-like cumulative intensity rate $\int_0^t \lambda(\tau)d\tau$ on a public operational loss data set from 1950 to 2002. Because of the long time period used, the chosen intensity rate fits much better than the linear cumulative intensity rate λt.

3.3.1 Over-dispersion and zero-inflation

Even though the goal of modelling loss frequency is to estimate or forecast the loss frequency distribution for the next one-year period, when calculating the historical loss frequency sample it usually requires a shorter unit period based on which the loss frequency are aggregated. For example, if we aggregate the loss frequency annually, because the historical data only includes few years data, we then will only have a very small sample. If we aggregate the loss frequency daily, then for most event types, there might be a serious problem of zero-inflation. Empirically, the loss frequency are usually calculated monthly or quarterly to achieve a trade-off between arriving at a enough sample and avoiding the zero-inflation problem.

In practice, there are several ways to test the over-dispersion effect. Let $X = (X_1, ..., X_n)^T$ be a sample of count data from unknown distribution with mean λ, and we need to test if it is over-dispersed. The null hypothesis and alternative hypothesis are:

$$\begin{aligned} H_0 &: X \sim Poisson(\lambda) \\ H_1 &: var(X) > \lambda \end{aligned}$$

One test statistic for over-dispersion suggested by Bohning (1994) is:

$$T = \frac{\frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2 - \overline{X}}{\sqrt{\frac{2}{n-1}\overline{X}}}$$

where $\overline{X} = 1/n \sum_{i=1}^n X_i$. Under the null hypothesis, the expected value of $\sum_{i=1}^n (X_i - \overline{X})^2/(n-1) - \overline{X}$ is zero, and variance is $2\lambda/(n-1)$.

There are several other test for over-dispersion, which can be found in Baksh, Bohning and Lerdsuwansri (2011), Rao and Chakravarthi (1956), and Best et al. (2007).

The zero-inflation might affect the test of over-dispersion. Since the operational losses for some event types are very rare especially when choosing a short aggregation period such as a month, or a quarter, it is usual that the observed frequency has a lot of zero values. For example, suppose the annual aggregated loss frequency does not fail the over-dispersion test, however, if we aggregate the loss frequency every month, or every day, then it is likely that there will be a lot of zero values in the sample, which would likely reject the Poisson distribution assumption. This phenomenon is also very typical in analysis of high frequency data.

In this case, a zero-inflated Poisson model may be better to accommodate the zeros that are not generated by Poisson distribution, but just structural zeros due to sampling frequency. The two components of the zero-inflated Poisson (ZIP) model (Lambert, 1992) are:

$$P(N = 0) = \pi_0 + (1 - \pi_0)e^{-\lambda}$$

$$P(N = n) = (1 - \pi_0)\frac{\lambda^n e^{-\lambda}}{n!}, n \geq 1$$

Greene (1994) later considered the zero-inflated negative binomial (ZINB) model.

3.4 Loss severity modelling

Lack of loss data is the main problem restricting the development of quantitative operational risk analysis. Operational risk capital is reserved for extreme losses, which occur very rarely. This poses difficulty in collecting the loss data especially the tail events which significantly impact fitting the loss distribution. On the other hand, to accurately estimate the probability and size of these extreme events requires a sufficient sample of loss data. Under LDA, the loss frequency distribution and loss severity distribution are estimated separately, and then aggregated. Compared to the loss frequency modelling, the loss severity distribution is much more difficult to fit and has more practical issues.

3.4.1 Operational loss distributions

Different from market risk and credit risk, a typical feature of operational losses is that the distribution has much heavier tails. The magnitude of a single loss usually varies from few hundreds to billion dollars. In addition, these tail events also increase the skewness of the distribution. Therefore, to find a good distribution with satisfying fit for operational losses is more difficult than for market risk. In this section, we discuss the usual loss distributions that have been suggested in the academic world and used prevalently in largest banks for operational losses.

In probability theory, heavy-tailed distributions are probability distributions whose tails are not exponentially bounded (Asmussen and Soren (2003)). In practice, we usually care about the the right tail of the distribution.

Definition 3.3 *(Heavy-tail)* The distribution of a random variable X with distribution F is said to have a heavy right tail if

$$\lim_{x\to\infty} e^{\lambda x} P(X > x) = \lim_{x\to\infty} e^{\lambda x} \bar{F}(x) = \infty$$

This is equivalent to the statement that the moment generating function of F, $M_F(t)$, is infinite for all $t > 0$.

In finance, there are frequently different terms that are actually sub-classes of heavy-tailed distributions: the fat-tailed distributions, the long-tailed distributions and the subexponential distributions. The fat-tailed belongs to subexponential class, subexponential belongs to long-tailed, and long-tailed belongs to heavy-tailed. Their mathematical definitions are described as follows.

Definition 3.4 *(Long-tail)* The distribution of a random variable X is said to have a long tail if

$$\lim_{x\to\infty} P(X > x + t | X > x) = 1,$$

or equivalently,

$$\bar{F}(x + t) \sim \bar{F}(x)$$

Definition 3.5 *(Subexponential distributions)* The distribution of a random variable X is said to be subexponential if for n independent random variables $X_1, ..., X_n$ with the same distribution F :

$$P(X_1 + X_2 + ... + X_n > x) \sim P(\max(X_1, ..., X_n) > x), \text{ as } x \to \infty$$

In convolution notations,

$$\overline{F^{*n}(x)} \sim n\overline{F(x)}, \text{ as } x \to \infty.$$

Definition 3.6 *(Fat-tail)* The distribution of a random variable X is said to have a fat tail if

$$\lim_{x\to\infty} P(X > x) = cx^{-\alpha}, \alpha > 0$$

where $c \in R$ is finite. Or, using the density function, if

$$\lim_{x\to\infty} f_X(x) = cx^{-(1+\alpha)}, \alpha > 0.$$

Usually, the "fat tail" distributions are for $0 < \alpha < 2$ (i.e., with infinite variance).

In the following, we summarize the usual candidate distributions that have been used for operational loss severity.

- Lognormal distribution $LN(\mu, \sigma^2)$

  pdf:

  $$f(x) = \frac{1}{\sqrt{2\pi}\sigma x} \exp\left\{-\frac{(\ln x - \mu)^2}{2\sigma^2}\right\}, x > 0, \sigma > 0$$

  mean: $e^{\mu+\sigma^2/2}$

  variance: $\left(e^{\sigma^2} - 1\right) e^{2\mu+\sigma^2}$

- Gamma distribution $\Gamma(\kappa, \theta)$

  pdf:

  $$f(x) = \frac{1}{\theta^\kappa \Gamma(\kappa)} x^{\kappa-1} \exp\left(-\frac{x}{\theta}\right), x > 0, \kappa > 0, \theta > 0$$

  mean: $\kappa\theta$

  variance: $\kappa\theta^2$

- Loggamma distribution $LG(\kappa, \theta)$

  pdf:

  $$f(x) = \frac{1}{\theta^\kappa \Gamma(\kappa)} \frac{(\ln x)^{\kappa-1}}{x} \exp\left(-\frac{\ln x}{\theta}\right), x > 1, \kappa > 0, \theta > 0$$

  mean: $(1-\theta)^{-\kappa}$,if $\theta < 1$

  variance: $(1-2\theta)^{-\kappa} - (1-\theta)^{-2\kappa}$, if $\theta < 1/2$.

- Weibull $Weib(\kappa, \lambda)$

  pdf:

  $$f(x) = \frac{\kappa}{\lambda}\left(\frac{x}{\lambda}\right)^{\kappa-1} \exp\left(-\left(\frac{x}{\lambda}\right)^\kappa\right), x > 0, \kappa > 0, \lambda > 0$$

  mean: $\lambda\Gamma(1 + 1/\kappa)$

  variance: $\lambda^2(\Gamma(1 + 2/\kappa) - \Gamma(1 + 1/\kappa)^2)$

- Logweibull $LW(\kappa, \lambda)$

  pdf:

$$f(x) = \frac{\kappa}{\lambda^\kappa} \frac{(\ln x)^{\kappa-1}}{x} \exp\left\{-\left(\frac{\ln x}{\lambda}\right)^\kappa\right\}, x > 1, \kappa > 0, \lambda > 0$$

- Loglogistic $LL(\alpha, \beta)$

  pdf:

$$f(x) = \frac{\left(\frac{\beta}{\alpha}\right)\left(\frac{x}{\alpha}\right)^{\beta-1}}{\left(1+\left(\frac{x}{\alpha}\right)^\beta\right)^2}, x > 0, \alpha > 0, \beta > 0$$

  The k-th moment exists when $k < \beta$:

$$E[X^k] = \alpha^k Beta(1 - k/\beta, 1 + k/\beta)$$

- Generalized Pareto distribution $GPD(\xi, \mu, \sigma)$

  pdf:

$$f(x) = \frac{1}{\sigma}\left(1 + \frac{\xi(x-\mu)}{\sigma}\right)^{-\frac{1}{\xi}-1}, \xi > 0, x \geq \mu, \sigma > 0$$

  mean: $\mu + \frac{\sigma}{1-\xi}$, if $\xi < 1$.

  variance: $\frac{\sigma^2}{(1-\xi)^2(1-2\xi)}$, if $\xi < 1/2$.

There are many other distributions such as Burr III, Burr XII, Lomax (Pareto Type II), Pearson VI, Generalized Beta Prime, Generalized Extreme Value distributions that have been used for loss distributions. There are also operational risk research using other heavy-tailed distributions or their variants such as alpha-Stable distribution, g-and-h distribution, and mixture distributions.

- Stable distribution $S_\alpha(\beta, \sigma, \mu)$

A random variable X is said to follow a stable distribution $S_\alpha(\beta, \sigma, \mu)$ if its characteristic function is

$$E(e^{itX}) = \begin{cases} \exp\left[it\mu - |\sigma t|^\alpha \left(1 - i\beta \text{sgn}(t) \tan(\pi\alpha/2)\right)\right], \alpha \neq 1 \\ \exp\left[it\mu - |\sigma t| \left(1 - i\beta \frac{2}{\pi} \text{sgn}(t) \ln|t|\right)\right], \alpha = 1 \end{cases}$$

where α is the index of stability ($0 < \alpha \leq 2$), β is the skewness parameter ($-1 \leq \beta \leq 1$), σ is the scale parameter ($\sigma > 0$), and μ is the location parameter ($\mu \in R$). When $\beta, \mu = 0$, it is called symmetric αStable distribution ($S\alpha S$). Extensive analysis of alpha-stable distributions and their properties can be found in Samorodnitsky and Taqqu (1994) and Rachev and Mittnik (2000). Introduction about the estimation methods and their application in operational risk can be found in Chernobai et al (2007) and references therein.

An introduction about the g-and-h distribution and its use in operational risk can be found in Degen, Embrechts and Lambrigger (2007). Since the density of g-and-h distribution is generally non available, MLE is in general not applicable as well.

3.4.2 Extreme Value distributions

Since for operational risk losses, the characteristics in the tail and body of the loss distribution are often quite different, a single distribution may not fit well for them at the same time. Therefore, a piecewise or mixture distribution is often used where a lighter-tailed distribution such as log-normal usually has a good fit for the body while a Generalized Pareto distribution (GPD) is better for the tail. The justification of using GPD for the tails comes from extreme value theory (EVT), which is used to model the probability of extreme events.

Theorem 3.1 *(Fisher-Tippett-Gnedenko)* Suppose $(X_1, X_2, ..., X_n)$ be a sequence of i.i.d random variables, and $M_n = \max\{X_1, ..., X_n\}$. If there exists (a_n, b_n) such that each $a_n > 0$ and $\lim_{n\to\infty} P\left(\frac{M_n - b_n}{a_n} \leq x\right) = F(x)$, where $F(x)$ is a non degenerate distribution function, then $F(x)$ belongs to one of the three types family:

1. Gumbel: $F(x) = \exp(-e^{-x}), x \in R$
2. Frechet: $F(x) = \begin{cases} 0 & x \leq 0 \\ \exp(-x^{-\alpha}) & x > 0 \end{cases}$
3. Weibull: $F(x) = \begin{cases} \exp(-(-x)^{\alpha}) & x \leq 0 \\ 1 & x > 0 \end{cases}$

The three types can be grouped into the generalized extreme value (GEV) distribution:

$$F(x;\xi) = \begin{cases} \exp\left(-\left(1+\xi x\right)^{-\frac{1}{\xi}}\right) & \xi \leq 0 \\ \exp\left(-e^{-x}\right) & \xi = 0 \end{cases}$$

where $1 + \xi x > 0$. When $\xi = 0$, it's Gumbel type; when $\xi > 0$, it's Frechet type; when $\xi < 0$, it's Weibull type distribution. A three parameter distribution can be defined as $F(x;\xi,\mu,\sigma) = F((x-\mu)/\sigma;\xi)$, where μ is a location parameter and $\sigma > 0$ is a scale parameter.

Consider the distribution of X conditionally on exceeding some high threshold u

$$F_u(y) = P(X - u \leq y | X > u) = \frac{F(u+y) - F(u)}{1 - F(u)},$$

the following theorem shows that for a large u, the distribution F_u can be approximated by a generalized Pareto distribution.

Theorem 3.2 *(Pickands-Balkema-de Haan)* Let $(X_1, X_2, ...)$ be a sequence of i.i.d r.v.'s, and let F_u be their conditional excess distribution function, then for a large class of underlying distribution functions F, and large u, F_u is well approximated by the generalized Pareto distribution:

$$F_u(y) \to G_{\xi,\sigma}(y), \text{ as } u \to \infty$$

where $G_{\xi,\sigma}(y) = 1 - (1 + \xi y/\sigma)^{-1/\xi}$,if $\xi \neq 0$ and $G_{\xi,\sigma}(y) = 1 - e^{-y/\sigma}$, if $\xi = 0$.

This provides a theoretical reason why GPD usually approximates the tails well.

3.4.3 Selection of tail threshold

There are several methods to diagnose whether the EVT can be applied and where is the tail threshold.

One of the methods is using the mean excess function

$$e(u) = E(X - u|X > u).$$

For GPD distribution, the mean excess function is a linear function of u :

$$e(u) = \frac{\sigma + \xi u}{1 - \xi}.$$

Another approach is using the Hill estimator, which is an estimator of the tail index.

Definition 3.7 *(Tail index)* F is said to have a regularly varying right tail of index $\alpha > 0$ if

$$\lim_{x\to\infty} \frac{\bar{F}(tx)}{\bar{F}(x)} = t^{-\alpha}$$

for all $t > 0$. The smaller is α, the heavier is the tail.

Hill (1975) introduces an estimator of the tail index. Assume $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ are order statistics of a positive sample (or from the positive part of the sample) $X_1, ..., X_n$. The Hill estimator based on k upper order statistics is defined as:

$$H_{k,n} = \frac{1}{k} \sum_{i=0}^{k-1} \left(\log X_{(n-i),n} - \log X_{(n-k),n} \right)$$

If the sample is from a distribution with a regularly varying right tail of index α, then under certain conditions, the Hill estimator converges to $1/\alpha$ in probability (Mason, 1982). In practice, when Hill estimator is applied on a finite sample, the choice of an appropriate k is a problem. A typical way to choose k is by plotting the Hill estimator for a range of k, and then choose one around which the estimator is stable.

Nguyen and Samorodnitsky (2011) introduced a sequential statistical testing method to select the tail threshold. The idea is to sequentially testing the samples $\{\log \frac{X_{(n-i)}}{X_{(n-k)}} : i = 0, 1, ...k-1\}$ for the null hypothesis of exponential distribution. The choice of k is the largest number that does not rejects the null hypothesis of exponentiality.

A moment statistic to test for expoentiality is:

$$Q_{k,n} = \frac{\sqrt{k}}{2} \left(\frac{\frac{1}{k}\Sigma_{i=0}^{k-1} \left(\log \frac{X_{(n-i),n}}{X_{(n-k),n}} \right)^2}{\left(\frac{1}{k}\Sigma_{i=0}^{k-1} \log \frac{X_{(n-i),n}}{X_{(n-k),n}} \right)^2} - 2 \right).$$

It asymptotically follows a standard normal distribution under null hypothesis assumption (Dahiya and Gurland (1972)). Then under confidence level α, then the k can be chosen as:

$$\hat{k}_n := \sup \{k : 1 \leq k \leq n, |Q_k| < Z_\alpha\}$$

However, the test statistic $Q_{k,n}$ grows as the sample size n increases, to take into this account, the critical value has to increase with the sample size as well. On the other hand, one has to avoid choosing very large k. Nguyen and Samorodnitsky considered to multiply the critical value by $\sqrt{\frac{\theta_n}{k}}$:

$$\hat{k}_n := \sup \left\{ k : 1 \leq k \leq n, |Q_k| < Z_\alpha \sqrt{\frac{\theta_n}{k}} \right\}$$

In their paper, a good suggestion for θ_n is $(\log n)^2$.

In practice, since the characteristics of operational loss distribution in the body and tail are usually very different, a single distribution might be unable to model well both of them or such a distribution is hard to find. Therefore, it is natural to consider two distributions separately for the body and the tail. Suppose the tail threshold is H, the distribution for the body is $F_1(x) : x \leq H$, and the distribution for the tails is $F_2(x) : x > H$, the total distribution is:

$$F(x) = \begin{cases} F_1(x) & x \leq H \\ F_1(H) + (1 - F_1(H))F_2(x) & x > H \end{cases}$$

For example, in operational loss modelling, the lognormal distribution $F_{LN}(x; \mu, \sigma^2)$ is usually good for the body, but may underestimate the tail. Therefore, a GPD distribution $F_{GP}(x; \xi, H, \beta)$ may be applied on the losses greater than the tail threshold H.

3.5 Reporting threshold problem

The loss distribution approach provides a way to estimate the aggregated loss distribution by modelling the loss frequency and loss severity separately. However, in the step of data collection of operational losses, for practical reasons, small losses below some specified thresholds are usually not collected, or just summarized as group records. In this case, the individual losses below the threshold are not available for severity distribution modelling, thus samples are left-truncated. This creates the main difficulty for operational loss distribution fitting. Typically, there are several approaches in fitting the loss severity distribution.

- Naive approach

Under this approach, the observed data are treated as complete and an unconditional distribution is directly fitted to the loss severity sample, and the observed loss frequency is treated as full loss frequency.

However, this approach neglects the left-truncated feature of the data and is clearly misspecified. For small truncation thresholds, this might be a minor issue. However, in practice, the truncation threshold is not negligible, a clearly left-truncated profile can be observed. Fitting an unconditional distribution clearly misspecifies both the loss frequency and loss severity distribution. Chernobai et al (2006) analyzed the bias of using the naive approach in estimating loss severity distribution's parameters and the impact on the expected loss and unexpected loss using the typical Poisson-Lognormal example.

- Truncated (or conditional) approach

Under this approach, a conditional distribution is fitted to the observed loss sample where the distribution function and density function are following:

$$F_u(x) = P(X < x | X \geq u) = \frac{F(x) - F(u)}{1 - F(u)} I_{\{x \geq u\}}$$

$$f_u(x) = \frac{f(x)}{1 - F(u)} I_{\{x \geq u\}}$$

This approach is often used when no prior information is available for losses below the threshold, and it relies on following assumptions:

(1) Losses below and above the threshold follows the same distribution.

(2) Loss frequency is independent of the loss severity. This is also an assumption of LDA.

After the parameters for the conditional distribution is estimated, the full distribution with the same parameters are used for the loss severity, and then the observed loss frequency is adjusted by the inferred truncated proportion to get the full frequency. For example, when the observed loss frequency follows a Poisson process with the observed intensity rate λ_T, the full loss frequency is adjusted as:

$$\hat{\lambda} = \frac{\lambda_T}{1 - F(u)}$$

- Shifted approach

Under this approach, the losses over the threshold are treated as complete data and all losses are assumed to be above the threshold. A distribution $f(y)$ is fitted to the shifted losses $Y = X - u$, which implies that:

$$f(X) = f(Y + u)$$

This approach circumvents the difficulty in estimating the parameters for truncated distributions. Although it underestimates the loss frequency, the impact of overall estimation of aggregated losses might be minor because it overestimates the loss severity.

3.5.1 Truncated or Shifted?

Rozenfeld (2010) proposed to use shifted distributions to deal with the reporting threshold problem. Given that it usually produces more stable parameter estimates than truncated distributions, it has been favored in practice. However, since it overlooks the fact that there are losses below the threshold but assumes that the losses are all above the reporting threshold, and in practice banks may try to intentionally use shifted distributions to lower their capital estimates, this approach has not been favored by the regulators in United States. However, the debates on whether using truncated or shifted distributions have never stopped in both academia and industry.

1. Distribution fitting:

In practice, the true model is unknown. Assuming either a truncated distribution or shifted distribution faces the risk of model misspecification. When the true model is a truncated distribution, fitting a shifted distribution will give different parameter estimates from the true parameters. Usually the full distribution has a lighter tail. However, one should not compare the parameters of the fitted distribution to claim that the shifted distribution underestimates the tail. Instead, what we need to compare is the performance of fitting for losses over the threshold. Both the truncated distribution and shifted distribution can arrive at the same fitting for the losses over the collection threshold. Comparing the parameters of the fitted distribution has no implication about whether it will underestimate the tail or not. Cavallo (2012) used Vuong's test to compare the closeness of both the fitted truncated distribution and the fitted shifted distribution to the true model, which are found can be indistinguishable. While the parameters are usually completely different. no conclusion can be made by comparing the parameters, because truncated distributions and shifted distributions are different models.

On the other hand, when the true model is from shifted distribution, likewise, both the truncated distribution and shifted distribution can arrive at the same fitting performance to

the true loss again, and the estimated parameters are different. Usually, in practice, we see that the parameters of truncated distribution become so unintuitive making the tail become a very small proportion. This is on one hand due to the estimation stability problem of truncated distribution, and on the other hand due to probably misspecification. However, using Vuong's test to compare the closeness of the two models with the true model on the loss data over threshold, they can be indistinguishable as well. Cavallo et al (2012)'s study provides strong experimental evidence that considering the misspecification issue, the truncated model and shifted model are equally valid or invalid.

In practice, this suggests that the shifted model should not be simply abandoned by taking for granted that the truncated distribution is the true model, as it has been done by the US regulators. They claim that the shifted distribution neglects the losses below the threshold, which will lead to underestimation of the capital charge, and this was deliberately used by some banks so that they moved toward requiring banks to use truncated models. However, the truncated distribution can have misspecification risk as well. In practice, a truncated distribution sometimes overestimates the probability of small loss amount of each event, and the tail becomes too small to be unintuitive. This is very unrealistic that the observed losses occurred with so low probability and there are so many small losses unobserved. In addition, since the adjust frequency for the full loss distribution becomes too large, it is very difficult to use Monte Carlo simulation to calculate the VaR of aggregated loss distribution, which requires a exceptionally large number of trails and sample size thus leading to substantial instability of the VaR estimates.

2. Calculation of capital charge using truncated and shifted distribution:

For truncated distributions, there are usually two ways to calculate the capital charge: (1) simulate losses from full distribution, and adjust the observed frequency; (2) simulate losses from truncated distribution, and use the observed frequency.

Proposition 3.2 When the true model is a truncated model, a relation between the quantile estimate of the two approaches is that the latter is less than or equal to the former.

Proof): This proof is based on private discussions with Johnson (2013). Let X be a random variable drawn from a loss distribution $f(x)$, and $X^{(T)}$ be a truncated random variable from $\frac{f(x)}{1-F(u)}$. $X^{(T)}$ can also be represented as a random variable $(X|X > u)$.

Assume that the aggregated loss by using full loss distribution $f(x)$ is L, and the aggregated loss by using conditional distribution $\frac{f(x)}{1-F(u)}$ is $L^{(T)}$, then

$$
\begin{aligned}
P(L &= \Sigma_{j=1}^{n} X_j > K) = \sum_{n=1}^{\infty} p(N=n) P(\Sigma_{j=1}^{n} X_j > K) \\
&= \sum_{n=1}^{\infty} p(N=n) P(\Sigma_{j=1}^{n} X_j > K | \Sigma_{j=1}^{n} (X_j | X_j > u) > K) P(\Sigma_{j=1}^{n} (X_j | X_j > u) > K \\
+ \sum_{n=1}^{\infty} p(N &= n) P(\Sigma_{j=1}^{n} X_j > K | \Sigma_{j=1}^{n} (X_j | X_j > u) \leq K) P(\Sigma_{j=1}^{n} (X_j | X_j > u) \leq K \\
&= \sum_{n=1}^{\infty} p(N=n) P(\Sigma_{j=1}^{n} (X_j | X_j > u) > K \\
+ \sum_{n=1}^{\infty} p(N &= n) P(\Sigma_{j=1}^{n} X_j > K | \Sigma_{j=1}^{n} (X_j | X_j > u) \leq K) P(\Sigma_{j=1}^{n} (X_j | X_j > u) \leq K
\end{aligned}
$$

Note that the first term

$$\sum_{n=1}^{\infty} p(N=n)P(\Sigma_{j=1}^{n}(X_j|X_j>u)>K = P(L^{(T)}>K)$$

and the second term is non-negative. Therefore, $P(L>K) \geq P(L^{(T)}>K)$, then $F_L(K) \leq F_{L^{(T)}}(K)$, and

$$F_L^{\leftarrow}(\alpha) \geq F_{L^{(T)}}^{\leftarrow}(\alpha).$$

For shifted distributions, only the latter way is applicable to calculate the capital charge. In practice, the shifted distribution is criticized that the tails are underestimated. This is true but should be understood correctly. This is not a statement resulting from comparing the parameters of shifted distribution and full distribution. It is comparing the quantile estimate of a shifted distribution with a full distribution. However, one should also note that, for truncated distributions, this applies as well. When the adjusted loss frequency is too big (meaning the loss above threshold from the fitted distribution is a very small proportion), simulating losses with the full frequency is neither reasonable nor practical. In this case, even a full distribution is estimated from the tail, using it to calculate capital estimate is impossible. One is still forced to use the quantiles of aggregated loss distribution using observed frequency and conditional distribution above the threshold, which is similar to the treatment for shifted distribution.

Rozenfeld (2010) shows that the annual aggregated loss amount below the threshold, which some banks collect as well, can be incorporated into the mean-corrected single loss approximation by writing the quantiles of the full distribution as quantiles of conditional distribution. Note that:

$$F_u(x) = \frac{F(x)-F(u)}{1-F(u)}$$

$$F(x) = F(u) + (1-F(u))F_u(x)$$

$$F^{-1}(\alpha) = F_u^{-1}\left(\frac{\alpha - F(u)}{1-F(u)}\right), \alpha \geq F(u)$$

$$\begin{aligned} VaR_\alpha(L) &\approx F^{\leftarrow}(1-\frac{1-\alpha}{\lambda}) + (\lambda-1)\mu \\ &= F_u^{\leftarrow}\left(\frac{1-\frac{1-\alpha}{\lambda}-F(u)}{1-F(u)}\right) + (\lambda-1)\mu \end{aligned}$$

where $\mu = \int_0^\infty x f(x)dx$ is the mean of severity distribution.

Let $\mu_N = \int_0^u x\frac{f(x)}{F(u)}dx$ be the mean of severity below threshold, and $\mu_T = \int_u^\infty x\frac{f(x)}{1-F(u)}dx$ be the mean of severity above threshold. Then the term $\lambda\mu$ can be written as:

$$\begin{aligned} \lambda\mu &= \lambda\int_0^\infty x f(x)dx \\ &= \lambda F(u)\int_0^u x\frac{f(x)}{F(u)}dx + \lambda(1-F(u))\int_u^\infty x\frac{f(x)}{1-F(u)}dx \\ &= \lambda_N\mu_N + \lambda_T\mu_T \end{aligned}$$

where λ_T is the frequency above threshold and λ_N is the non-observed frequency below threshold

$$\begin{aligned} \lambda_T &= \lambda(1 - F(u)) \\ \lambda_N &= \lambda F(u). \end{aligned}$$

Therefore,

$$VaR_\alpha(L) \approx F_u^{\leftarrow}\left(1 - \frac{1-\alpha}{\lambda_T}\right) + \lambda_T \mu_T + \lambda_N \mu_N - \mu$$

where the term $\lambda_N \mu_N$ is the unobserved annual losses below the threshold, μ is the mean of loss severity distribution. This shows that the single loss approximation applied to the full distribution can be written as the SLA applied to conditional distribution with adjustment by $(\lambda_N \mu_N - \mu)$.

3.6 Goodness-of-fit test

Goodness-of-fit test is an important step in model selection. There are usually in-sample goodness-of-fit tests and out-of-sample goodness-of-fit tests (or backtesting). In operational risk modeling given that 99.9% quantiles of losses are calculated and the loss data is limited, out-of-sample testing is usually not so meaningful for a short horizon. In this section, we will discuss several usual goodness-of-fit methods.

3.6.1 Visual testing

Visual testing is a very important step because it provides the most intuitive picture how the fitted distribution matches with the sample. There are many approaches of visual testing including the CDF plot, PDF plot, QQ plot etc.

3.6.2 Pearson's Chi-Squared test

Simply put, this test first divides the sample values into k classes, and then compares the sample frequency and theoretical frequency of the fitted distribution. Formally, the test statistic is defined as:

$$T = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

where O_i is the observed frequency in the i-th group, E_i is the expected frequency inferred from the fitted distribution.

The test statistic asymptotically approaches a χ^2 distribution. The critical values and p-value of the test can then be obtained from χ^2 distribution. The deficiency of this test statistic is that it is sensitive to the choice of groups.

3.6.3 Empirical distribution function (EDF) -based tests

EDF-based tests compare the fitted cumulative distribution function with the empirical distribution function. The empirical distribution function is defined as

$$F_n(x) = \frac{\Sigma_{i=1}^n I\{X_i \leq x\}}{n} = \begin{cases} 0 & x < X_{(1)} \\ 1/n & X_{(1)} \leq x < X_{(2)} \\ \\ (n-1)/n & X_{(n-1)} \leq x < X_{(n)} \\ 1 & x \geq X_{(n)} \end{cases}$$

where $X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}$ are the order statistics of the sample.

There are various popular tests using different definitions of the distance between two cumulative distribution functions. Typically, there are two classes of distances: supremum-type and quadratic-type. The frequently used EDF-based tests include Kolmogorov-Smirnov (KS) test, Kuiper test, Cramer-von Mises test and Anderson-Darling test.

- Kolmogorov-Smirnov (KS) test

Kolmogorov (1933) proposed the following form of test statistic:

$$D_n = \sqrt{n} \sup_x |F_n(x) - F(x)|$$

For a sample $X_1, ..., X_n$, it is calculated by following formula:

$$\begin{aligned} D_n &= \sqrt{n} \max(D_n^+, D_n^-) \\ D_n^+ &= \max_i (F(X_i) - \frac{i-1}{n}), i = 1, 2, ..., n \\ D_n^- &= \max_i \left(\frac{i}{n} - F(X_i) \right), i = 1, 2, ..., n \end{aligned}$$

An important property is that the distribution of KS test statistic depends on n but is independent of the distribution $F(x)$. For large n, the probability distribution of KS is given by:

$$\lim_{n \to \infty} P(\sqrt{n} D_n < z) = L(z) = 1 - 2 \sum_{j=1}^{\infty} (-1)^{j-1} e^{-2j^2 z^2} = \frac{\sqrt{2\pi}}{z} \sum_{i=1}^{\infty} e^{-(2i-1)^2 \pi^2 / (8z^2)}$$

New proof of this can be seen in Feller (1948) and Doob (1949). The function $L(z)$ has been tabulated by Smirnov (1948).

Kolmogorov (1933) also provided a recursion formula for calculating the distribution with finite n. Birnbaum (1952) tabulated the numerical values for finite n. A practical way of calculate the distribution is provided in Marsaglia et al (2003).

- Kuiper test:

Kuiper test (Kuiper (1960)) is closely related to KS test but defined by the sum of D_n^+ and D_n^- :

$$V = \sqrt{n}(D_n^+ + D_n^-)$$

- Cramér-von Mises test

Cramér (1928) suggested a following test statistic:

$$\int_{-\infty}^{\infty}(F_n(x)-F(x))^2dK(x)$$

where $K(x)$ is a suitable nondecreasing weight function. Von Mises independently made an equivalent suggestion and developed several properties of the test.

Smirnov (1937) modified the test as a distribution-free form if $F(x)$ is continuous:

$$W_n^2 = n\int_{-\infty}^{\infty}(F_n(x)-F(x))^2\psi(F(x))dF(x)$$

where $\psi(t), 0 \le t \le 1$, is a nonnegative weight function. W_n^2 can also be written as:

$$\begin{aligned} W_n^2 &= n\int_{-\infty}^{\infty}(\frac{1}{n}\sum_{j=1}^{n} I_{\{X_j \le x\}} - F(x))^2\psi(F(x))dF(x) \\ &= n\int_0^1(\frac{1}{n}\sum_{j=1}^{n} I_{\{F(X_j)\le t\}} - t)^2\psi(t)dt. \end{aligned}$$

The usual so-called Cramer-von Mises test refers to the test when $\psi(t) \equiv 1$:

$$W_n^2 = n\int_{-\infty}^{\infty}(F_n(x)-F(x))^2dF(x)$$

Smirnov (1937) showed that in this case, under null hypothesis when $n \to \infty$, $W^2 = \lim W_n^2$ has in the limit "omega-squared" distribution independent of the hypothetical distribution function $F(x)$.

For finite samples, its computation formula is:

$$W_n^2 = \frac{1}{12n} + \sum_{i=1}^{n}\left(F(X_{(i)}) - \frac{2i-1}{2n}\right)^2$$

- Anderson-Darling test

While KS test put most weight on the body of the distribution, Anderson-Darling (1954) test tries to put more emphasis on the tails of the distribution. It is defined as:

$$AD_n^2 = n\int_{-\infty}^{\infty}\frac{(F_n(x)-F(x))^2}{F(x)(1-F(x))}dF(x)$$

This is a special case of the general Cramér-von Mises test when $\psi(t) = \frac{1}{t(1-t)}, 0 \le t \le 1$.

The limiting distribution of AD_n^2 has following representation (Anderson and Darling, 1952):

$$P(AD_n^2 \le z) = \frac{\sqrt{2\pi}}{z}\sum_{j=0}^{\infty}\binom{-1/2}{j}(4j+1)e^{-((4j+1)^2\pi^2)/(8z)}\int_0^{\infty}e^{z/(8(\omega^2+1))-((4j+1)^2\pi^2\omega^2)/(8z)}d\omega$$

The computing formula of AD_n^2 for finite sample is (Anderson and Darling, 1952):

$$\begin{aligned} AD_n^2 &= -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\left[\log F(X_{(i)}) + \log(1-F(X_{(n+1-i)}))\right] \\ &\quad -n - \frac{1}{n}\sum_{i=1}^{n}[(2i-1)\log F(X_{(i)}) + (2(n-i)+1)\log(1-F(X_{(i)}))] \end{aligned}$$

Marsaglia (2003) introduces a recursion-based algorithm to calculate the limiting distribution of AD_∞^2 and a simulation-based algorithm to compute the distribution of AD_n^2 for finite n.

Anderson Darling also introduced a supremum-type test (see Anderson and Darling, 1952):

$$K_n = \sup_x \sqrt{n}\,|F_n(x) - F(x)|\,\sqrt{\psi(F(x))}.$$

- Upper tail AD test:

Chernobai et al (2005) introduced a class of "upper tail" Anderson-Darling test statistic by defining $\psi(t) = 1/(1-t)$ in the supremum-type and $.\psi(t) = 1/(1-t)^2$ in the quadratic-type. This result in following statistics:

$$AD_{up} = \sqrt{n}\sup_x \left|\frac{F_n(x) - F(x)}{1 - F(x)}\right|$$

and

$$AD_{up}^2 = n\int_u^\infty \frac{(F_n(x) - F(x))^2}{(1-F(x))^2} dF(x)$$

For truncated samples, the $F_n(x)$ is the empirical distribution of truncated data, and $F(x)$ is the fitted conditional distribution $\frac{F(x)-F(u)}{1-F(u)}$where u is the truncation threshold. The computing formula for discrete samples for truncated data can be found in Chernobai et al (2005) as well.

3.7 Capital charge calculation

Under LDA, the operational loss distribution is an aggregated distribution of the loss frequency distribution with loss severity distribution:

$$L = \sum_{i=1}^{N} X_i$$

$$P(L < l) = \sum_{n=0}^{\infty} p(n) P(\sum_{i=0}^{n} X_i < l)$$

However, since the aggregated loss distribution could not be represented in an analytical form, numerical methods are needed to calculate the capital charge.

3.7.1 Monte Carlo Simulation

Monte Carlo simulation is the simplest approach to calculate the capital charge. It is implemented as follows:

1. Simulate the loss frequency $n_1, ..., n_s$ from the frequency distribution;
2. Simulate n_1,...,n_s losses from the severity distribution $\left\{X_1^{(1)}, .., X_{n_1}^{(1)}\right\}, ..., \{X_1^{(s)}, ..., X_{n_s}^{(s)}\}$;
3. Sum them up to obtain s aggregated loss scenarios $\{L_i = \sum_{k=1}^{n_i} X_k^{(i)}\}_{i=1}^{s}$.
4. Find the α-th quantile of the simulated aggregated loss sample $\{L_i\}_{i=1}^{s}$.

For operational risk capital charge, the 99.9% quantile needs to be calculated. Therefore, in practice, a large number of trials are needed to obtain a relatively stable estimate.

3.7.2 Panjer recursion

Definition 3.8 *(Panjer class)* A probability distribution of a discrete random variable p_k is said to be a member of Panjer class if for some $a, b \in R$ following holds for $k \geq 1$:

$$p_k = \left(a + \frac{b}{k}\right) p_{k-1}$$

The Poisson distribution and negative binomial distributions, which are usually used to model operational loss frequency, both belong to this family.

Poisson distribution is a Panjer$(0, \lambda)$ class:

$$p_k = \frac{\lambda^k}{k!} e^{-\lambda} = \left(\frac{\lambda}{k}\right) \frac{\lambda^{k-1}}{(k-1)!} e^{-\lambda}.$$

Negative binomial distribution is a Panjer $(1-p, (r-1)(1-p)/k)$ class:

$$\begin{aligned} p_k &= \left(\begin{array}{c} r+k-1 \\ k \end{array} \right) p^r (1-p)^k \\ &= \frac{(r+k-1)!}{k!(r-1)!} p^r (1-p)^k \\ &= \left((1-p) + \frac{(r-1)(1-p)}{k}\right) p_{k-1}. \end{aligned}$$

Panjer recursion is recursive method to calculate the compound distribution if the frequency distribution belongs to a Panjer class..

Theorem 3.3 *(Panjer recursion)* If the frequency distribution satisfies

$$p_k = (a + b/k)\, p_{k-1}, k \geq 1$$

for some $a, b \in R$, let X_i be i.i.d r.v.'s with $P(X_i = j) = f_j$, then the compound distribution $g_K = P(L = \sum_{i=1}^{N} X_i = K)$ satisfies the recursion

$$g_K = \frac{1}{1 - a f_0} \sum_{j=1}^{K} \left(a + \frac{bj}{K}\right) f_j g_{K-j}, K \geq 1$$

Proof): See McNeil, Frey, and Embrechts (2005).

For operational loss distribution, the aggregated loss distribution can be approximated by Panjer recursion:

1. Set a maximum loss M, choose a step size $\Delta x = \frac{M}{n}$to discretize the loss severity distribution such that

$$f_j = F((j+1)\Delta x) - F(j\Delta x), j = 1, ..., n.$$

. For operational loss severity distributions, we usually have $f_0 = 0$.

2. $g_0 = P(N = 0)$,

3. $g_j = \frac{1}{(1-af_0)} \sum_{i=1}^{j} \left(a + \frac{bi}{j}\right) f_i g_{j-i}$.

3.7.3 Single loss distribution Approximation (SLA)

For subexponential distributed i.i.d r.v's X_i, the single loss distribution and the aggregated loss distribution have following asymptotic relation:

$$P(L = \sum_{i=1}^{N} X_i > x) \sim E[N]P(X_i > x), x \to \infty$$

This suggests that the operational VaR is asymptotically given by (see Bocker and Klupelburg (2005)):

$$VaR_\alpha(L) = F^{\leftarrow}(1 - \frac{1-\alpha}{\lambda})$$

where $\lambda = E[N]$ is the expectation of frequency.

However, this representation usually underestimates the VaR when the frequency is very high. Thus, by utilizing the long-tailed property of sub-exponential distributions, Bocker and Sprittulla (2006) suggested a mean-corrected approximation:

$$VaR_\alpha(L) = F^{\leftarrow}(1 - \frac{1-\alpha}{\lambda}) + (\lambda - 1)\mu$$

This approximation greatly improves the accuracy, and has been used as an alternative to numerical method in calculating capital charge.

3.8 Aggregation of capital across categories

According to the Basel II, operational losses are required to be classified into 7 event types and 8 business lines. An intuitive way to evaluate the total risk would be to sum up all the risk measures for each cell, or group all the losses together and then to estimate risk for the total sample. However, either way would lead to biased estimates. The first approach (referred to as LDA methodology) assumes the losses from different cells are perfectly dependent. According to the diversification rule, this actually gives an upper bound of the total capital charge, and overestimates the total risk. The second approach in fact assumes that the losses from various groups are independent, which might not be true. For instance, the losses of some type of events in the same business line could be dependent with each other.

Giacometti et.al (2008) applied copula to model the dependence between 3 business lines, which has greatly reduced the required risk capital compared with the perfect dependent approach. However, without granulating the losses further, it assumes the independence

of all losses in the same business line. From the correlation between the weekly loss frequency and weekly aggregated losses, we see that for some business lines, the correlation between different event types are significant. Therefore, it is necessary to further study the dependence structure for each cell not only for each business line or event type.

3.8.1 Frequency or severity dependence?

The dependence between the aggregated loss of each cell are believed to be from two parts: the frequency and the severity. An assumption behind compound models is that the individual losses within a same cell are independent. Frachot et al. (2004) showed that it is conceptually difficult to assume simultaneously severity independence in each cell and severity dependence between different cells. It also shows mathematically that it cannot be true in general but by chance. Therefore, based on the LDA model, an assumption was proposed that the correlation of aggregated losses are actually conveyed by the correlation between frequencies.

$$\left.\begin{array}{c} Corr(N_1, N_2) \neq 0 \\ Corr(X, Y) = 0 \end{array}\right\} \Rightarrow Corr(L_1, L_2) \neq 0 \tag{1}$$

However, a trivial result will be obtained that:

$$Corr(L_1, L_2) \leq Corr(N_1, N_2) \tag{2}$$

In practice, this inequality might not be true. Following two models are suggested to be investigated:

(1) The severities are independent: $Corr(X, Y) = 0$. The only dependence are conveyed by the loss frequencies.

(2) The severities are not independent: $Corr(X, Y) \neq 0$. Given that it is hard to estimate the intrinsic dependence between the severities, dependence is modelled on the aggregated losses.

Following the basic assumption of LDA model, serial correlation of the losses are ignored here.

3.8.2 Copula models

Copula approach is a flexible tool to capture the dependence and is widely used in modeling multivariate joint distributions. According to Sklar's theorem, for a d-dimensional cdf F with marginals $F_1, ..., F_d$, there exists a copula C, such that:

$$F(x_1, x_2, ..., x_d) = C(F_1(x_1), ..., F_d(x_d))$$

for all $x_i \in [-\infty, +\infty], i = 1, 2, ..., d$.And the copula function can be obtained as following:

$$C(u_1, ..., u_d) = F(F_1^{-1}(u_1), ..., F_d^{-1}(u_d))$$

There are various types of copula functions including Gaussian, student t and Archimedean copulas. Gaussian copula is pertaining to a multivariate normal distribution with standard normal marginals and correlation matrix R, which has cdf:

$$C_G(u_1, ..., u_d) = \Phi_d(\Phi^{-1}(u_1), ..., \Phi^{-1}(u_d); R)$$

where $u = (u_1, ..., u_d)$ is the vector of marginal probabilities, Φ_d denotes the cdf for the d-variate standard normal distribution with correlation matrix R, and Φ^{-1} is the inverse of the cdf for the univariate standard normal distribution.

Copula maintains the dependence structure under monotonic transformation, which is called invariance property, and is frequently used in modeling multivariate portfolio risk. Embrechts et al. (2003) might be an excellent guidance for applying copulas to risk management. Further discussions on copula and its properties, and the dependence measures can be found in McNeil et al. (2005).

3.8.3 Simulation approaches

Based on different assumptions that the dependence exists in frequency only or exists for severities as well, there are two copula simulation methods — frequency copula approach and aggregated loss copula approach.

- Frequency copula simulation

1. Generate p correlated vectors $U_{n\times p}$ from U(0,1) by the loss frequency copula C_f and obtain the loss numbers $N_{n\times p}$by inverting the vector with the Poisson distribution.

2. For each scenario of loss frequency $N_{i,j}$, simulate that number of loss severities $\{X_{i,:}\}$ by the loss severity distribution, as the loss severities are assumed independent.

3. Sum up the generated series $\{X_{i,j}\}$ as the aggregated losses during the i-th scenario of cell j.

- Aggregated-loss copula simulation

1. Generate p correlated vectors $U_{n\times p}$ from U[0,1] with the estimated total copula C_{total}.

2. Transfer each column U_j into simulated loss by inverting via the aggregated loss cumulative distribution L_j:

$$F_j(L_j) = \sum_{n=0}^{\infty} P(N_j = n)P(\sum_{l=1}^{n} X_j \leq L_j) = U_j$$

where X_j follows the loss severity distribution in the j-th cell and N_j follows the frequency distribution in the j-th cell.

The aggregated-loss copula copula approach requires calculating the compound distribution and simulating random variables from this compound distribution, which is computationally very intensive as the number of simulation needs to be very large in the heavy tailed modeling. Besides the Monte Carlo method, fast Fourier transformation (FFT) method and Panjer recursion are two popular alternatives to calculate the compound distribution (Shevchenko, 2010).

In addition to these methods of using general copulas, Bocker and Kluppelberg (2008) introduced a way of using Levy copula to model the frequency and severity dependency together. More approaches of modeling the dependence can be found as a summary in Shevchenko(2010).

3.9 Practical Issues with operational risk modelling

3.9.1 Categorization of loss data

Basel II Accord requires "sufficient granularity to capture the major drivers of operational risk affecting the shape of the tail of the loss estimates" when selecting unit of measures

(UoMs, or cells), and then the capital charges of each cell are aggregated by summation or justifiable dependence models. However, there are a lot of issues need to be considered in the selection of UoMs.

1. Data availability:

Ideally, the UoMs should follow the definition of Basel II as seven event types by eight business lines, and each intersection should be modelled separately. However, this is often impractical because some of the cells may not have enough loss events. Figure 3.1 shows a profile of loss frequency and loss amount of the internal losses from a European bank. It shows that most losses come from business line 3, 4, 5 and 8. The corresponding definitions of these business lines in the bank are: retail banking, commercial banking, payment & settlement, and retail brokerage. By event types, the losses mostly come from event type 2, 4, 7. The definition of the event types are consistent with Basel II categories: external fraud (EF), client, products & business practices (CPBP), and execution, delivery, & process management (EDPM).

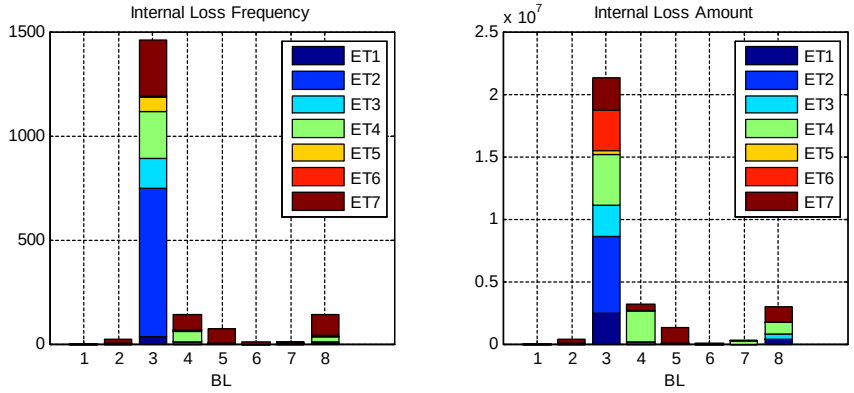


Figure 3.1: The loss frequency and loss amount in each event type/business line

Since the involved business lines for each bank are usually different, in a supervisory guidelines document for operational risk, BIS (2011) allows much flexibility in choosing UoMs by "taking into account the bank's idiosyncrasies including the business profile, risk profile, history of operational losses, business environment and other factors". The bank needs to seek a balance between granularity and data availability.

On one hand, Basel II Accord requires that the bank to capture the major risk drivers that affect the shape of the tail of loss estimates. For some UoMs, even though the loss events are relatively infrequent, their loss amount of each individual event (i.e., loss severity) is high. Therefore, pooling these losses with other losses would mask the risk drivers.

On the other hand, if UoMs are too fine, then, for some event type, the number of losses are too few and it may be impossible to fit a severity distribution.

2. Homogeneity:

Under LDA, the losses inside each cell are assumed to be independent and identically distributed. Compared to the dependence issue, the profile and shape of the loss distributions are more important. In BIS (2011), it also mentioned that "it is important that risks sharing common factors are grouped together". The bank should examine the homogeneity across categories and inside categories. Two-sample statistical tests including Kolmogorov-Smirnov test, Andersan-Darling test, Cramer–von Mises test, and Wilcoxon signed-rank test can be used to test the similarity between the losses from two categories.

For the losses inside one category, provided that data is sufficient, one can test the losses from level 2 categories or use other reasonable factors to further categorize the losses and test the similarity of them.

3. Aggregation:

Another possible factor that forces bank to seek a balance between granularity and data availability is the issue of aggregation across categories. When the loss categories are too many, summing up of the capital charges of each category may greatly overestimate the total capital charge, however, it may also be difficult to develop any dependence model or validate the dependence assumptions when the loss data in each category are too few. The usual correlation estimates may underestimate the true dependence due to the well-known Epps (1979) effect in econometrics and time series analysis — the empirical correlation tends to decrease as the sampling frequency increases.

3.9.2 Loss frequency modelling

In practice, the usual models for loss frequency modelling include the Poisson distribution and negative binomial distribution. However, there are several practical issues in loss frequency modelling.

1. Reporting latency

Operational loss records usually have several different reference dates: "date of occurrence, date of discovery, date of contingent liability, date of accounting (first financial impact), and date of settlement". The lag between date of occurrence and date of discovery could be up to years. Therefore, the loss frequency in the past historical period in an operational loss database usually keep evolving. There is often a declining trend of loss frequency in recent periods due to the lag of data collection and reporting (see Figure 3.2). This should be noticed when modelling the loss frequency process. A practical way is to exclude the recent periods in frequency modelling because including them would generally underestimate the loss frequency.

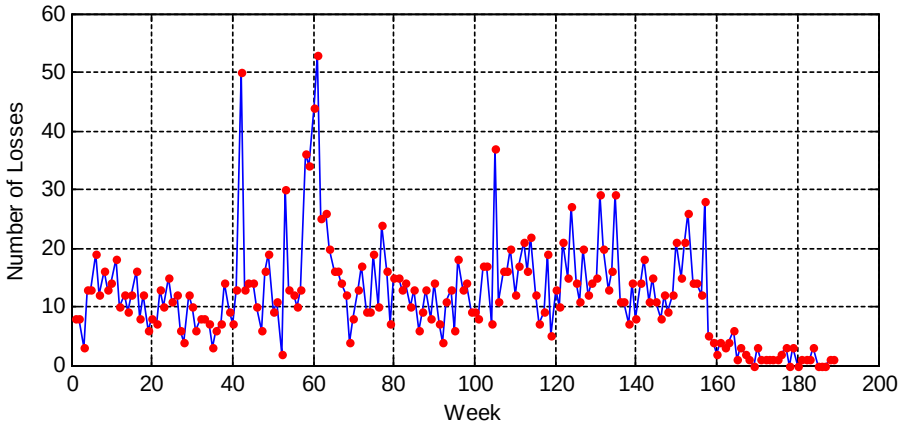


Figure 3.2: The declining trend of loss frequency due to reporting latency

2. Role of external loss data

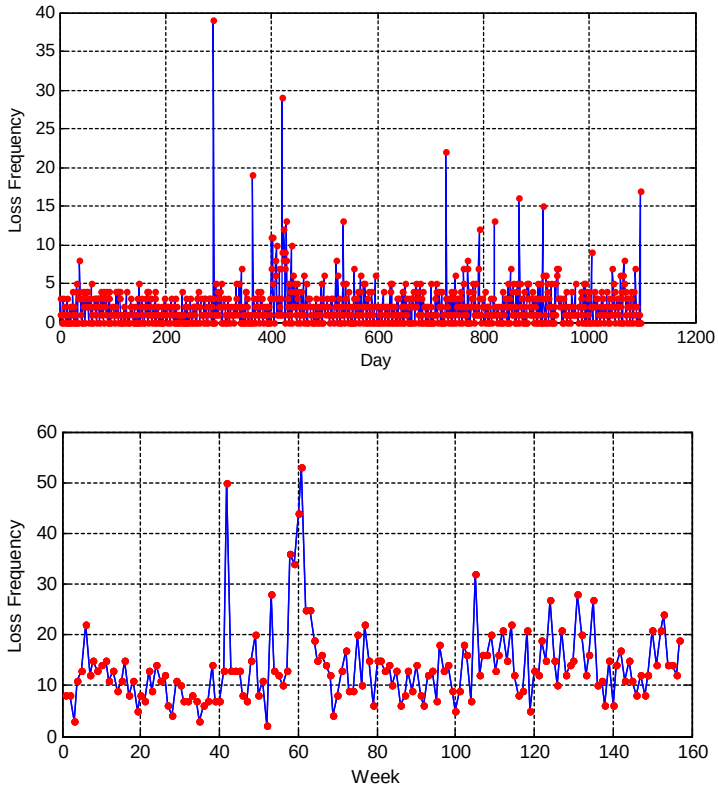
Since the external loss data are pooled from various institutions, the loss frequency is much higher than the internal loss data. Directly including external loss data for loss

frequency modelling is not appropriate. The widely accepted way is discarding the external loss frequency information and using internal loss frequency. However, the loss frequency profile among different UoMs in the external loss data may still shed some light on the internal loss frequency.

3. Selection of aggregation period

The final capital charge requires modelling the annual loss frequency. However, if using one year as the aggregation period, the total number of years in historical data is very limited for direct modelling. A usual approach is to select a sampling frequency such as quarterly or monthly and then aggregate them to obtain an annual loss frequency process. The selection of an aggregation period should be based on the length of historical data and the horizon of loss frequency forecasting. When a Poisson process is applied, the difference between choosing different aggregation period on the estimate of annual frequency should be very small.

However, when there is a trend observed in the loss frequency, nonhomogeneous Poisson process needs to be used instead and a short aggregation period is needed. However, the aggregation period should not be too short otherwise there will be a lot of zero frequencies. In fact, except for few outliers, the loss frequency is often close to a Poisson process and the over-dispersion effect is often closely related to use of short aggregation period (see Figure 3.3).



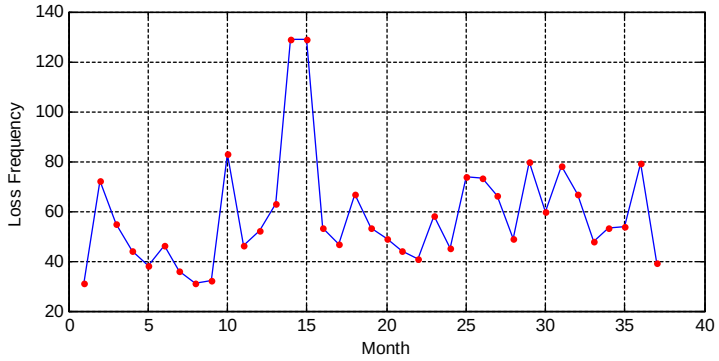


Figure 3.3: The daily, weekly and monthly loss frequency

The dependence of loss frequency across UoMs also relies on the selection of aggregation period. The following plot shows the dependence of loss frequency, average losses, and aggregated losses when choosing different aggregation period (daily, weekly, and monthly). It shows that with longer aggregation period, the correlation of loss frequency and aggregated losses are higher. A short aggregation period may lead to underestimated dependence due to the Epps effect.

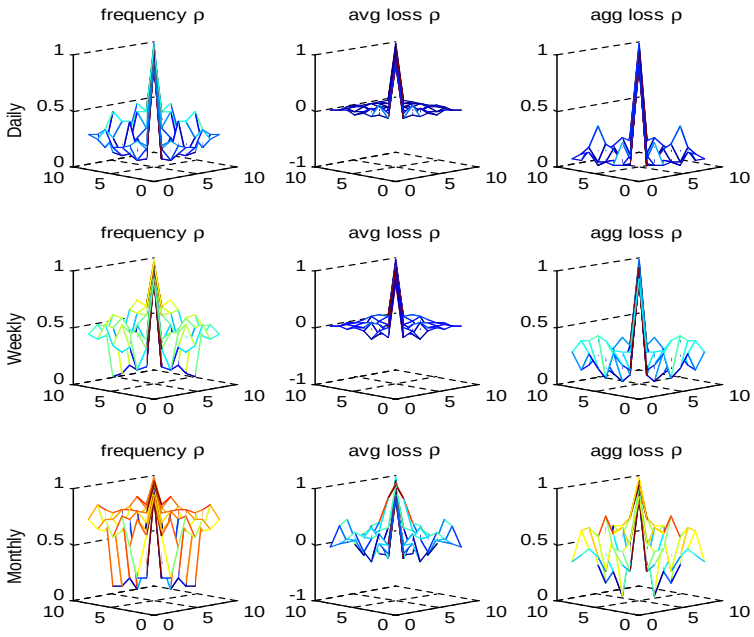


Figure 3.4: Correlation of frequency, average losses, and aggregate losses with various time intervals

3.9.3 Loss severity modelling

Compared to loss frequency modelling, loss severity modelling is much more important and is the key of operational risk measurement. Given the small sample size and data truncation, there are many practical issues with loss severity distribution fitting.

1. Parameter estimation

Due to the small sample size and truncation of data, the estimation of the distribution parameters for operational loss severity distribution becomes very difficult. Maximum likelihood estimates often become highly unstable due to the distortion of the log-likelihood by the denominator in the conditional distribution density. There are various approaches proposed for dealing this problem such as constraining the parameter space, constraining the truncated proportion, EM algorithm, and penalized likelihood method. We review the cause of difficulty in loss severity distribution fitting, and propose using Bayesian estimation method with non-informative priors to reduce the parameter instability. A complete and detailed study is presented in Section 3 (Bayesian estimation of truncated data).

2. Goodness of fit test

The goodness of fit of a distribution should be evaluated through both visual and quantitative approaches. The visual testing usually include comparing the fitted CDF and the empirical CDF, comparing the fitted PDF and the histogram, and comparing the fitted quantiles and the sample quantiles. Figure 3.5 shows an example of fitting lognormal distribution to (ET2, BL3) of internal loss data.

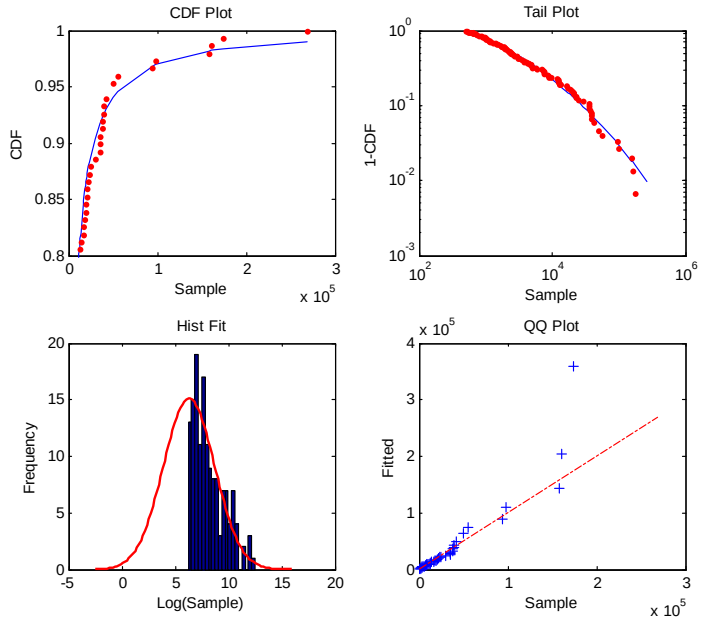


Figure 3.5: Examples of visual tests (CDF plot, Tail plot, Hist fit and QQ plot)

There are also other quantitative goodness-of-fit test method such as Chi-square test and EDF-based test (KS test, CvM test and AD test). Chernobai et al (2007) proposed a one-side AD test for loss distributions to put more weight on the upper tails. However, one difficulty of using these EDF-based tests is that the distribution of the test statistics

are hard to obtain. For example, for the upper-tail AD test, there is not a known limiting distribution. Alternatively, an approach to calculate the critical values and p-values for a given statistical test is by simulation methods. The general approach is described as following:

(1) Estimate the parameters $\hat{\theta}$ of the assumed distribution $f(x;\theta)$ using sample data, and calculate the test statistics value $\hat{T}$.

(2) Simulate M samples from $f(x;\hat{\theta})$ and re-estimate the parameters $\{\hat{\theta}_i\}_{i=1}^{M}$;

(3) Calculate the test statistics for each sample $\{\hat{T}_i\}_{i=1}^{M}$, thereby obtain a simulated distribution $\hat{G}(T)$ for the test statistics;

(4) Calculate the p-value as $p(\hat{T}_i > \hat{T})$, and the critical value of significance level $(1-\alpha)\%$ be $CV = G^{\leftarrow}((1-\alpha)/100)$. If $\hat{T} > CV$, reject the null hypothesis that the sample comes from the assumed distribution $f(x;\hat{\theta})$.

The simulation-based EDF test relies on the consistency and stability of re-estimated parameters to achieve a consistent and stable critical value. However, for small samples and truncated distributions, due to the large variability and difficulty in re-estimation, this approach may often appear to be too strong and even often reject the true parameters. Therefore, when selecting distributions, this approach should not be relied on and need to be incorporated with other visual testing approaches.

3. Distribution selection

In general, the selection of distributions have several principles as suggested in Dutta and Perry (2006) ordered by importance: (1) good fit - how well the model fits the data; (2) realistic - is the capital estimate realistic; (3) well-specified - are the characteristics of fitted data similar to loss data and logically consistent; (4) flexible - is it flexible to accommodate a wide variety of empirical loss data; (5) simple - is the method easy to use in practice.

Above all, the goodness of fit is the most important, but not the only criterion. For example, when the sample data are lack of extreme events, one may find that a lighter-tailed distribution (such as Weibull distribution) fits the data better than other heavier-tailed distributions (such as lognormal and loglogistic). However, it is often found that Weibull distribution usually produces a very low capital estimate and underestimates the tail risk beyond the sample. In addition to goodness-of-fit test for each category, it is often useful to fit all the loss data without categorization or for external data to provide a benchmark estimate or overall analysis.

Second, the reality of capital estimates needs to be considered. If there are multiple distributions having comparable goodness-of-fit test performance for the empirical data, but some of the capital estimates are unrealistic (such as exceed total asset value of the company). Using these capital estimates would be meaningless.

Third, whether the fitted data has similar characteristics with empirical data. Some distributions may have excellent in-sample goodness-of-fit test performance, but simulated losses from the fitted distribution may exhibit completely different characteristics from the empirical data. For example, the simulated losses may have excessively higher variability and result in non-robust capital estimates. It is worth resampling losses from the fitted distribution and analyzing the stability of the fitted model.

Fourth, the flexibility of a distribution should be considered. Theoretically, we should not believe the loss generating mechanism greatly varies across different categories. A distribution which is able to accommodate more types of losses is preferred. There are occasionally some distributions suitable for a specific type, but it should be examined whether they are sustainable across other categories or over time.

Finally, the brevity of the model should be considered as a factor for practical reasons.

A complex model with more parameters or a complicated representation may face difficulty and instability in parameter estimation. If there are multiple models with comparable goodness-of-fit and similar characteristics, a simpler, faster and well-understood model is preferred.

4 Bayesian estimation of truncated data with applications to operational risk measurement

4.1 Introduction

Operational risk management has become an increasingly popular topic because commercial banks have been required to calculate the operational risk capital charges based on the Basel II Accord. Among the current advanced measurement approaches, the loss distribution approach (LDA) is one of the standard approaches to quantify the operational risk and capital charge. Essentially, LDA involves the modelling of both the frequency and severity distribution for each cell, where there are 56 cells because the losses are categorized into seven event types and eight business lines as mandated by Basel II.

Although the idea of LDA is relatively easy in theory, there are still many thorny problems when estimating the loss severity distributions in practice. One of the major problems is the shortage of loss data for estimating the risk profile. The usual solution is to pool with external data. One way used by Baud et al (2002) and Aue and Kalkbrener (2006) is to adjust the probabilities of external losses under what the former refer to as a fair mixing assumption — external data are assumed to have the same distribution as internal data except for a different threshold. When this assumption is not valid, the external losses are usually scaled by size in order to be comparable to the internal losses. Dahen and Dionne (2007) and Cope and Labbi (2008) scaled the external losses by exposure indicators such as firm size and geographical region. However, the external data may neither have the same profile as internal data, nor be able to be scaled by a few observable factors. Moreover, even if pooled with external data, the cells with scarce data may still have few observations. Therefore, the small sample inference is still inevitable given that the cells with few observations may contain the most serious losses.

In addition to the problem of a small sample size, operational losses are often only collected above a reporting threshold. Estimating unconditional distributions and overlooking the threshold will produce biased estimates (see Chernobai et al (2005a, 2005b, 2006)) and consequently left-truncated distributions need to be used instead. However, estimating parameters for a truncated distribution is much more complicated than estimating parameters for a complete distribution. Typically, the maximum likelihood estimation (MLE) method has been used to estimate the loss distribution due to its asymptotic efficiency. The estimation could be accomplished using a direct numerical optimization or an expectation-maximization (EM) iterative algorithm (see Chernobai (2007) for a general formulation and Bee (2005) for the case of a lognormal distribution). However, it has been observed that the maximum likelihood estimates may suffer from substantial variability and sensitivity to outliers (see Huber (2010)). To overcome this issue, Cope (2011) developed a penalized likelihood estimator (PLE) for truncated data by adding a linear penalty term to the log-likelihood to decrease the variability. This method greatly reduces the mean-squared error of estimation by adding a small amount of bias. However, this method relies on the calculation of the inverse of Hessian matrix of the log-likelihood function at pre-estimated parameters, and this could be a very challenging task when the likelihood surface is too flat and the Hessian matrix is ill-conditioned.

In this paper, we propose a more reliable and flexible approach for truncated data inference, especially where there are small samples. The Bayesian method has been suggested as an alternative to the MLE method due to its underlying idea of combining expert prior information and its superiority to the MLE method in small sample estimation. Dalla Valle and

Giudici (2008) introduced a Bayesian approach to estimate the marginal loss distributions in operational risk management. However, in their study they reported little difference in the estimates compared to the MLE method, and therefore the superiority of the Bayesian method was not found to be significant. This finding is not surprising since Dalla Valle and Giudici (2008) did not consider the truncation problem.

There are several other theoretical contributions and simulation work applying Bayesian inference in operational risk (see Cruz (2002), Shevchenko (2011), and Peters and Sisson (2006)). One of the issues in using the Bayesian method is the selection of a prior distribution. There are usually three types of priors: elicited prior, vague (or "flat") prior, and noninformative prior such as the Jeffreys' prior. Since there is generally no proper expert information available and vague priors may become informative under a different parametrization, we propose employing the Jeffreys' prior, which is invariant under any transformation.

Using a simulation example based on the truncated lognormal distribution, we show in this paper that the Bayesian approach greatly reduces the variance compared to the MLE method. Even with a small sample size, the Bayesian approach gives stable and relatively reasonable estimates and is very useful for improving the credibility of parameter estimates when faced with scarce loss data. Using a real loss data set, we estimate the parameters for each intersection of event-type/business line with loss data and find that it is much more reliable and gives more reasonable estimates when the other methods fail. Applying a bootstrap method, we show how to obtain the confidence intervals for the estimated parameters and value-at-risk measure.

The rest of the paper is structured as follows. In Section 2, we describe the problem of using the MLE method for truncated data and the usual approaches for improvement such as the EM algorithm, constrained MLE, and penalized likelihood estimates. In Section 3, we introduce the Jeffreys' prior for truncated normal and truncated gamma distributions, because we find lognormal and loggamma distributions are useful for characterizing loss distributions. The general routines of our proposed estimation procedure is also briefly described. We provide an extensive simulation study based on the lognormal distribution, and compare the performance of MLE, EM, PLE, and Bayesian methods in Section 4. The effect of various priors and different criteria to choose Bayesian estimates are also discussed. An application to real operational losses data of a European bank is provided in Section 5 where the estimation of parameters and value-at-risk (VaR) measure is obtained for each cell with loss data based on both internal and external data. The last section summarizes our principal findings.

4.2 Usual methods for truncated data inference

The MLE method is widely used in statistical inference due to its asymptotic normality and asymptotic efficiency properties. It performs well when there is a large sample and no truncation of the data because the likelihood surface is generally well-shaped and has good convex properties. However, the practical situation in operational risk measurement is that the losses are generally only reported above a threshold, which is assumed in this paper to be known for both the internal data and external data. Directly fitting an unconditional distribution for the loss severity will clearly produce biased estimates for the loss distribution. (For a detailed discussion of this topic, see Chernobai et al (2005a, 2005b, 2006).) In order to accommodate truncated data, a conditional distribution is fitted instead of an unconditional distribution.

4.2.1 Truncated MLE

Let the truncated threshold be denoted by u, and the assumed density function of the loss severity distribution be $f(x;\theta)$, then the MLE of θ for the truncated data is:

$$\widehat{\theta} = \arg\max_{\theta\in\Theta} \sum_{k=1}^{n} \log \frac{f(x_k;\theta)}{1-F(u;\theta)}$$

However, directly solving the MLE for a truncated distribution is much more difficult than the corresponding complete distribution. Generally, no analytical solutions are available. Therefore, numerical optimization methods are needed, which might themselves have some problems. On one hand, singularity issues arise because of the term $(1-F(u;\theta))$ in the denominator. Notice that when the estimated truncated area approaches 100%, the denominator goes to zero. On the other hand, as shown by Cope (2011), the log-likelihood surface is greatly distorted by the conditional term $\log(1-F(u;\theta))$. As a result, the log-likelihood surface becomes very flat over a broad region in the parameter space, and the credibility of the estimated parameters is very low. Figure 4.1 shows the log-likelihood surface for the truncated normal distribution and complete normal distribution with respect to different parameter μ and σ. From the figure we can see that the log-likelihood of the truncated distribution is greatly distorted.

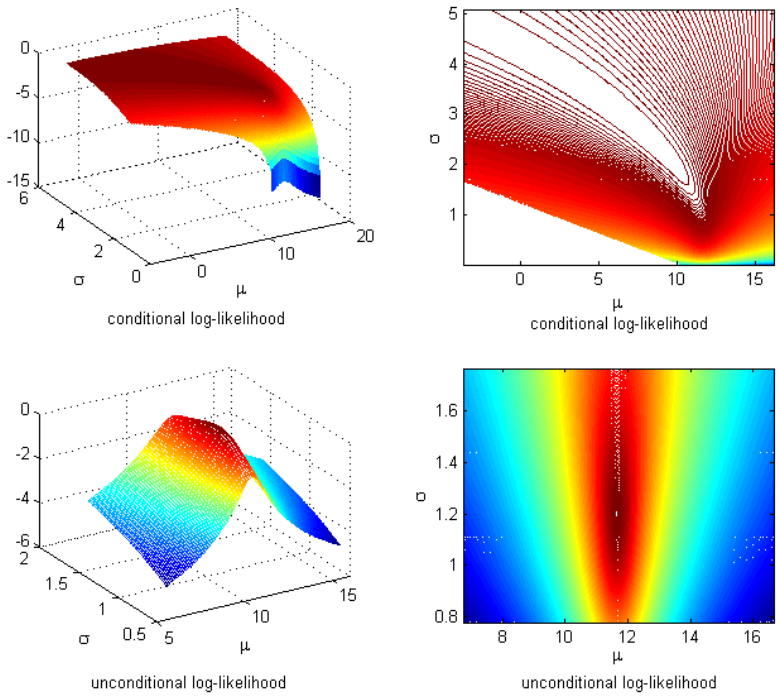


Figure 4.1: Conditional and unconditional log-likelihood

4.2.2 EM algorithm

To circumvent the difficulty of searching for a solution through numerical optimization (hereafter referred to as the MLE method), the EM algorithm has been suggested. Instead of

dealing with the conditional distribution, this method takes a more straightforward approach by optimizing the complete log-likelihood function. The basic idea is that given an initial guess for the distribution's parameters, in each iteration the next step's parameters are obtained such that the expectation of the complete log-likelihood with respect to the current parameters is maximized. The routine is described below (see Chernobai et al (2007) and Bee (2005) for more details).

Let X be the missing sample, and Y be the observed sample. Given an initial guess for the parameters denoted by $\theta^{(0)}$, and let $\theta^{(p)}$ denote the value of parameters after p cycles, then the iteration in the next cycle is as follows:

- *E-step*: Calculate the expected log-likelihood function of the complete data (X, Y)

$$E_{\theta^{(p)}}[\log L_\theta(X, Y)|Y].$$

- *M-step*: Find the parameter $\theta^{(p+1)}$ that maximizes the log-likelihood:

$$\theta^{(p+1)} := \arg\max_\theta E_{\theta^{(p)}}[\log L_\theta(X, Y)|Y].$$

If X is from exponential family, then the iteration steps can be simplified as following (Dempster et al. 1977):

E-step: Estimate the complete data sufficient statistics $T(X)$ by finding

$$T^{(p)} = E_{\theta^{(p)}}[T(X)|Y].$$

M-step: Determine $\theta^{(p+1)}$ as the solution of following:

$$E_\theta[T(X)] = T^{(p)}.$$

Specifically, for the truncated normal distribution, Bee (2005) derived the explicit iteration formula for the parameters, which will be used in this paper in order to compare the parameters estimated by the other methods. Let $u^{*(t)} = \frac{u-\mu^{(t)}}{\sigma^{(t)}}$, $\pi = 1 - \Phi(u^{*(t)})$, $\alpha(\theta^{(t)}) = \frac{\phi(u^{*(t)})}{\Phi(u^{*(t)})}$, then

$$N_T^{(t+1)} := E[N_T|Y] = N_U(1-\pi)/\pi,$$

$$E[X|\theta^{(t)}, Y] = \mu^{(t)} - \sigma^{(t)}\alpha(\theta^{(t)}),$$

$$E[X^2|\theta^{(t)}, Y] = (\sigma^{(t)})^2(1 - u^{*(t)}\phi(u^{*(t)}) - \alpha^2(u^*)) + E[X|\theta^{(t)}, Y]^2,$$

$$\mu^{(t+1)} = \frac{1}{N_U + N_T^{(t)}}\left(\sum_{i=1}^{N_U} y_i + N_T^{(t)}E[X|\theta^{(t)}, Y]\right),$$

$$(\sigma^{(t+1)})^2 = \frac{1}{N_U + N_T^{(t)}}\left(\sum_{i=1}^{N_U} y_i^2 + N_T^{(t)}E[X^2|\theta^{(t)}, Y]\right) - (\mu^{(t+1)})^2.$$

The EM algorithm is superior to the usual numerical optimization because the log-likelihood increases at each iteration and the method is more reliable. However, it is essentially solving the same optimization and accordingly may converge very slowly, producing unrealistic estimates (e.g, negative location parameters) as well when the log-likelihood surface is very flat.

Another solution to mitigate the risk of obtaining unrealistic estimates is to restrict the parameter space to a reasonable domain. Usually there are two approaches. First, constraints are imposed on the parameters' domain. Transformations of parameters are often applied in order to avoid adding constraints in the optimization routine. For example, an exponential transformation is often helpful to restrict the parameters to be positive. Second, constraints may be put on the truncated area to prevent it from becoming unreasonable. However, the choice of the critical proportion is highly subjective and, as a result, we do not recommend using this approach if the result is very sensitive to the constraints selected.

4.2.3 Penalized likelihood estimate

From a theoretical point of view, the reason that maximum likelihood estimates exhibit high variability is because the Fisher information matrix of a truncated distribution is usually much more ill-conditioned than the complete distribution. As a result, the likelihood surface is too flat for one to locate the maximum and the estimated parameter has a large variance and a low confidence level. Cope (2011) introduced a penalized estimator by adding an additional linear term into the log-likelihood function to reduce the distortion caused by the truncation.

A summary of the PLE method is as follows. Let the penalized log-likelihood function be:

$$P_n(X;\theta,\kappa,v) = \sum_{i=1}^{n} l(x_i;\theta) - \kappa v^T \theta$$

where

$$l(x_i;\theta) = \log \frac{f(x_i;\theta)}{1-F(u;\theta)}$$

An approximate value of κ is given by:

$$\kappa \approx \frac{\nabla J(\theta_0) v}{2\lambda}$$

where θ_0 is the true parameter without the penalized term, $J(\theta_0) = trace(V_{\theta_0}^{-1})$, where V_{θ_0} is the Hessian matrix of unpenalized log-likelihood. $\nabla J(\theta_0)$ is its first gradient with respect to θ at θ_0, v is the first principal component of $V_{\theta_0}^{-1}$ with eigenvalue λ and $V_{\theta_0}^{-1} v = \lambda v$. Cope (2011) demonstrates that using this approach when there is a small sample size reduces the distortion of the log-likelihood surface and the maximum value becomes localized to its true parameter. The bias of this estimate is approximately $-\frac{\kappa}{n} V_{\theta}^{-1} v$ and therefore it is asymptotically close to the estimate obtained by using the MLE method. However, it may be challenging to apply this method when the Hessian matrix is ill conditioned or even not positive definite. Under such conditions, the PLE method may produce highly biased estimates when the size of the penalty term is overestimated, or fail when the direction of the penalty vector is opposite to the direction of the true parameter.

4.3 The Bayesian method

The Bayesian method has been suggested as an alternative to the MLE method as a way to incorporate prior information. Dalla Valle and Giudici (2008) apply the Bayesian approach using various frequency and severity distributions to a small sample of only 407 loss events distributed over eight intersections. However, they did not consider the reporting threshold

problem. Not surprisingly, their results suggest that the Bayesian method gives results that are similar to the MLE method. However, in the presence of a truncation threshold, the MLE method may suffer from the instability and non-robustness problems, and, to the best of our knowledge, in the literature little attention has been paid to applying the Bayesian method to truncated data inference, especially in the context of operational risk modelling.

4.3.1 Jeffreys' priors for truncated distributions

One important issue in using the Bayesian approach is the selection of a proper prior distribution. Using flat priors or non-informative priors without expert opinions is suggested because informative priors will lead to biased estimates with strong prior information. Cruz (2002) briefly discusses the problem of choosing a prior distribution. There are typically three types of priors: elicited priors, vague (or "flat") priors, and non-informative priors such as the Jeffreys' prior (Jeffreys (1967)). As we do not have any expert information and the vague priors may become very informative through a change of variables, we propose using the Jeffreys' prior which has a nice invariant property under any transformation. The Jeffreys' prior is defined as the determinant of the Fisher information matrix, and is a uniform density on the space of probability distributions in the sense that it assigns equal mass to each "different" distribution (Myung and Navarro (2004)). In addition, Firth(1993) notes that the Jeffreys' prior corrects the first-order bias of the MLE method for exponential family distributions with canonical parametrization.

Nevertheless, there is a vast literature addressing how to choose a better prior distribution. It has been argued that Jeffreys' prior may be deficient in multi-parameter problems when only a subset of the parameters are of inferential interest and the remaining are nuisance parameters (see Bernardo and Berger et al (1992) and Datta and Ghosh (1996)). Then, ad-hoc modifications need to be made on the prior. Alternatively, the "reference prior" approaches introduced by Bernardo (1979) and Berger and Bernardo (1989) have overcome this issue. Recently, Berger et al (2009) provides the formal definition of reference priors as well as a simple constructive formula for a reference prior. Applying the reference priors to truncated data may be a very interesting topic in Bayesian estimation of operational losses. However, for the truncated distributions in this study, say, the truncated normal distribution, both of the mean and variance parameter could be of primary interest and neither of them could be treated as a nuisance parameter. In addition, the computation for the reference priors in multi-parameter cases is extremely complex. Therefore, we do not address this issue in this paper.

In the following, we give the form of Jeffreys' prior for the truncated normal and truncated gamma distribution by working with their respective Fisher information matrix. The reason why we choose the two distributions is as follows. On one hand, the lognormal distribution has been widely used as the primary example in theoretical discussions as well as real data applications (see Chernobai (2006), Frachot (2001), Bee (2005), and Cope (2011)). In our simulation study, we follow this tradition. On the other hand, it is also noted that the lognormal distribution may underestimate the tails and consequently many alternative heavier-tailed distributions have been proposed such as lognormal, Weibull, log-Weibull, gamma, loggamma loglogistic, g-and-h, alpha-stable, Burr, generalized Pareto, extreme value mixture and piecewise distributions (see De Fontnouvelle et al (2007), Dutta and Perry (2006), Chernobai et al (2007), Giacometti (2008), and Aue and Kalkbrener (2006)). Among the candidate distributions, we choose the loggamma distribution as a representative of the heavier-tailed distribution because we find through our empirical study (discussed in

Section 5) that, along with the lognormal distribution, it is useful for characterizing losses, as well as being relatively flexible and robust to the various tails of each cell.

We denote the Fisher information matrix for the two-parameter distributions by

$$\begin{aligned} I(\boldsymbol{\theta}) &= -E[\frac{\partial^2}{\partial \boldsymbol{\theta}^2} \log f(X;\boldsymbol{\theta})|\boldsymbol{\theta}] \\ &= \left[\begin{array}{cc} A & B \\ B & C \end{array} \right]. \end{aligned}$$

By plugging in the log-density of the truncated normal distribution,

$$l(x;\boldsymbol{\theta}) = \log f(x;\mu,\sigma^2) = -\log(\sqrt{2\pi}\sigma) - \frac{(x-\mu)^2}{2\sigma^2} - \log(1-\Phi(\frac{u-\mu}{\sigma}))$$

we have:

$$A = \left(1-\alpha^2(u^*)+u^*\alpha(u^*)\right)/\sigma^2,$$

$$B = \alpha(u^*)(1-u^*\alpha(u^*)+u^{*2})/\sigma^2,$$

$$C = (2+u^*\alpha(u^*)-\alpha^2(u^*)u^{*2}+u^{*3}\alpha(u^*))/\sigma^2,$$

$$\text{where } \alpha(u^*) = \frac{\phi(u^*)}{1-\Phi(u^*)}, u^* = \frac{u-\mu}{\sigma}.$$

The derivation is provided in Appendix A. (For a general derivation, see Roehr (2002) and Appendix E, where the Fisher information matrix is derived to analyze the sensitivity of model parameters to various thresholds and examples of the truncated normal, generalized Pareto, and one-parameter Weibull distributions are given.)

For the truncated gamma distribution:

$$f(x;\kappa,\lambda) = \frac{x^{\kappa-1}e^{-x/\lambda}}{\Gamma(\kappa)\lambda^\kappa}/(1-G(u;\kappa,\lambda))$$

$$l(x;\boldsymbol{\theta}) = \log f(x;\kappa,\lambda) = (\kappa-1)\ln x - \frac{x}{\lambda} - \ln\Gamma(\kappa) - \kappa\ln\lambda - \ln(1-G(u;\kappa,\lambda)).$$

Let $g(x;\kappa,\lambda)$ be the complete gamma density, $H_{\theta_i}(u) = \frac{\partial}{\partial\theta_i}\int_0^u g(x;\kappa,\lambda)dx$, where $\theta_i = \kappa,\lambda; \overline{G}(\kappa,\lambda) = (1-G(u;\kappa,\lambda))$,then

$$\begin{aligned} A &= \Psi'(\kappa) - H_{\kappa\kappa}/\overline{G}(\kappa,\lambda) - H_\kappa^2/\overline{G}(\kappa,\lambda)^2, \\ B &= \frac{1}{\lambda} - H_{\kappa\lambda}/\overline{G}(\kappa,\lambda) - H_\kappa H_\lambda/\overline{G}(\kappa,\lambda)^2, \\ C &= \frac{2\kappa\overline{G}(\kappa+1,\lambda)}{\lambda^2\overline{G}(\kappa,\lambda)} - \frac{\kappa}{\lambda^2} - H_{\lambda\lambda}/\overline{G}(\kappa,\lambda) - H_\lambda^2/\overline{G}(\kappa,\lambda)^2. \end{aligned}$$

where $\Psi(\kappa) = (\ln\Gamma(\kappa))'$ is the digamma function. Letting $M_n = \int_0^u (\ln\frac{x}{\lambda} - \Psi(\kappa))^n g(x;\kappa,\lambda)dx$, the $H_\kappa, H_{\kappa\kappa}, H_\lambda, H_{\kappa\lambda}, H_{\lambda\lambda}$ in the above are given by:

$$H_\kappa = M_1,$$

$$H_{\kappa\kappa} = -\Psi'(\kappa)G(u) + M_2,$$

$$H_\lambda = -\frac{u}{\lambda} g(u;\kappa,\lambda)$$

$$H_{\kappa\lambda} = -\frac{u}{\lambda}(\ln\frac{u}{\lambda} - \Psi(\kappa)) g\,(u;\kappa,\lambda)$$

$$H_{\lambda\lambda} = \frac{\kappa(\kappa+1)}{\lambda^2}(g(u;\kappa+1,\lambda) - g(u;\kappa,\lambda))$$

The more detailed steps are described in Appendix B.

The Jeffreys' prior is therefore:

$$\pi(\theta) \sim |\det I(\theta)|^{1/2} = \sqrt{(AC - B^2)},$$

and the posterior distribution of θ is:

$$p(\theta|X) \propto p(X|\theta)\pi(\theta) = p(X|\theta)\,|\det I(\theta)|^{1/2}\,.$$

Theoretically, the Fisher information matrix is always positive semidefinite. However, it should be noted that in practice the negative values attributable to machinery precision level needs to be ruled out.

4.3.2 Behavior of Jeffreys' priors for truncated distributions

In this subsection, we analyze the behavior of the Jeffreys' priors for the truncated normal and gamma distributions with respect to the parameters under different truncations. This provides an intuitive explanation as to how the Jeffreys' priors correct the estimation error of MLE.

In Figure 4.2, we calculate the Jeffreys' prior for truncated normal $N(\mu,\sigma)$ under truncation thresholds $u = -\infty$, 6, 8, 10, 11. Note that, when $u = -\infty$, it corresponds to the complete normal distribution, whose Fisher information matrix is $[1/\sigma^2, 0; 0, 2/\sigma^2]$.It is interesting to note that the Jeffreys' prior for the truncated distribution is always below the Jeffreys' prior for the complete distribution. To illustrate the behavior of the Jeffreys' prior clearly, we plot the change of Jeffreys' prior with respect to different $\mu \in [5, 15]$ first for a fixed $\sigma = 2$. It is found that when the truncation threshold increases, the Jeffreys' prior tends to favor larger μ to a greater extent. Then for a fixed $\mu = 10$, we plot the Jeffreys' prior versus different $\sigma \in [0.5, 3.5]$, and find that the Jeffreys' prior always favors smaller σ, though to different extents under different truncations. Opdyke and Cavallo (2012) recently applied the influence functions to analyze the robustness of parameter estimates for the truncated loss severity distribution. In short, the influence function assesses the impact on the parameter estimates of an infinitesimal deviation at a specific severity value. They showed that the MLE tends to underestimate μ and overestimate σ greatly under high truncation thresholds. This explains analytically why the negative location parameters are frequently observed by using MLE in that case. Here, we show that this estimation error due to the truncation threshold could be corrected by the Jeffreys' prior, and the extent to which it makes the correction depends on the truncation threshold. (The surface plots of the Jeffreys' priors over the space of (μ,σ) under various truncation thresholds are shown in the Appendix C.)

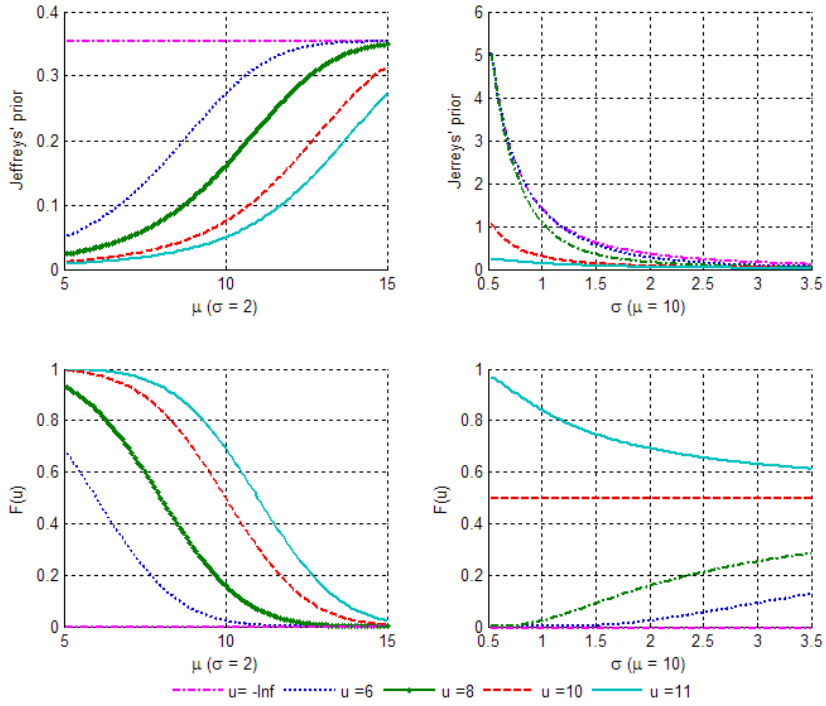


Figure 4.2: Jeffreys' prior for truncated normal distribution

We also analyze the Jeffreys' prior for the truncated gamma distribution $\Gamma(\kappa, \lambda)$ under different truncation thresholds $u = 0, 5, 6, 8, 10$, as shown in Figure 4.3. When $u = 0$, it corresponds to the complete gamma distribution, whose Fisher information matrix is $\left[\Psi'(\kappa), 1/\lambda; 1/\lambda, \kappa/\lambda^2\right]$. For a fixed $\lambda = 0.4$, we plot the Jeffreys' prior versus different $\kappa \in [5, 60]$. For fixed $\kappa = 20$, we plot the Jeffreys' prior versus different $\lambda \in [0.2, 1]$. Interestingly, unlike the complete distribution's Jeffreys' prior, which is a decreasing function of κ and λ, the truncated distribution's Jeffreys' prior is not a monotone function of either κ or λ. Moreover, we found that when the truncation threshold increases, the peak of the Jeffreys' prior also shifts to the right. Opdyke and Cavallo (2012) showed that the MLE tends to underestimate κ and overestimate λ (in their paper, they used a different parametrization from ours by denoting $b = 1/\lambda$), and this effect is more sensitive under higher truncation thresholds. Similarly as in the case of the normal distribution, the Jeffreys' prior corrects this effect by favoring larger κ and smaller λ. Note that even though for a fixed κ the Jeffreys' prior favors a larger λ, when checking the surface of Jeffreys' prior over the space of (κ, λ) we found that the maximum is achieved at a large κ and a small λ (see the surface in the Appendix D). The shape of the Jeffreys' prior determining how much it corrects the estimation error of MLE varies with the truncation thresholds.

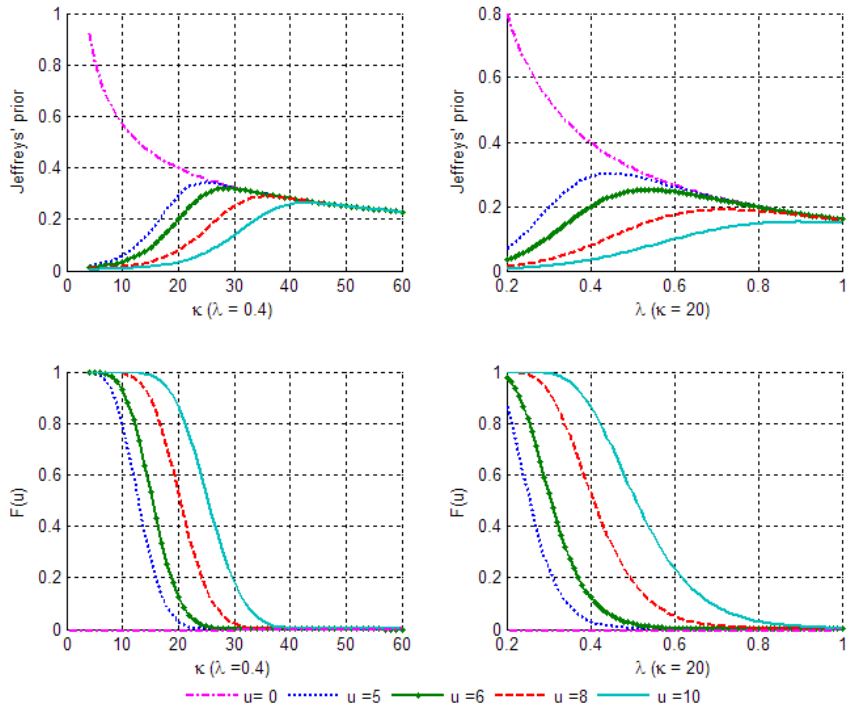


Figure 4.3: Jeffreys' prior for truncated gamma distribution

4.3.3 Markov chain Monte Carlo method

The Markov chain Monte Carlo (MCMC) method is the standard approach for simulating draws from the posterior distribution when a closed-form solution is unavailable. The method is well known so below we only briefly describe the algorithm used in this study:

- *Step 1*: Start with an initial guess about the parameter θ_0 such that the posterior distribution is well-defined.
- *Step 2*: Choose a random walk distribution, referred to as the proposal density, to sample a candidate point θ^*. Generally, a Gaussian random walk or uniform distribution random walk can be used. A judgement is needed to ensure the parameter would not jump out of the domain of θ.
- *Step 3*: Given the candidate point θ^*, calculate the acceptance ratio

$$\alpha = \min\{\frac{p(X|\theta^*)\pi(\theta^*)}{p(X|\theta_{t-1})\pi(\theta_{t-1})}, 1\}.$$

  There are several acceptance rules that can be used. Here we use the Metropolis algorithm. Notice that when the proposal density is not symmetric, the Metropolis-Hastings algorithm needs to be used instead.

For the truncated distributions discussed in this paper, the likelihood function is very flat, especially for small samples and the posterior distribution may often exhibit multi-modal properties. In this case, the MCMC methods would be very inefficient. To ensure the convergence of the chain, a very large number of calculations are needed. In the case of point estimation, the maximum a posteriori (MAP) estimate could be used to speed up the convergence. In this case, a simulated annealing approach is very helpful to decrease the jump probability gradually and find the global maximum. The procedure is to modify the acceptance ratio as

$$\alpha = \min\left\{\left[\frac{p(X|\theta^*)\pi(\theta^*)}{p(X|\theta_{t-1})\pi(\theta_{t-1})}\right]^{1/T(t)}, 1\right\}$$

where the function $T(t)$ is called the cooling schedule which can be chosen as an exponential decay function of t, such as $T(t) = p^t, p < 1$ (e.g, p=0.999).

- *Step 4*: Generate a random number u from $U(0,1)$, if $u < \alpha$, set $\theta_t = \theta^*$; otherwise set $\theta_t = \theta_{t-1}$.
- *Step 5*: Define a burn-in period, where the elements are excluded, and calculate the posterior estimates based on the remaining elements of the simulated chain.

The two commonly used rules to choose the Bayesian estimate are the MAP and the minimum mean square error (MMSE), which are respectively the posterior mode and mean. In this study, when a weak annealing rate is used, the mean of the posterior could be treated as an estimate weighted between MMSE and MAP. In Section 4.3, we also discuss the effect of choosing MMSE or MAP on the parameter estimates.

An appealing feature of the MCMC approach is that it offers the flexibility to impose constraints on the parameters. When using numerical optimization, adding constraints often results in hitting the bounds and, as a consequence, the output parameters may not be the optimum. The MCMC approach allows the parameters to traverse the entire specified parameter space. For example, in this study, when the truncated area exceeds $1 - 10^{-3}$, due to the rounding error during calculation, the Jeffreys' prior may become imaginary. Therefore, the parameters which lead to unreasonable truncated area need to be excluded to avoid numerical error.

4.4 Simulation study

4.4.1 Stability of estimates

We first simulate a random sample of size 100 from the LN(10,2) distribution with truncation above 20000, and estimate the parameters of μ and σ using the MLE and Bayesian methods. In order to test the stability of the two methods, for each sample size from 2 to 100, we randomly draw a subsample of this size from the simulated sample, and estimate the parameters for the subsample.

From Figure 4.4, we can see that the MLE method exhibits a large variance when the sample size is small and may produce unrealistic estimates (see the negative location parameters and the consequent enormous scale parameters). For an analytical explanation of the sensitivity of parameters to data contamination under various truncation thresholds, see Opdyke and Cavallo (2012). In contrast, the Bayesian method is much less sensitive to the sample size and produces more stable estimates than the MLE method.

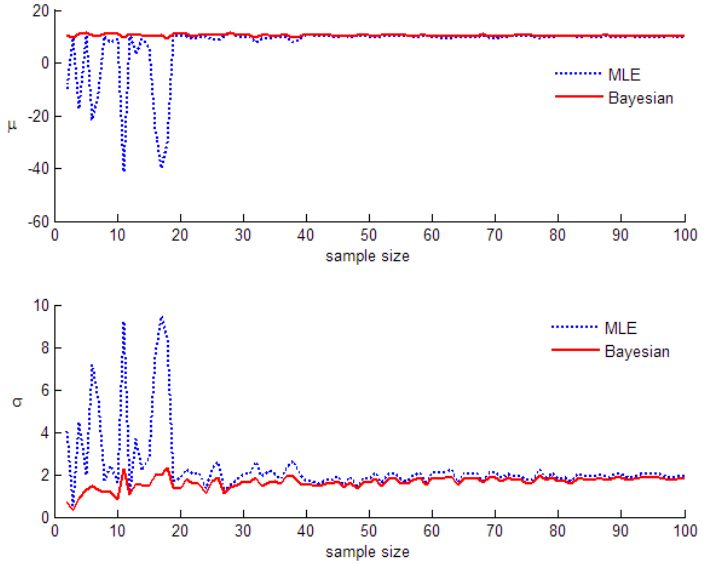


Figure 4.4: Stability of estimates by MLE and Bayesian methods

To further explore this phenomenon, we repeat the above experiment 1000 times, and plot the 5%, 20%, 50%, 80% and 95% quantiles of the parameter estimates for each sample size. From the results, as shown in Figure 4.5, we see that the Bayesian method greatly reduces the variance of the estimates and provides much more stable and reliable estimates at the expense of a small bias. When the sample size increases, the Bayesian estimates converge to the ML estimates.

Moreover, in risk measurement, one may also be interested in the variability of VaR estimates based on the two approaches. Figure 4.6 shows the boxplot of the logarithm of 99.9% VaR of the loss severity distribution at 10 different sample sizes (10,20,...,100). Clearly, the Bayesian method produces estimates that have much less variation and are more reliable.

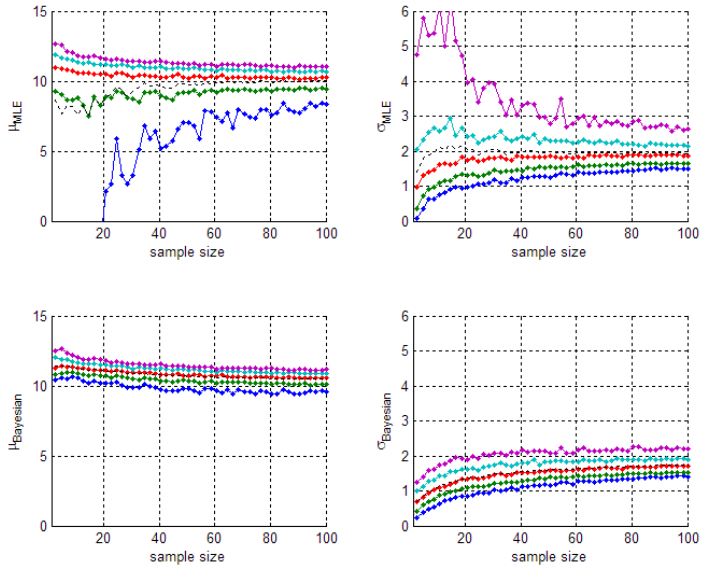


Figure 4.5: Quantiles of parameter estimates by MLE and Bayesian methods

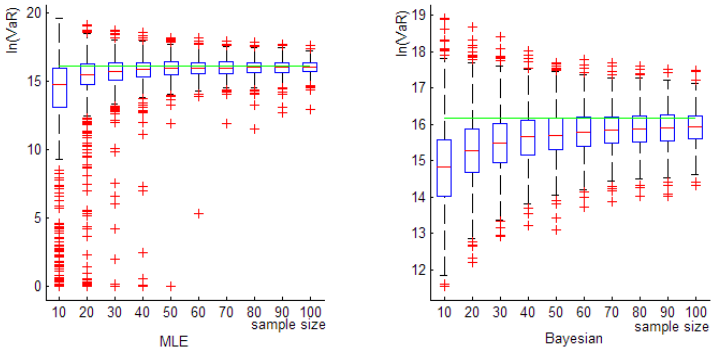


Figure 4.6: Boxplot of ln(VaR) by MLE and Bayesian methods

4.4.2 Bias and variance of estimates

To further evaluate the performance of Jeffreys' prior estimate, we investigate the squared bias, variance, and mean-squared error (MSE) for different sample sizes and different truncation levels from the $LN(10,2)$ distribution. For each case, 1000 independent samples are generated to calculate the average squared bias, variance, and MSE of the four estimation methods: MLE, EM algorithm, PLE, and Bayesian estimate. Table 4.1 and Table 4.2 report the result of each experiment: sample size N ranging from 20, 50, 100 to 500, truncation level p from 25%, 50%, 75% to 90%.

For the MLE method, the standard Matlab optimization toolbox with default settings of termination rules is applied. For the EM algorithm, we terminate the iteration if adjacent parameters' difference is less than 10^{-5} or the likelihood difference is less than 10^{-6} within at

most 20,000 iteration steps. For the PLE method, the penalty term is approximately chosen as $\nabla J(\theta_0)v/2\lambda$ without knowing the true parameter but based on the parameters estimated by the MLE method. In the Bayesian estimation, Jeffreys' prior for the truncated normal is used. The length of the chain is 10000 and the burn-in period is the first 1000 samples. The estimated parameter is chosen as the mean of the posterior distribution. Notice that to accelerate the simulation we apply the simulated annealing with a weak decay factor 0.9995. Therefore, the estimates generated are in fact neither an MMSE nor MAP estimate. (When the decay factor is 1, the estimate is the MMSE; and when the decay factor is smaller, the estimate becomes closer to MAP.)

Compared to the MLE method, the Bayesian method provides a considerable reduction for the MSE for all sample sizes and at all truncated levels considered. (See the "Redux%" column, in which we partly follow the format of showing simulation results by Cope (2011), but we list the bias, variance, and MSE for both μ and σ). On a relative basis, the EM algorithm provides the least bias with a greater variance than the Bayesian method. The PLE method also performs well, but at a high truncation level, the estimates become worse than that obtained from the MLE method because of a much higher bias. In addition, when the sample size is small, the PLE method fails to produce an estimate for some of the samples, and we removed these failures when evaluating the results for the PLE method. Perhaps this explains why the bias and MSE are higher than the results shown in Cope (2011).[1]

4.4.3 Selection of priors and criteria

From the findings reported above, we can see that Jeffreys' prior asymptotically approaches the MLE when the sample size increases, suggesting that this method generates estimates that are consistent. However, in the presence of a high truncation level, especially when there is a small sample size, the estimates using Jeffreys' prior will exhibit a large bias and small variance, which indicates that the bias is unidirectional (see the results in Table 4.1 at the 90% truncation level with the sample size of 20). One may wonder if this is attributable to the prior distribution or the criterion used to select the estimate. In this section, we investigate these two issues based on a simulation study. Specifically, we concentrate on the high truncation level and small sample size case, where the estimates obtained from the Bayesian method are highly biased.

We begin by discussing the two different criteria used to choose posterior estimates by evaluating the bias, variance, and MSE. For the experiment in Section 4.2, to save the number of computation steps, we incorporated the simulated annealing (SA) into our MCMC to decrease the transition probability. This may be the primary reason that the variance of the estimates becomes much smaller. Therefore, in order to investigate the effect of these two criteria clearly, we present the MMSE estimates without SA as well as the MAP estimates.

Then we discuss the sensitivity of the estimates to the selection of various priors. Apparently, given more accurate prior information about the true parameters, the posterior estimates would be much better; therefore, we do not discuss the elicited priors. Instead, we discuss vague priors, Jeffreys' priors, and no prior (or constant prior).

[1] In a private communication with Professor Eric Cope, he pointed out that the PLE method may be challenging when the likelihood surface is very flat and the Hessian matrix is ill-conditioned, and working in a transformed variable space may increase the convexity of the Hessian. However, we did not investigate this issue further in this paper because our main goal focus was on the Bayesian approach.

Table 4.1: Squared bias, Variance and MSE of estimates by MLE and Bayesian

			MLE			Bayesian			
F(u)%	*N*		Bias^2	Var	MSE	Bias^2	Var	MSE	Redux(%)
25	20	μ	4.895	50.882	55.726	0.146	0.325	0.471	99.15
25	20	σ	0.228	1.837	2.064	0.112	0.144	0.256	87.61
25	50	μ	0.179	4.813	4.987	0.036	0.259	0.295	94.10
25	50	σ	0.010	0.285	0.295	0.031	0.088	0.119	59.78
25	100	μ	0.047	0.377	0.424	0.001	0.198	0.198	53.22
25	100	σ	0.005	0.099	0.104	0.003	0.064	0.067	35.26
25	500	μ	0.001	0.055	0.056	0.000	0.053	0.053	5.97
25	500	σ	0.000	0.019	0.019	0.000	0.018	0.018	3.55
50	20	μ	20.973	112.140	133.000	0.897	0.426	1.323	99.01
50	20	σ	0.438	3.034	3.469	0.300	0.138	0.439	87.36
50	50	μ	1.789	20.817	22.585	0.240	0.574	0.814	96.40
50	50	σ	0.050	0.859	0.909	0.082	0.119	0.200	77.96
50	100	μ	0.307	4.414	4.716	0.018	0.674	0.692	85.33
50	100	σ	0.012	0.305	0.317	0.015	0.106	0.121	61.90
50	500	μ	0.009	0.236	0.245	0.003	0.215	0.218	11.06
50	500	σ	0.001	0.036	0.037	0.000	0.034	0.034	9.15
75	20	μ	25.794	107.360	133.040	3.249	0.400	3.648	97.26
75	20	σ	0.296	2.640	2.933	0.587	0.118	0.705	75.96
75	50	μ	7.858	54.738	62.540	1.226	0.822	2.047	96.73
75	50	σ	0.121	1.530	1.649	0.203	0.134	0.337	79.54
75	100	μ	2.295	21.081	23.355	0.267	1.006	1.272	94.55
75	100	σ	0.054	0.710	0.764	0.049	0.113	0.162	78.82
75	500	μ	0.027	0.819	0.845	0.002	0.568	0.570	32.56
75	500	σ	0.001	0.064	0.065	0.000	0.049	0.049	24.79
90	20	μ	29.580	98.359	127.840	7.139	0.456	7.595	94.06
90	20	σ	0.281	2.380	2.658	0.765	0.124	0.890	66.54
90	50	μ	14.486	62.981	77.404	3.631	0.845	4.475	94.22
90	50	σ	0.171	1.563	1.732	0.349	0.115	0.464	73.23
90	100	μ	4.994	35.279	40.238	1.672	1.128	2.799	93.05
90	100	σ	0.062	0.910	0.971	0.156	0.102	0.258	73.42
90	500	μ	0.205	3.917	4.118	0.026	1.123	1.148	72.13
90	500	σ	0.004	0.157	0.161	0.004	0.066	0.071	56.23

Table 4.2: Squared bias, Variance and MSE of estimates by EM and PLE

			EM				PLE			
F(u)%	*N*		Bias^2	Var	MSE	Redux(%)	Bias^2	Var	MSE	Redux(%)
25	20	μ	0.150	5.318	5.463	90.20	0.451	0.176	0.627	98.88
25	20	σ	0.000	0.629	0.629	69.53	0.147	0.106	0.252	87.77
25	50	μ	0.008	0.832	0.839	83.18	0.120	0.094	0.214	95.70
25	50	σ	0.000	0.176	0.176	40.15	0.045	0.054	0.099	66.54
25	100	μ	0.007	0.303	0.310	26.74	0.014	0.086	0.100	76.37
25	100	σ	0.000	0.089	0.089	14.42	0.007	0.041	0.048	54.01
25	500	μ	0.000	0.053	0.053	5.34	0.000	0.046	0.046	18.69
25	500	σ	0.000	0.019	0.019	1.96	0.000	0.017	0.017	11.50
50	20	μ	0.643	10.102	10.735	91.93	2.269	0.115	2.384	98.21
50	20	σ	0.003	0.893	0.895	74.20	0.501	0.059	0.560	83.85
50	50	μ	0.130	3.914	4.040	82.11	1.426	0.069	1.494	93.38
50	50	σ	0.000	0.394	0.394	56.61	0.290	0.029	0.319	64.92
50	100	μ	0.065	1.863	1.927	59.14	0.760	0.049	0.808	82.86
50	100	σ	0.001	0.211	0.212	33.22	0.154	0.021	0.175	44.73
50	500	μ	0.003	0.223	0.226	7.57	0.057	0.053	0.111	54.87
50	500	σ	0.000	0.035	0.035	4.11	0.011	0.013	0.024	35.17
75	20	μ	0.138	9.483	9.611	92.78	6.149	0.080	6.229	95.32
75	20	σ	0.014	0.742	0.754	74.28	0.989	0.043	1.032	64.82
75	50	μ	0.142	6.396	6.531	89.56	5.240	0.065	5.305	91.52
75	50	σ	0.001	0.491	0.491	70.21	0.746	0.021	0.768	53.45
75	100	μ	0.199	4.136	4.331	81.45	4.174	0.055	4.228	81.90
75	100	σ	0.003	0.302	0.304	60.16	0.535	0.015	0.550	27.98
75	500	μ	0.010	0.718	0.727	14.02	1.338	0.012	1.349	-59.66
75	500	σ	0.000	0.060	0.060	7.36	0.149	0.005	0.154	-137.49
90	20	μ	0.003	9.488	9.482	92.58	12.642	0.088	12.730	90.04
90	20	σ	0.024	0.717	0.740	72.16	1.356	0.048	1.404	47.18
90	50	μ	0.002	6.224	6.219	91.97	11.030	0.131	11.161	85.58
90	50	σ	0.010	0.410	0.419	75.80	1.117	0.026	1.143	34.00
90	100	μ	0.000	4.141	4.137	89.72	10.005	0.050	10.055	75.01
90	100	σ	0.006	0.256	0.261	73.09	0.963	0.011	0.974	-0.34
90	500	μ	0.007	1.497	1.503	63.50	5.849	0.021	5.870	-42.55
90	500	σ	0.000	0.086	0.086	46.62	0.486	0.005	0.491	-204.05

Table 4.3: The effect of various priors and criteria

Combination	$\text{Bias}^2(\hat{\mu})$	$\text{Var}(\hat{\mu})$	$\text{MSE}(\hat{\mu})$	$\text{Bias}^2(\hat{\sigma})$	$\text{Var}(\hat{\sigma})$	$\text{MSE}(\hat{\sigma})$
No Prior, MMSE	6.16	1.33	7.49	0.30	0.26	0.56
No Prior, MAP	2.22	18.02	20.23	0.00	0.91	0.91
Jeffreys' Prior, MMSE	1.17	1.27	2.44	0.16	0.21	0.37
Jeffreys' Prior, MAP	7.83	0.32	8.15	0.94	0.09	1.04
Vague Prior, MMSE	1.22	2.36	3.58	0.02	0.31	0.33
Vague Prior, MAP	4.59	1.44	6.03	0.60	0.18	0.78

To be consistent with the experiments performed earlier, we continue to focus on the lognormal distribution. By simulating 1000 samples of size 20 from the truncated $LN(10,2)$ distribution with truncation point at the 90th percentile, we estimate the parameters using the following combinations of the prior distributions and criteria for choosing estimates:

1. No prior information (or constant prior) and MMSE;
2. No prior information and MAP;
3. Jeffreys' prior and MMSE;
4. Jeffreys' prior and MAP;
5. Vague prior ($\mu \sim N(0, 100)$, $\sigma^2 \sim IG(0.01, 0.01)$) and MMSE;
6. Vague prior and MAP.

Notice that the second combination, as a special case in our Bayesian estimation study, is in fact a simulated annealing approach to solve the MLE, which can perform better than using the Hessian-matrix based algorithms because it allows the parameters to jump out of the local optima. For the specification of the vague priors, we follow the tradition that an inverse gamma distribution with small scale and shape parameter is used as the prior for the variance, and a flat and wide normal distribution is used as the prior for the mean.

The bias, variance, and MSE of the estimates for the above six combinations are reported in Table 4. 3. We see that at this high truncation level (90%) and for this small sample size (20), using either Jeffreys' prior or vague prior, the MMSE estimates greatly reduces the bias and MSE at the expense of a small increase in variance compared to MAP estimates. The difference between the Jeffreys' prior and vague prior is not so significant. However, both of them substantially outperform the no prior information case. This finding suggests that the Bayesian estimation is not so sensitive to the prior distribution selected once it is properly specified. The significance of using Jeffreys' prior lies in its invariance of noninformative property under different parametrization, while the vague prior does not have this property and may have a potentially strong impact on the posterior estimates.

In the above experiment, we have shown that the MMSE estimates reduces the bias at the high truncation level. As a further comparison, we investigate the performance in the case of a medium truncation level. To do so, we repeat the second experiment in Section 4.1: simulate 1000 samples from $LN(10,2)$ with truncation at 20000 (or 48th percentile) for each sample size from 2 to 100, and plot the 5%, 20%, 50%, 80%, and 95% quantiles of the MMSE estimates. The results are reported in Figure 4.7. In comparison to the MAP estimates which exhibit a very small variance but an obvious bias, the MMSE estimates greatly reduce the bias at the expense of a greater variance. However, when there is a small

sample size, no matter which criterion is chosen, the estimates generated by the Bayesian method are more stable than those generated by the MLE method.

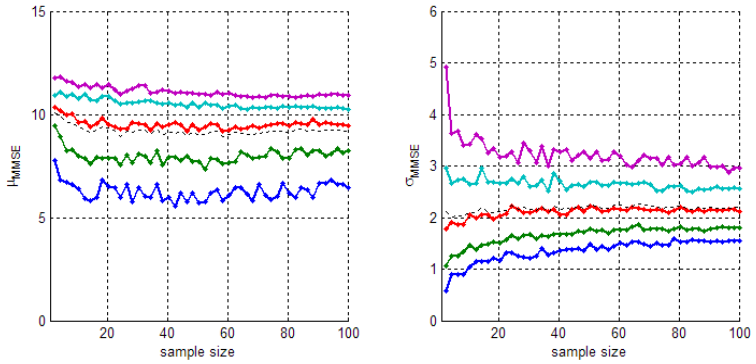


Figure 4.7: Quantiles of MMSE estimates

4.5 Empirical study

The Basel II Accord requires that banks calculate operational risk capital charges. According to the LDA, the loss frequency and severity distributions for each intersection of the seven event types and eight business lines need to be modelled. Without considering the time dependence, the loss frequency distribution is usually assumed to be a Poisson distribution and the loss intensity parameter is easily estimated from the average number of losses.

However, when estimating the parameters for the loss severity distribution, data insufficiency and the reporting threshold have become two main issues in using the usual MLE method because the estimates produced often suffer from substantial instability and may appear to be highly unrealistic. As demonstrated by the simulation studies in Section 4, for truncated data, especially with a small sample size, applying the Bayesian method would greatly relieve the difficulty in parameter estimation, and improve the credibility of the produced parameters and VaR measure.

Using real data, in this section we provide an empirical study of applying the Bayesian method to estimate the loss severity distribution parameters and the operational risk capital charge. The data used in this study consist of two datasets: internal losses from a European bank (internal data) and external consortium data (external data). Both of the datasets are provided by the same bank.[2]

According to Basel II, external data are required to be incorporated with internal data to evaluate a bank's exposure to high-severity events. Consequently, many approaches have been proposed to combine the two datasets. Typically, the weights of data points in the external dataset need to be adjusted (see Baud et al (2002) and Aue and Kalkbrener (2006)), or the sizes of the loss amount need to be scaled by some exposure factors such as firm size and geographical region (see Dahen and Dionne (2007) and Cope and Labbi (2008)). Nevertheless, the first approach relies on the so-called "fair mixing assumption", which may not be valid, and the second approach is based on the existence of the scaling relationships, and more importantly, the availability of exposure factors data.

As we do not have any further exposure factors information, and do not know about the validity of the "fair mixing assumption", we attempt to deal with the internal data and

[2] For confidentiality issues, we do not disclose any specific information about the bank.

external data separately without risking the naive combining of them. We suggest that the external data should be used as a reference when estimating the capital charge, but not be pooled directly with the internal data because it might mask the effect of the original internal losses. Moreover, in most of the cells where there is a lack of internal losses, the same applies to the external data as well. There is little benefit of pooling them together when the pooled data still are scarce.

Specifically, in this section, we investigate the loss severity distribution for each cell with observations for both the internal and external data, estimate the parameters using the Bayesian method, calculate the 99.9% VaR for a single loss event and the 99.9% VaR for the compound losses over a one-year period (as required by Basel II), and compare the risk profile for internal and external data in each cell.

4.5.1 Internal data

The internal data consist of the reported losses of a European bank for the period January 2003 to August 2005. For this period, there were 1850 observations. The reporting threshold is €500. The losses are categorized as required by Basel II into seven event types and eight business lines. Due to confidentiality issues, we assign a random code to each intersection. Most studies only focus on the cells with "enough" data. However, the cells with few data but high loss amounts are very important for measuring a bank's total risk profile. The total number of losses and loss amount for each cell are reported in Table 4. 4. Using the MLE, EM, PLE, and Bayesian methods, we estimated the parameters for each cell with loss data using the internal data only under the assumption of truncated lognormal distributions.

Table 4. 4 lists the estimated parameters by the four methods. We can see that when the MLE method produces negative location parameter estimates, the PLE method is likely to fail as well. Compared to the MLE method, the EM algorithm provides higher location parameter estimates, which looks more reasonable. However, for several cells, the location parameters are still negative. The Bayesian method is the most reliable approach and always provides reasonable parameter estimates, with the exception of cell 2. In this cell, same as the EM and MLE methods, the Bayesian method produces a negative location parameter as well. The estimated truncated area is also extremely high (0.98). Because the sample size is not small, this finding is probably due to a misspecification of the distribution because of a very heavy tail. (Although not shown here, we confirmed this by looking at the histogram for the losses in this cell.) However, as in the LDA, the underestimated severity will be compensated by adjusting the frequency by multiplying $\frac{1}{1-F(u)}$.Therefore, these misspecifications will not have much of an impact on the estimation of the compound losses. In fact, one will never know how much is truncated below the threshold. There is no way to test whether the estimated truncated area is correct or not. However, these extremely high truncations should be avoided because in practice it is unrealistic that almost 100% of the losses are below the threshold.

Various tests of quantiles and empirical distribution function tests are often suggested as goodness-of-fit tests for testing an hypothesized distribution. However, these tests are limited to situations when there are enough data. Since in most of the cells there are not enough observations, goodness-of-fit testing methods provide little meaning. Thus, the selection of the distribution type is mainly determined by the cells with sufficient data. We provide the fitting performance of the lognormal distribution to all the internal loss observations (see Figure 4.8). The left plot shows the fitted cumulative distribution function (cdf) and the empirical cdf, and the right plot shows the fitted histogram in a logarithm

Table 4.4: Estimated lognormal parameters by MLE, EM, PLE and Bayesian methods

Cell	N	Amount	MLE		EM		PLE		Bayesian		$F(u)$
1	100	1.17E+06	3.04	3.23	4.25	2.93	8.27	1.59	5.49	2.50	0.63
2	270	2.57E+06	-17.87	6.02	-2.56	3.92	7.26	1.55	-1.70	3.75	0.98
3	1	1032.5	/	/	/	/	/	/	6.75	0.25	0.02
4	3	21943	-31.84	7.13	1.47	3.55	/	/	7.87	1.16	0.09
5	64	1.16E+06	-1.80	4.44	0.68	3.95	/	/	4.77	2.84	0.73
6	48	2.52E+06	8.42	2.35	8.46	2.32	8.48	2.31	8.64	2.17	0.14
7	68	2.60E+05	7.41	1.17	7.44	1.15	7.45	1.15	7.47	1.13	0.13
8	7	47309	-33.07	6.44	-4.56	3.71	/	/	5.75	1.71	0.61
9	10	1.11E+05	7.95	1.76	8.34	1.55	8.47	1.47	8.53	1.29	0.04
10	1	4.01E+05	/	/	/	/	/	/	11.11	2.20	0.01
11	76	5.25E+05	-7.72	4.67	-1.67	3.71	7.80	1.43	3.57	2.57	0.85
12	25	9.03E+05	8.56	2.26	8.59	2.24	8.59	2.24	8.86	1.99	0.09
13	9	27379	7.25	1.12	7.42	1.03	7.51	0.97	7.60	0.83	0.05
14	5	6898.7	7.06	0.43	7.17	0.35	7.17	0.35	7.08	0.41	0.02
15	17	2.92E+05	-34.53	8.10	-6.27	4.93	/	/	5.45	2.39	0.64
16	34	2.41E+06	7.76	2.91	8.21	2.71	8.63	2.49	8.63	2.41	0.16
17	709	6.15E+06	6.96	1.92	6.98	1.91	7.02	1.89	6.99	1.90	0.35
18	2	2.97E+05	-65.14	15.77	-10.77	8.59	/	/	9.67	2.20	0.06
19	5	3.30E+06	-82.15	18.80	-12.63	9.84	/	/	9.42	3.28	0.17
20	145	2.52E+06	7.55	1.82	7.61	1.79	7.64	1.77	7.66	1.76	0.21
21	7	88140	-33.87	8.27	5.11	2.92	/	/	7.91	1.55	0.15
22	1	39785	/	/	/	/	/	/	9.43	1.47	0.02
23	7	3.63E+05	9.32	1.93	9.62	1.72	9.57	1.75	9.60	1.60	0.02
24	3	62270	-35.29	9.22	5.07	3.34	/	/	8.55	1.43	0.06
25	224	4.06E+06	8.38	1.54	8.38	1.54	8.38	1.54	8.38	1.54	0.08
26	3	37570	4.94	2.64	8.71	1.29	/	/	8.62	1.14	0.02
Total	1850	2.94E+07	6.08	2.43	6.08	2.43	6.19	2.39	6.07	2.44	0.52

Note: The two columns in each method are $\hat{\mu}$ and $\hat{\sigma}$, and "/" denotes that the method fails to produce an estimate.

scale. It can be seen that the lognormal distribution provides a good overall fit as well as realistic estimates in the tails. For the cells with a reasonable amount of data, we also tested their lognormal fitting performance and found that the lognormal distribution fits well for almost all of them. Notice that it could not provide perfect fitting to every cell so as not to be rejected by the goodness-of-fit tests because of the relatively small sample size and the data collection issues such as rounding error, repeatedly occurred losses with the same amount, and outlier losses. Generally, the lognormal distribution provides a good overall fit, as well as generating reasonable tail estimates.

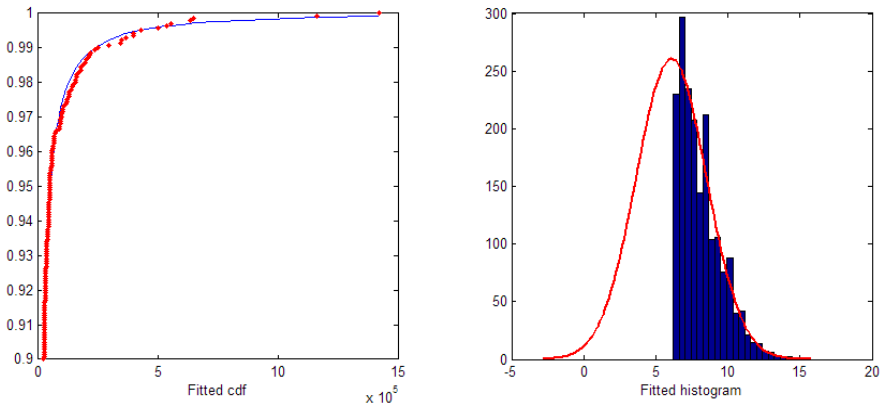


Figure 4.8: Fitted lognormal cdf and histogram for all internal data

4.5.2 External data

The external data include 14824 losses from a consortium of banks for the period January 2003 to June 2005, with a threshold of €5000. As the lognormal distribution provides a good overall fitting for the internal data, we fit the lognormal distribution to the external data as well. However, as shown in Figure 4.9, the lognormal distribution clearly underestimates the tail. This suggests that the external data has a heavier tail than the internal data.

After trying several alternative distributions — including the Weibull, log-Weibull, gamma, loggamma, alpha-stable, and extreme value distributions — we found that the loggamma distribution is relatively flexible and easy to estimate. It does not require one to deal with the tail and body separately as needed to apply the extreme value approach. From the fitting performance shown in Figure 4.9, we observe that the loggamma distribution provides a much better fitting than the lognormal distribution.

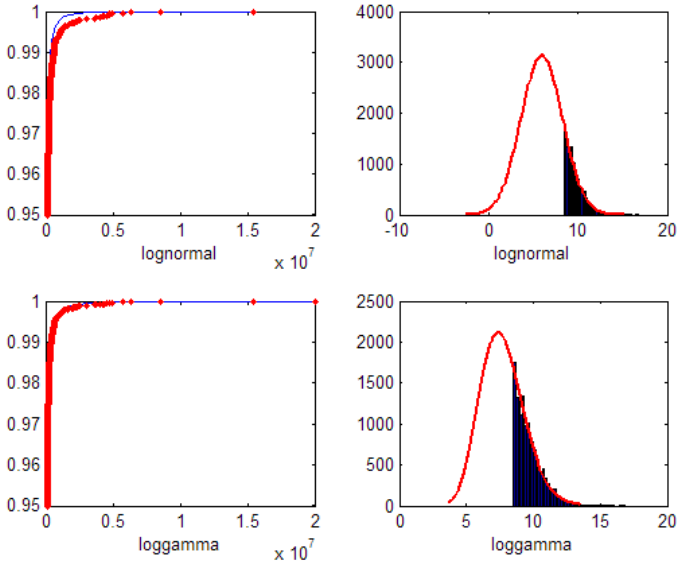


Figure 4.9: Fitted lognormal and loggamma cdf and histogram for all external data

Without making a hasty decision — using loggamma distribution for all the cells of external data — we looked into the data in each cell. We found that the effect that the lognormal distribution underestimates the tails for the external data is mainly caused by the losses in cell 12. A comparison of the fitted lognormal cdf and loggamma cdf for this cell is shown in Figure 4.10. We see that the lognormal distribution deviates substantially from the empirical cdf starting from the 80-th percentile. In contrast, using the loggamma distribution provides a much better fitting.

We also investigate the performance of the loggamma distribution for the other cells, and find that it too provides good fitting to the other cells with a little bit higher estimate in the tails. Therefore, we tend to believe that the loggamma distribution is more flexible than the lognormal distribution because it can capture the heavier tails, but at the same time it also adapts to the lighter tails where the lognormal distribution provides a satisfactory fitting. One may argue that fitting different distributions to the various cells gives better in-sample fittings. Nevertheless, we would prefer a more universal and robust distribution that applies to every cell with a relatively satisfactory fitting.

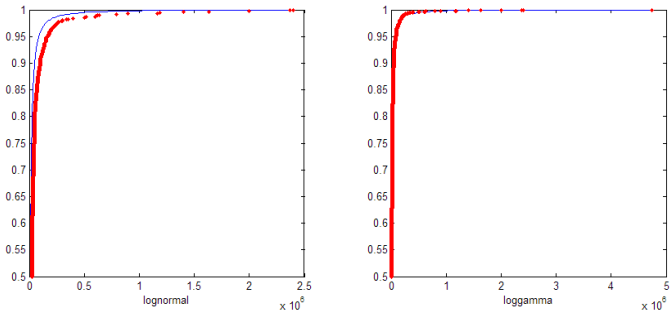


Figure 4.10: Fitted lognormal and loggamma cdf for cell 12

Table 4.5: Estimated loggamma parameters for external data by MLE and Bayesian methods

				MLE				Bayesian			
Cell	*N(In)*	*N(Ex)*	Amount	$\hat{\kappa}$	$\hat{\lambda}$	mean	*F(u)*	$\hat{\kappa}$	$\hat{\lambda}$	mean	*F(u)*
1	100	449	1.52E+07	18.85	0.38	7.16	0.80	20.63	0.37	7.55	0.74
2	270	1423	4.90E+07	6.44	0.66	4.25	0.98	7.41	0.63	4.69	0.97
3	1	1	5.71E+03	/	/	/	/	218.37	0.04	9.29	0.11
4	3	108	1.04E+07	61.91	0.17	10.52	0.06	66.84	0.16	10.78	0.04
5	64	53	1.59E+06	/	/	/	/	69.09	0.14	9.72	0.15
6	48	147	1.07E+07	25.65	0.35	8.98	0.42	27.63	0.34	9.46	0.32
7	68	472	9.07E+06	94.63	0.10	9.46	0.17	94.90	0.10	9.33	0.20
8	7	71	4.56E+06	0.01	1.37	*0.01*	*1.00*	6.82	0.75	5.11	0.94
9	10	244	9.70E+06	0.13	1.17	*0.15*	*1.00*	4.12	0.82	3.39	0.99
10	1	82	4.68E+07	36.57	0.31	11.34	0.06	36.04	0.32	11.43	0.05
11	76	490	1.59E+07	17.68	0.40	7.07	0.81	18.97	0.39	7.37	0.76
12	25	3447	9.80E+07	17.91	0.38	6.81	0.86	17.69	0.38	6.76	0.86
13	9	9	1.51E+05	1.10	0.62	*0.68*	*1.00*	69.66	0.13	9.00	0.34
14	5	75	2.79E+06	46.18	0.20	9.24	0.31	49.21	0.19	9.50	0.24
15	17	373	1.40E+07	16.46	0.42	6.91	0.83	17.46	0.42	7.26	0.78
16	34	209	4.83E+07	21.65	0.46	9.96	0.26	23.03	0.44	10.20	0.22
17	709	5580	1.61E+08	47.64	0.19	9.05	0.36	49.91	0.18	8.91	0.39
18	2	5	9.59E+04	1.68	0.64	*1.08*	*1.00*	153.13	0.07	11.20	0.00
19	5	67	2.97E+06	9.80	0.59	5.78	0.92	21.96	0.36	7.96	0.65
20	145	654	4.11E+07	32.04	0.29	9.29	0.34	32.39	0.29	9.37	0.32
21	7	25	2.25E+05	1.44	0.49	*0.71*	*1.00*	186.85	0.05	8.93	0.27
22	1	2	1.97E+04	/	/	/	/	391.68	0.02	9.46	0.02
23	7	15	6.42E+05	65.85	0.15	9.88	0.13	93.74	0.11	10.56	0.02
24	3	4	4.63E+05	26.50	0.36	9.54	0.31	95.81	0.12	11.43	0.00
25	224	563	2.77E+07	/	/	/	/	14.87	0.47	7.04	0.80
26	3	33	6.76E+06	0.03	1.74	*0.05*	*1.00*	13.19	0.59	7.78	0.66
Total	1850	14824	6.10E+08	22.21	0.35	7.77	0.69	22.77	0.34	7.74	0.70

Note: "/" denotes that the method fails to produce an estimate.

To be conservative, we fit both the lognormal and loggamma distribution for the external data because each of them has its advantage. The former fits very well for most of the cells but fails seriously for cell 12, while the latter slightly overestimates the tails for most of the cells but is capable of capturing varying tail conditions. Using the Jeffreys' prior for the truncated gamma distribution provided in Section 3, we estimate the parameters for the loggamma distribution. To compare the external VaR with internal VaR (shown in Section 5.3), here we only focus on the cells that have internal losses.

Table 4. 5 lists the loggamma parameters estimated by the MLE and Bayesian methods. The estimated mean of the complete distribution is the product of the two estimated parameters. By comparing the estimated mean for the full distribution and the truncated proportion, it is found that the MLE method would likely underestimate the mean and overestimate the truncated area. In contrast, the Bayesian method provides a more reasonable estimate for the mean for every cell, suggesting that it is more reliable. (To save space, a comparison of the MLE and Bayesian estimates for the lognormal distribution is not reported here because the comparison is already shown for the internal data. However, the lognormal distribution parameters estimated by the Bayesian method are presented in Table 4. 7 in Section 5.3.)

4.5.3 VaR and confidence interval estimates

In this subsection, we present the calculation of VaR and show how to obtain the confidence intervals for the estimates.

After obtaining the loss severity distribution, the 99.9% VaR of each loss event can be simply obtained by taking the 99.9% quantile. The estimated parameters of loss severity distributions and 99.9% VaR of a single event for internal data and external data are given in Tables 4.6, 4.7, and 4.8. The lognormal distribution is used for the internal data, while for the external data we investigate both the loggamma distribution and lognormal distribution.

To obtain an estimate for the compound operational VaR, such as the 99.9% VaR for the losses over a one-year period, several approximation methods have been proposed such as Panjer recursion, fast Fourier transform, and closed-form approximations. (See McNeil et al (2005), Shevchenko (2010), and Bocker and Kluppelberg (2005).) However, these methods are either difficult to implement or not robust to different tails, so we employ Monte Carlo simulation to obtain the estimates of VaR.

Assuming that there are on average λ losses reported in one period, a Poisson distribution is often assumed for the loss frequency distribution. Then, there are usually two approaches for simulating a compound loss scenario:

- *Approach 1*: Simulate a number N from the Poisson distribution with the adjusted loss intensity $\frac{\lambda}{1-\widehat{F}(u)}$, and simulate N losses from the unconditional distribution $F(x)$ and then aggregate the N losses as the total loss in one period;
- *Approach 2*: Simulate a number N from the Poisson distribution with the unadjusted loss intensity λ, and simulate N losses from the conditional distribution $F_u(x) = \frac{F(x)-F(u)}{1-F(u)}$and then aggregate the N losses as a scenario for the compound loss in one period.

Although theoretically the first approach leads to the true compound distribution, there are two problems with this approach. First, potential model risk may be introduced when the estimated truncation area $\widehat{F}(u)$ is very large. For example, if $\widehat{F}(u) > 0.999$, when using the unconditional distribution, the 99.9% VaR for the compound losses may become less than the observed losses. This is unacceptable. Second, a large number of small losses below the threshold may be unnecessarily generated. For example, when $\widehat{F}(u)$ is very large, the losses above the threshold account for only a very small region of the full distribution. Then a huge amount of small losses are generated during the simulation, but become useless for the estimation of VaR which is only concerned with the tails. Moreover, because the adjusted loss frequency is very large, it becomes difficult via simulation to obtain an accurate estimate of VaR.

Therefore, we employ the second approach in this paper. Notice that it uses the conditional distribution when computing the 99.9% VaR, so it may produce biased estimates compared to the full distribution approach. However, we found the bias not to be significant based on several simulation experiments. Moreover, this approach has several advantages. First, because it only simulates the losses above the threshold, the estimated VaR can be tested against the real losses, which are also only known above the threshold. Second, it avoids the simulation of small losses and reduces the computational effort. Third, it is robust to misspecification error for the distribution below the threshold because the estimated VaR is not affected by the shape of the distribution below the threshold. This is important because one may never know about the true distribution below the truncation point. Therefore, it is better to focus on the region above the threshold.

After obtaining the point estimates for the loss distribution parameters and VaR measures, we investigate their confidence intervals by applying a bootstrap method. The process is as follows. From the estimated distribution, simulate a number of samples of the same size as the original sample and then re-estimate the parameters for each sample. Hence, the confidence interval of any estimates can be obtained from the bootstrapped distribution of the parameters.

By bootstrapping 200 samples, we estimate the 90% confidence interval (left 5% to right 5%) of the parameters. Then the 90% confidence interval of 99.9% VaR for each single event is also calculated. The standard deviation of the estimated parameters and the 90% confidence intervals of the 99.9% VaR of a single event and 99.9% annual VaR are obtained by the bootstrap method (see Tables 4.6, 4.7, and 4.8).

In the following, we present the analysis and main findings from the estimation results shown in the three tables.

The comparison of internal VaR and external VaR By computing the correlation coefficient of the 99.9% VaR (by taking the logarithm to base 10) of each single loss for the internal and external data, we find a significant correlation close to 0.5 between them (see Figure 4.11, where to be intuitive the VaRs shown are logarithm to base 10 in this figure and the others to follow). This means that the external losses do play a reference role for the internal losses. However, it may also be noted that in some special cases, the internal losses and external losses can be totally different. For example, in cell 19 (see Table 4. 6), the average loss amount by single event for the internal data is 6.6×10^5, and the calculated VaR is 3.07×10^8, which may be the most important loss events for the bank. However, from the external losses for this cell (see Table 4. 7), the average amount is only 4.4×10^4. When pooling the internal data and external data without consideration, the estimated risk profile will be totally different because the observations from external losses will dominate the estimation of the severity distribution.

In Figure 4.12, we compare the internal VaR and external VaR for each cell (sorted increasingly by the number of internal losses). It is observed that when the internal and external data both have many loss observations, the estimated VaR are very close to each other. This suggests that the internal and external data may have the same profile for the frequently occurred loss events, which coincides with the "fair mixing assumption" by Baud et al (2002). However, when the number of internal losses is small, the estimated VaR for internal and external data are usually different. There are two possible explanations for this phenomenon. One is that the internal losses in these cells may exhibit a different profile from the external losses. The other explanation is that in the small sample cases, the estimated VaR is more volatile. The estimated VaR of external losses could be used as a reference to adjust the internal VaR for the cells where more external loss observations are possible. When both the internal and external data are scarce, the latter are also helpful for one to know more about the variation of the estimated VaR.

Table 4.6: Estimated parameters and VaR for internal data

Cell	Annual Num	Avg Loss	Annual Loss	$\hat{\mu}$	$\hat{\sigma}$	$F(u)$	99.9% VaR of single event			99.9% Annual VaR		
							VaR	Lower	Upper	VaR	Lower	Upper
1	38.2	1.17E+04	4.48E+05	5.51(0.91)	2.49(0.4)	0.61	5.49E+05	1.62E+05	1.06E+06	1.66E+07	1.78E+06	4.36E+07
2	103.2	9.53E+03	9.83E+05	-1.54(2.23)	3.72(0.5)	0.98	2.13E+04	1.26E+04	1.63E+05	2.96E+07	3.03E+06	6.26E+07
3	0.4	1.03E+03	3.95E+02	6.76(0.19)	0.26(0.09)	0.02	1.91E+03	6.73E+02	3.08E+03	3.34E+03	1.76E+03	4.30E+03
4	1.1	7.31E+03	8.39E+03	7.87(0.54)	1.17(0.26)	0.08	9.58E+04	3.77E+03	2.90E+05	1.32E+05	1.05E+04	4.11E+05
5	24.5	1.82E+04	4.45E+05	4.95(1.04)	2.78(0.45)	0.68	7.49E+05	1.29E+05	1.61E+06	2.05E+07	9.28E+05	6.19E+07
6	18.3	5.25E+04	9.64E+05	8.64(0.39)	2.17(0.27)	0.13	4.59E+06	1.14E+06	1.03E+07	3.30E+07	4.88E+06	1.39E+08
7	26.0	3.83E+03	9.94E+04	7.47(0.23)	1.12(0.14)	0.13	5.59E+04	3.10E+04	8.88E+04	2.85E+05	1.59E+05	4.87E+05
8	2.7	6.76E+03	1.81E+04	5.79(0.5)	1.7(0.36)	0.60	6.28E+04	4.20E+03	1.73E+05	2.01E+05	1.11E+04	3.03E+05
9	3.8	1.11E+04	4.25E+04	8.53(0.36)	1.29(0.2)	0.04	2.70E+05	5.75E+04	1.38E+06	5.21E+05	1.30E+05	1.71E+06
10	0.4	4.01E+05	1.53E+05	11.17(1.31)	2.23(0.57)	0.01	7.03E+07	1.01E+05	3.88E+09	3.43E+07	3.64E+03	9.27E+07
11	29.0	6.91E+03	2.01E+05	3.67(1.18)	2.55(0.41)	0.84	1.04E+05	4.46E+04	2.51E+05	3.20E+06	3.27E+05	5.64E+06
12	9.6	3.61E+04	3.45E+05	8.86(0.56)	1.99(0.43)	0.09	3.34E+06	4.97E+05	1.34E+07	1.25E+07	1.20E+06	4.86E+07
13	3.4	3.04E+03	1.05E+04	7.6(0.23)	0.83(0.12)	0.05	2.56E+04	8.32E+03	4.73E+04	5.02E+04	2.06E+04	1.12E+05
14	1.9	1.38E+03	2.64E+03	7.07(0.16)	0.41(0.08)	0.02	4.17E+03	1.82E+03	6.72E+03	1.07E+04	6.79E+03	1.58E+04
15	6.5	1.72E+04	1.12E+05	5.43(0.72)	2.4(0.42)	0.63	3.75E+05	3.13E+04	7.67E+05	3.28E+06	8.01E+04	5.17E+06
16	13.0	7.09E+04	9.21E+05	8.62(0.49)	2.41(0.31)	0.16	9.48E+06	5.68E+05	2.14E+07	5.26E+07	7.90E+06	2.30E+08
17	271.0	8.68E+03	2.35E+06	6.98(0.19)	1.9(0.09)	0.34	3.86E+05	3.00E+05	5.16E+05	1.06E+07	7.02E+06	1.55E+07
18	0.8	1.49E+05	1.14E+05	9.68(1.18)	2.21(0.53)	0.06	1.47E+07	6.12E+03	7.26E+07	1.51E+07	2.12E+04	6.31E+08
19	1.9	6.60E+05	1.26E+06	9.4(1.03)	3.28(0.67)	0.17	3.07E+08	1.34E+05	1.82E+09	1.15E+09	2.05E+05	8.61E+09
20	55.4	1.74E+04	9.65E+05	7.66(0.25)	1.74(0.18)	0.20	4.67E+05	2.06E+05	7.65E+05	4.23E+06	1.60E+06	1.03E+07
21	2.7	1.26E+04	3.37E+04	7.91(0.55)	1.55(0.28)	0.14	3.29E+05	1.44E+04	1.06E+06	6.94E+05	3.77E+04	1.91E+06
22	0.4	3.98E+04	1.52E+04	9.41(0.88)	1.47(0.4)	0.01	1.16E+06	5.22E+03	3.66E+06	9.67E+05	3.84E+03	1.50E+07
23	2.7	5.19E+04	1.39E+05	9.6(0.54)	1.61(0.23)	0.02	2.12E+06	1.78E+05	1.36E+07	3.95E+06	4.14E+05	3.45E+07
24	1.1	2.08E+04	2.38E+04	8.54(0.58)	1.43(0.3)	0.05	4.24E+05	1.11E+04	1.82E+06	7.12E+05	2.05E+04	1.83E+06
25	85.6	1.81E+04	1.55E+06	8.37(0.14)	1.54(0.11)	0.08	5.07E+05	3.20E+05	7.56E+05	4.26E+06	3.07E+06	8.67E+06
26	1.1	1.25E+04	1.44E+04	8.63(0.47)	1.14(0.21)	0.02	1.87E+05	1.87E+04	4.93E+05	2.48E+05	1.57E+04	1.38E+06
Total	707.1	1.59E+04	1.12E+07	6.07(0.28)	2.44(0.12)	0.52	8.13E+05	5.07E+05	1.32E+06	6.79E+07	4.19E+07	1.24E+08

Table 4.7: Estimated parameters and VaR for external data using lognormal

Cell	Annual Num	Avg Loss	Annual Loss	$\hat{\mu}$	$\hat{\sigma}$	$F(u)$	99.9% VaR of single event			99.9% Annual VaR		
							VaR	Lower	Upper	VaR	Lower	Upper
1	179.7	3.38E+04	6.08E+06	5.83(1.25)	2.29(0.3)	0.88	4.04E+05	1.80E+05	6.24E+05	2.65E+07	1.37E+07	5.66E+07
2	569.5	3.45E+04	1.96E+07	-0.59(1.65)	3.34(0.29)	1	1.67E+04	1.17E+04	1.56E+05	1.19E+08	4.34E+07	1.84E+08
3	0.4	5.71E+03	2.29E+03	8.62(0.04)	0.05(0.02)	0.02	6.45E+03	5.22E+03	6.95E+03	1.77E+04	1.57E+04	2.11E+04
4	43.2	9.59E+04	4.15E+06	10.52(0.17)	1.4(0.13)	0.08	2.80E+06	1.45E+06	4.46E+06	1.65E+07	8.81E+06	3.16E+07
5	21.2	2.99E+04	6.34E+05	9.22(0.35)	1.22(0.19)	0.28	4.43E+05	1.99E+05	6.72E+05	2.11E+06	1.06E+06	4.02E+06
6	58.8	7.25E+04	4.26E+06	8.68(0.54)	2(0.24)	0.47	2.85E+06	1.43E+06	4.30E+06	4.59E+07	1.31E+07	1.15E+08
7	188.9	1.92E+04	3.63E+06	8.96(0.15)	1.05(0.07)	0.34	1.97E+05	1.65E+05	2.39E+05	4.98E+06	4.41E+06	5.53E+06
8	28.4	6.42E+04	1.82E+06	3.31(1.46)	3.02(0.45)	0.96	3.04E+05	2.43E+05	1.51E+06	4.37E+07	2.51E+06	5.30E+07
9	97.7	3.98E+04	3.88E+06	-0.08(2.12)	3.36(0.46)	0.99	3.00E+04	2.94E+04	3.99E+05	5.95E+07	8.49E+06	7.14E+07
10	32.8	5.70E+05	1.87E+07	11.18(0.27)	1.98(0.21)	0.09	3.21E+07	1.02E+07	6.68E+07	2.13E+08	8.22E+07	7.98E+08
11	196.1	3.24E+04	6.34E+06	5.3(1.36)	2.41(0.34)	0.91	3.50E+05	1.72E+05	5.74E+05	4.40E+07	1.25E+07	8.92E+07
12	1379.6	2.84E+04	3.92E+07	4.13(1)	2.5(0.2)	0.96	1.42E+05	6.04E+04	2.36E+05	9.38E+07	6.71E+07	1.34E+08
13	3.6	1.68E+04	6.05E+04	8.38(0.25)	1.02(0.19)	0.55	1.03E+05	1.88E+04	2.07E+05	2.64E+05	1.00E+05	3.84E+05
14	30.0	3.71E+04	1.11E+06	9.27(0.34)	1.37(0.2)	0.29	7.27E+05	3.00E+05	1.25E+06	4.22E+06	1.66E+06	8.32E+06
15	149.3	3.76E+04	5.61E+06	4.74(1.66)	2.53(0.39)	0.93	2.80E+05	1.06E+05	5.90E+05	3.48E+07	1.06E+07	6.99E+07
16	83.6	2.31E+05	1.93E+07	9.36(0.4)	2.44(0.23)	0.36	2.18E+07	8.23E+06	4.13E+07	5.27E+08	1.67E+08	1.43E+09
17	2233.2	2.89E+04	6.46E+07	8.38(0.09)	1.52(0.03)	0.54	4.77E+05	4.45E+05	5.12E+05	7.40E+07	6.99E+07	8.02E+07
18	2.0	1.92E+04	3.84E+04	9.57(0.21)	0.62(0.11)	0.04	9.72E+04	3.12E+04	1.85E+05	1.90E+05	7.28E+04	2.87E+05
19	26.8	4.44E+04	1.19E+06	6.54(1.1)	2.17(0.39)	0.82	5.57E+05	2.37E+05	1.31E+06	1.20E+07	1.49E+06	1.63E+07
20	261.7	6.29E+04	1.65E+07	8.8(0.33)	1.92(0.14)	0.44	2.49E+06	1.83E+06	3.44E+06	7.26E+07	3.60E+07	1.08E+08
21	10.0	9.00E+03	9.01E+04	8.6(0.2)	0.6(0.12)	0.44	3.51E+04	1.76E+04	5.31E+04	2.05E+05	1.59E+05	2.36E+05
22	0.8	9.84E+03	7.88E+03	9.06(0.12)	0.26(0.05)	0.02	1.91E+04	9.39E+03	2.50E+04	4.54E+04	3.10E+04	6.22E+04
23	6.0	4.28E+04	2.57E+05	9.98(0.3)	1.08(0.19)	0.09	6.07E+05	1.62E+05	1.16E+06	1.65E+06	4.74E+05	4.15E+06
24	1.6	1.16E+05	1.85E+05	10.84(0.46)	1.14(0.22)	0.02	1.75E+06	1.59E+05	6.43E+06	2.37E+06	1.85E+05	8.32E+06
25	225.3	4.92E+04	1.11E+07	3.67(1.76)	2.87(0.37)	0.95	2.77E+05	7.57E+04	6.36E+05	9.14E+07	2.52E+07	2.02E+08
26	13.2	2.05E+05	2.71E+06	6.73(0.8)	2.65(0.39)	0.75	3.04E+06	5.15E+05	7.61E+06	5.60E+07	2.06E+06	1.38E+08
Total	5932.9	4.12E+04	2.44E+08	5.97(0.26)	2.32(0.07)	0.86	5.06E+05	3.30E+05	7.95E+05	3.67E+08	3.15E+08	4.17E+08

Table 4.8: Estimated parameters and VaR for external data using loggamma

Cell	Annual Num	Avg Loss	Annual Loss	$\hat{\kappa}$	$\hat{\lambda}$	*F(u)*	99.9% VaR of single event			99.9% Annual VaR		
							VaR	Lower	Upper	VaR	Lower	Upper
1	179.7	3.38E+04	6.08E+06	20.63(6.66)	0.37(0.08)	0.74	9.36E+05	6.90E+05	1.81E+06	7.42E+07	2.62E+07	3.57E+08
2	569.5	3.45E+04	1.96E+07	7.41(3.16)	0.63(0.11)	0.97	1.37E+05	4.52E+04	4.95E+05	8.61E+08	1.85E+08	2.37E+09
3	0.4	5.71E+03	2.29E+03	218.37(56.92)	0.04(0.01)	0.11	8.51E+04	5.28E+04	1.20E+06	7.77E+04	5.79E+04	9.78E+05
4	43.2	9.59E+04	4.15E+06	66.84(13.39)	0.16(0.03)	0.04	4.46E+06	2.52E+06	1.16E+07	3.00E+07	1.53E+07	1.12E+08
5	21.2	2.99E+04	6.34E+05	69.09(24.12)	0.14(0.04)	0.15	9.31E+05	5.96E+05	3.87E+06	4.79E+06	2.68E+06	2.35E+07
6	58.8	7.25E+04	4.26E+06	27.63(6.9)	0.34(0.08)	0.32	8.95E+06	4.81E+06	2.92E+07	1.96E+08	6.27E+07	1.23E+09
7	188.9	1.92E+04	3.63E+06	94.9(13.19)	0.1(0.01)	0.20	2.89E+05	2.32E+05	4.31E+05	6.11E+06	5.59E+06	8.30E+06
8	28.4	6.42E+04	1.82E+06	6.82(12.79)	0.75(0.22)	0.94	6.13E+05	3.82E+05	6.93E+06	2.31E+08	8.46E+06	1.92E+09
9	97.7	3.98E+04	3.88E+06	4.12(7.19)	0.82(0.19)	0.99	5.63E+04	3.13E+04	1.26E+06	5.37E+08	2.94E+07	1.41E+09
10	32.8	5.70E+05	1.87E+07	36.04(8.23)	0.32(0.07)	0.05	8.18E+07	3.08E+07	4.59E+08	1.02E+09	2.25E+08	1.00E+10
11	196.1	3.24E+04	6.34E+06	18.97(5.56)	0.39(0.08)	0.76	9.09E+05	7.79E+05	1.93E+06	1.04E+08	4.02E+07	4.64E+08
12	1379.6	2.84E+04	3.92E+07	17.69(3.02)	0.38(0.05)	0.86	3.75E+05	3.01E+05	5.03E+05	1.39E+08	1.05E+08	3.55E+08
13	3.6	1.68E+04	6.05E+04	69.66(53.66)	0.13(0.04)	0.34	3.27E+05	9.41E+04	3.00E+06	7.08E+05	2.18E+05	5.74E+06
14	30.0	3.71E+04	1.11E+06	49.21(17.5)	0.19(0.05)	0.24	1.53E+06	8.55E+05	5.33E+06	9.41E+06	4.50E+06	5.91E+07
15	149.3	3.76E+04	5.61E+06	17.46(6.84)	0.42(0.1)	0.78	1.02E+06	7.63E+05	2.23E+06	1.12E+08	3.35E+07	4.92E+08
16	83.6	2.31E+05	1.93E+07	23.03(5.05)	0.44(0.08)	0.22	6.89E+07	2.74E+07	2.19E+08	3.16E+09	7.00E+08	2.89E+10
17	2233.2	2.89E+04	6.46E+07	49.91(3.37)	0.18(0.01)	0.39	6.10E+05	5.61E+05	7.18E+05	8.44E+07	7.65E+07	9.71E+07
18	2.0	1.92E+04	3.84E+04	153.13(62.34)	0.07(0.03)	0.00	1.47E+06	6.14E+05	2.50E+07	2.21E+06	9.51E+05	4.01E+07
19	26.8	4.44E+04	1.19E+06	21.96(20.1)	0.36(0.14)	0.65	1.55E+06	5.52E+05	5.52E+06	4.27E+07	3.25E+06	2.85E+08
20	261.7	6.29E+04	1.65E+07	32.39(4.62)	0.29(0.04)	0.32	4.36E+06	3.18E+06	7.40E+06	1.54E+08	8.80E+07	5.68E+08
21	10.0	9.00E+03	9.01E+04	186.85(45.37)	0.05(0.02)	0.27	6.53E+04	6.20E+04	6.20E+05	2.87E+05	2.93E+05	1.90E+06
22	0.8	9.84E+03	7.88E+03	391.68(46.98)	0.02(0.01)	0.02	6.04E+04	5.96E+04	1.16E+06	8.61E+04	8.20E+04	1.28E+06
23	6.0	4.28E+04	2.57E+05	93.74(50.16)	0.11(0.05)	0.02	1.55E+06	4.83E+05	1.27E+07	3.33E+06	1.25E+06	3.37E+07
24	1.6	1.16E+05	1.85E+05	95.81(64.48)	0.12(0.05)	0.00	4.75E+06	6.34E+05	6.38E+07	7.14E+06	9.13E+05	8.31E+07
25	225.3	4.92E+04	1.11E+07	14.87(5.59)	0.47(0.13)	0.80	1.26E+06	7.46E+05	2.50E+06	1.97E+08	7.81E+07	2.84E+09
26	13.2	2.05E+05	2.71E+06	13.19(18.65)	0.59(0.22)	0.66	9.90E+06	1.41E+06	6.14E+07	2.33E+08	7.46E+06	4.54E+09
Total	5932.9	4.12E+04	2.44E+08	22.77(1.28)	0.34(0.01)	0.70	9.14E+05	8.50E+05	9.96E+05	6.05E+08	4.54E+08	8.65E+08

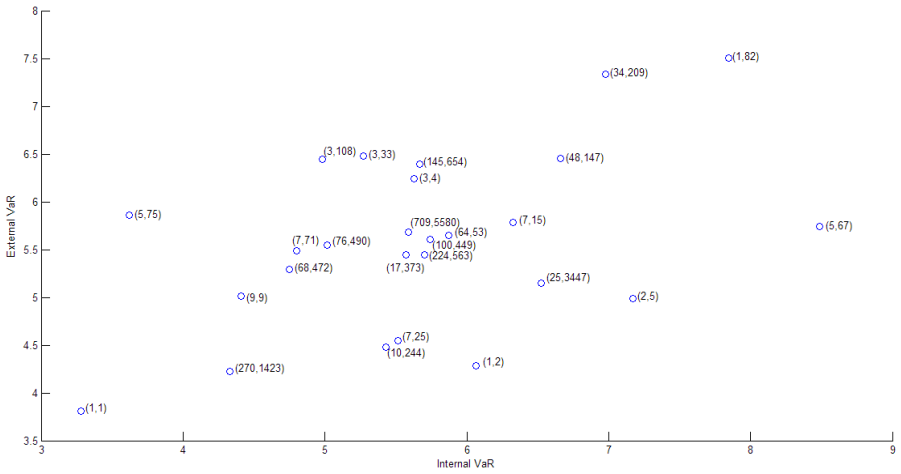


Figure 4.11: Scatter plot of internal and external 99.9% VaR

Note: The VaRs shown are logarithm to base 10. The numbers denote the sample sizes of internal and external data for each cell.

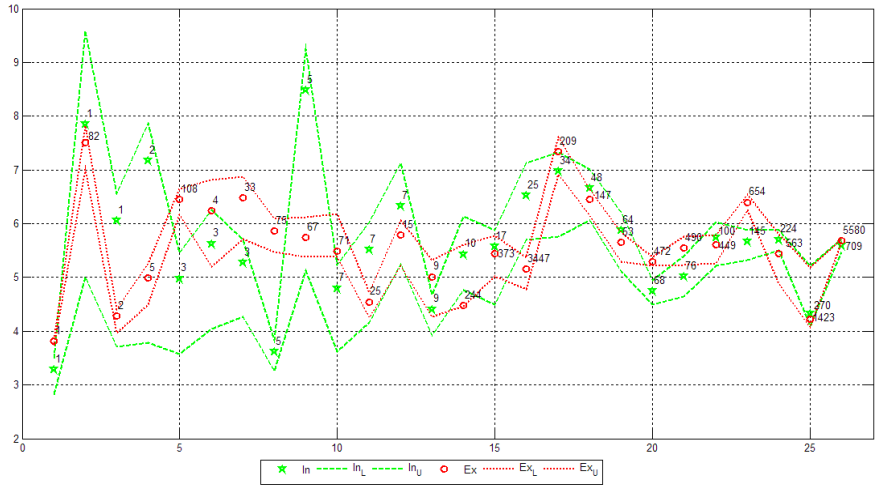


Figure 4.12: Comparison of internal and external 99.9% VaR

Note: The VaRs in the plot are logarithm to base 10. The numbers denote the sample sizes of internal and external data for each cell (sorted by internal sample size increasingly).

The selection of the loss severity distribution There are a number of studies that examine the selection of the loss severity distribution. In this study, we investigate both the lognormal and loggamma distribution for the external data, because the latter is more flexible to capture the extreme heavy tails as it includes the infinite-mean and infinite-variance cases. By comparing the 99.9% VaR of a single event and its confidence intervals for the external data using the lognormal and loggamma distributions, we find that using the loggamma distribution leads to a higher VaR estimate for every cell than using the lognormal

distribution (see Figure 4.13). For most of the cells, the loggamma VaR is approximately two to three times greater than the lognormal VaR. This suggests that without sufficient data even with similar in-sample fitting performance, using the loggamma distribution would lead to a higher estimate of the 99.9% VaR.

In Figure 4.14, we compare the 99.9% annual VaR for the external data using the lognormal and loggamma distributions. We find that using the lognormal distribution leads to more practical estimates, while the estimates for the loggamma distribution are more conservative and therefore may not be practical. In addition, we observe that when there is a large sample and the estimated proportion above the threshold is not small, the confidence interval of the estimates are very narrow. For example, for cell 7 (the 20-th point in Figure 4.14) with 472 observations, and cell 17 (the 26-th point in Figure 4.14) with 5580 observations, the estimated area above the threshold is respectively 66% and 46%. We find that for these two cells the 90% lower and upper bound of the 99.9% VaR estimates are very close, indicating that with less data truncated and more data reported, the estimates become more credible.

Notice that the estimated truncated area depends on the selection of the loss severity distribution. For cell 2 (the 24-th point in Figure 4.14) with 1423 observations, the estimated truncated area by using the lognormal and loggamma distributions are both higher than 0.97. This may suggest a misspecification of the distribution, and using a different distribution that leads to a smaller truncated proportion may improve the credibility.

In general, the estimation results of choosing different types of distribution are somewhat foreseeable. For example, in this paper, we see that using the loggamma distribution leads to a higher estimate of 99.9% VaR. Therefore, in practice, one may decide the selection of the distribution based on both the in-sample fitting performance and the practicality of the estimated VaR measures.

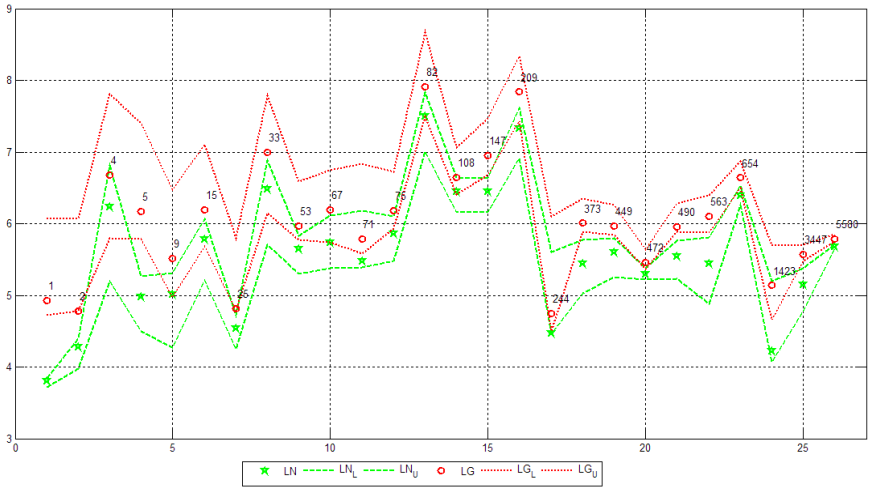


Figure 4.13: External 99.9% VaR using lognormal and loggamma distribution
Note: The VaRs are shown in logarithm scale to base 10. The numbers denote the sample sizes of external data for each cell (sorted increasingly).

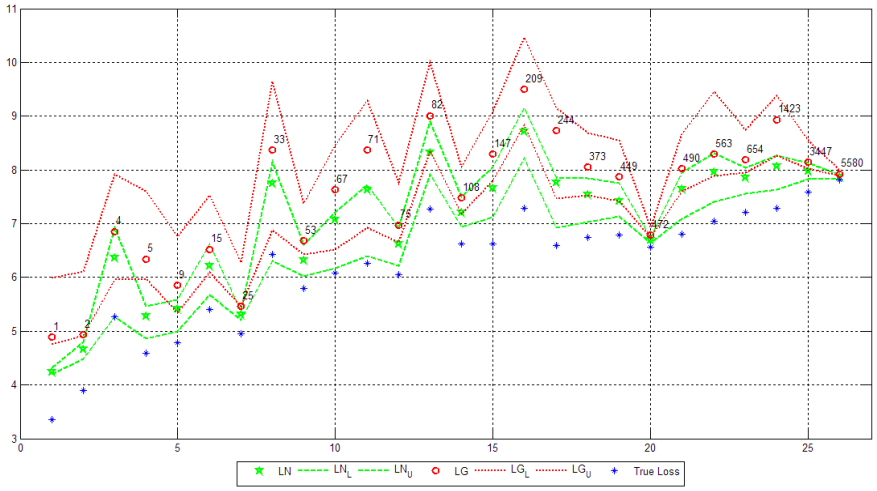


Figure 4.14: External 99.9% annual VaR using lognormal and loggamma distribution
Note: The annual VaRs and true losses are logarithm to base 10. The numbers denote the sample sizes of external data for each cell (sorted increasingly).

Granularity and credibility It is well known that the cells with low frequency could not be overlooked as they may contain extreme losses. In Figure 4.15, we estimate for each cell the 99.9% annual VaR for the internal data, and compare them with the average annual losses occurred. For cell 19 of the internal data (the 9-th point in Figure 4.15), we see that there are only five losses in this cell but it accounts for more than 10% among all the cells (see the loss figures in Table 4. 6). Neglecting these cells with few data will miss the most important losses.

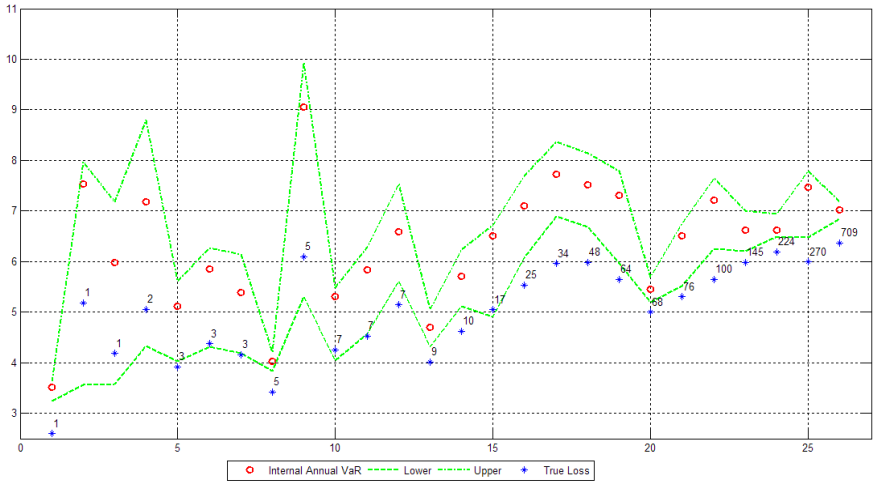


Figure 4.15: Internal 99.9% annual VaR for each cell using lognormal distribution
Note: The annual VaRs and true losses are logarithm to base 10. The number at each point denotes the sample size of internal data for each cell (sorted increasingly).

However, many studies overlook these cells or combine them with other cells with sufficient data because of the difficulty in estimating the parameters. The last rows in Tables 4.6, 4.7, and 4.8 report the estimation results by grouping all the losses into one cell. We see that the estimated VaR for the total losses is even smaller than that for the losses in a single cell because the effect of extreme losses are attenuated when pooled with numerous small losses. To avoid merging the cells, the Bayesian method provides a way to estimate the distributions for the cells with few data directly, and helps to pinpoint the cells with the highest potential risk. Therefore, the granularity of classification of the loss distributions is improved. For the internal data, we have studied the risk profile for each of the 26 cells that have loss observations, which is closer to the 56-cell requirement by Basel II than by combining the cells.

From Figure 4.15, we see that generally with more than 10 losses available, using the Bayesian method produces a VaR estimate with a reasonable confidence interval and the results seem to be reliable when compared to the true losses. For the extremely scarce data with one to five observations, however, we see that the confidence intervals of the estimates seem to be too wide to be useful. In this case, one should also refer to the severity distribution estimated by grouping the cells as an overall profile. The trade-off between improving the granularity and narrowing the confidence intervals is still an issue, which requires more data. After all, the sample of internal data we used in this paper only covers less than three years. However, it is seen that applying the Bayesian method leads to a more credible estimate than using the MLE method.

4.6 Conclusion

In this paper, we focus on the problem of parameter estimation for left-truncated data, especially for small samples. This problem is important in the estimation of operational risk. Having investigated a set of estimation methods such as the MLE method (using direct optimization, the EM algorithm, and a special case of the Bayesian method with "no prior"), the PLE method, and the main focus — Bayesian approach with the Jeffreys' prior — we present an extensive study on the parameter estimation problem for truncated data. Through a series of simulation studies based on the lognormal distribution, we find that using the Bayesian method greatly reduces the mean squared error of the parameters. The effect of selecting various priors and criteria to choose the Bayesian estimates are also discussed.

We apply the Bayesian method using the Jeffreys' prior for the truncated lognormal and loggamma distributions in an empirical study using real operational risk data from a European large bank. By estimating the loss distribution parameters for each cell using the internal and external data separately, we avoid naively pooling these two datasets together. Compared to the MLE method, the Bayesian method greatly relieves the difficulty of parameter estimation for cells with few data. By comparing the internal and external VaR estimates, a significant correlation between the internal and external loss severity distribution is found. Finally, by applying a bootstrap method, the confidence intervals of the estimates are obtained and are reasonable even for small samples. However, for the cells with extremely scarce data, we suggest that the results based on merging cells should be used as an overall profile.

5 Covariance Estimation

Covariance is an important input for a wide variety of statistical, engineering, and financial applications. Estimating covariance for high dimensional data is a classically difficult problem, which can be explained from various points of view. First of all, the number of parameters to be estimated grows as the squared dimension, which is often referred to as the curse of dimensionality. This results in great difficulty in parameter estimation. Therefore, many approaches try to reduce the dimension by assuming some structure on the covariance of random variables. This gives rise to a large branch in multivariate statistical analysis - dimension reduction techniques such as principal component analysis (PCA) and factor analysis (FA). It also leads to flourish research of covariance selection models such as sparse inverse covariance in statistics. A parallel study in engineering is called Gauss-Markov random field (GMRF) models which are closely related to graph models. Second, the sample covariance also leads to large estimation errors in the eigenvalue spectrum, thus performs poorly in many applications such as portfolio selection. A class of estimators is by shrinking the covariance estimate toward some pre-estimators, and is found to have smaller estimation error than sample covariance. These shrinkage estimators have found successful applications in finance by shrinking the covariance (e.g., towards identity matrix or a factor model) and then finding an optimum shrinkage coefficient. Third, the essential reason of why sample covariance performs poorly in high dimensions is that the noise in each element in the covariance matrix are not random but interacted with each other. Classical results from random matrix theory have shown that the eigenvalue distribution is even inconsistent to the true spectrum and has an asymptotically determined shape with the ratio of sample size to dimension being a constant. Therefore, many researchers have been trying to adjust the sample eigenvalues distribution in order to achieve improved covariance estimates. Given the enormous research directions behind the covariance estimation problem, it is beneficial to review the literature from multiple angles and understand the relation of them to each other.

5.1 Why sample covariance is bad?

Before reviewing the vast literature in covariance estimation method, it's better to know why sample covariance is bad and when it is bad. We would like to start from some classical results from random matrix theory (RMT). There are several important laws in RMT about eigenvalues' distribution for large random matrix such as Wigner's semicircle law and Marcenko-Pastur law.

Theorem 5.1 *(Wigner's semicircle law)* Consider an $n \times n$ matrix A with entries from $N(0, \sigma^2)$. Define

$$A_n = \frac{1}{\sqrt{2n}}(A + A^T)$$

then the asymptotic distribution of A_n's eigenvalues has the density

$$\rho(\lambda) = \begin{cases} 1/(2\pi\sigma^2)\sqrt{4\sigma^2 - \lambda^2}, & \text{if } |\lambda| \leq 2\sigma \\ 0, & \text{otherwise.} \end{cases}$$

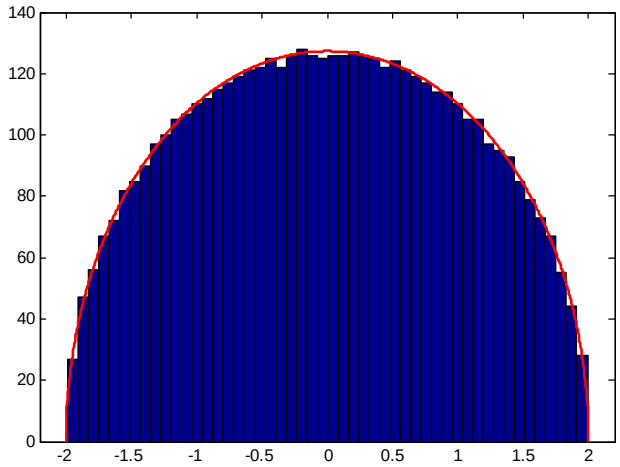


Figure 5.1: Wigner's semicircle law

Figure 5.1 shows an example plot of Wigner's semicircle law.

Marcenko-Pastur law is another more general theorem that describes the asymptotic spectrum distribution of large random matrices.

Theorem 5.2 *(Marcenko-Pastur)* If X is a $p \times n$ random matrix whose entries are independent and identically distributed random variables from $N(0, \sigma^2)$, let S be the sample covariance of $X : S_n = \frac{1}{n} X X^T$, and $\lambda_1, ...\lambda_p$ be the eigenvalues of S_n.Define the empirical distribution of λ by

$$F_p(x) = \frac{1}{p} \sum_{i=1}^{p} I_{\{\lambda_i \leq x\}}.$$

Let $n, p \to \infty$, and keep the ratio $q = n/p \geq 1$ a constant, then the asymptotic distribution of eigenvalues of S_n has density

$$\rho(\lambda) = \frac{q}{2\pi\sigma^2} \frac{\sqrt{(\lambda_+ - \lambda)(\lambda_- - \lambda)}}{\lambda}$$

where the maximum and minimum eigenvalues are given by

$$\lambda_\pm = \sigma^2 \left(1 \pm \sqrt{\frac{1}{q}}\right)^2$$

Here the $\rho(\lambda)$ is known as the Marcenko-Pastur density.

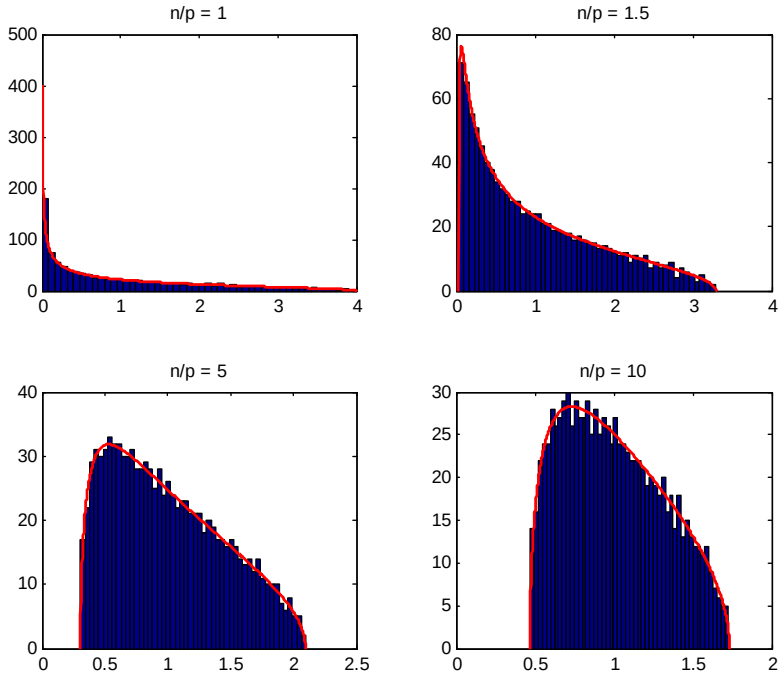


Figure 5.2: Marcenko-Pastur density under different n/p ratios

Figure 5.2 shows the shape of Marcenko-Pastur density under different ratio of n/p. From the fact that the samples are i.i.d, the true spectral distribution is a density with mass 1 at 1 (the covariance is an identity matrix). This theorem implies that the spectral distribution of high-dimensional sample covariance does not converge to the true spectral distribution when n and p both are large. The eigenvalues of sample covariance are dispersed and form a distribution with shape determined by the Marcenko-Pastur law.

There are also other important laws in random matrix theory such as Tracy-Widom law which describes the limiting distribution of the largest eigenvalues (Tracy and Widom, 2002).

Ledoit and Wolf (2003) showed that eigenvalues of sample covariance matrix are more dispersed than those of true covariance matrix.

Theorem 5.3 *(Ledoit & Wolf)* The eigenvalues are the most dispersed diagonal elements that can be obtained by rotation. Let R be a p-dimensional symmetric matrix and $V = (v_1, ..., v_p)$ be a p-dimensional rotation matrix: $VV^T = V^TV = I$ where v_i are the column vectors of V. The mean of the diagonal elements is $r = \frac{1}{p}tr(V^TRV) = \frac{1}{p}tr(R)$. Consider the dispersion of the diagonal elements of V^TRV:

$$d = \frac{1}{p}\sum_{i=1}^{p}\left(v_i^TRv_i - r\right)^2$$

Then d is maximized if and only if v_i are the eigenvectors ($v_i^TRv_i$ are the eigenvalues).

Proof): The proof is in Ledoit and Wolf (2003). Here we briefly restate the idea of the proof.

$$\begin{aligned} & \sum_{i=1}^{p} \left(v_i^T R v_i - r\right)^2 + \sum_{i=1}^{p} \sum_{j=1, j\neq i}^{p} \left(v_i^T R v_j\right)^2 \\ = \quad & tr\left((V^T R V - rI)^2\right) = tr((R - rI)^2) \end{aligned}$$

Then d is maximized when $\sum_{i=1}^{p} \sum_{j=1,j\neq i}^{p} \left(v_i^T R v_j\right)^2$ is minimized. This can be achieved when $v_i^T R v_j = 0$ (for $i \neq j$). Then $V^T R V = D$ is a diagonal matrix. Since $VV^T = I$, therefore V must contain all the eigenvectors of R and $R = VDV^T$.

Corollary 5.1 *(Ledoit & Wolf)* The eigenvalues of sample covariance matrix are more dispersed than those of true covariance matrix.

Proof): Assume Σ is the true covariance matrix, and its eigenvalue decomposition is $\Sigma = \Gamma^T \Lambda \Gamma$. Also, assume the sample covariance matrix is S, and its eigenvalue decomposition is $S = G^T L G$. From the above theorem, the dispersion of $L = GSG^T$ is the greatest among all rotations of S. Denote the dispersion of the diagonal elements of a matrix A be $d(A)$, then

$$d(GSG^T) > d(\Gamma^T S \Gamma)$$

Notice that sample covariance matrix S is an unbiased estimate of the true covariance matrix Σ. The dispersion of $\Gamma^T \Sigma \Gamma$ is approximately equal to the dispersion of $\Gamma^T S \Gamma$ (However, GSG^T is not at all an unbiased estimate of $\Gamma^T \Sigma \Gamma$ because the errors of G and S interact).Therefore

$$d(\Gamma^T S \Gamma) \approx d(\Gamma^T \Sigma \Gamma)$$

$$d(GSG^T) > d(\Gamma^T \Sigma \Gamma)$$

This means the eigenvalues of sample covariance matrix are more dispersed than the eigenvalues of true covariance matrix.

Now let's look at some examples. Consider the true covariance being an auto-covariance of an AR(1) process:

$$\Sigma = \begin{bmatrix} 1 & \rho & \rho^2 & \dots & \rho^{p-1} \\ \rho & 1 & \rho & & \\ \rho^2 & \rho & 1 & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \rho \\ \rho^{p-1} & \dots & & \rho & 1 \end{bmatrix}$$

Consider when ρ =0, 0.25, 0.5 and 0.75. The spectral distributions of the sample covariance of samples $X_{p\times n}$ from $N(0, \Sigma)$ with n/p =1.5, 5 and 10 against the true spectral distribution is shown in Figure 5.3. When ρ is higher, the eigenvalues of sample covariance become closer to the true eigenvalues. For the small eigenvalues close to 1, it's better to view in log-scale (see Figure 5.4). We see that the eigenvalues of sample covariance are always dispersed than the true covariance. The logarithm of eigenvalues actually are related to the information distance (Atkinson and Mitchell, 1981).

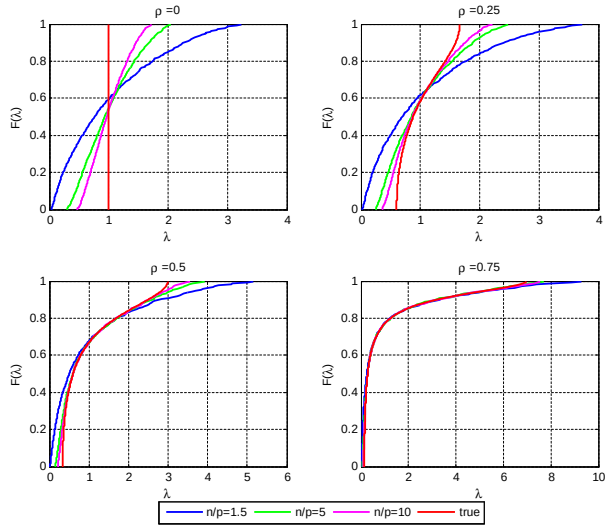


Figure 5.3: Empirical CDF of eigenvalues for AR(1) covariance with differnt ρ's

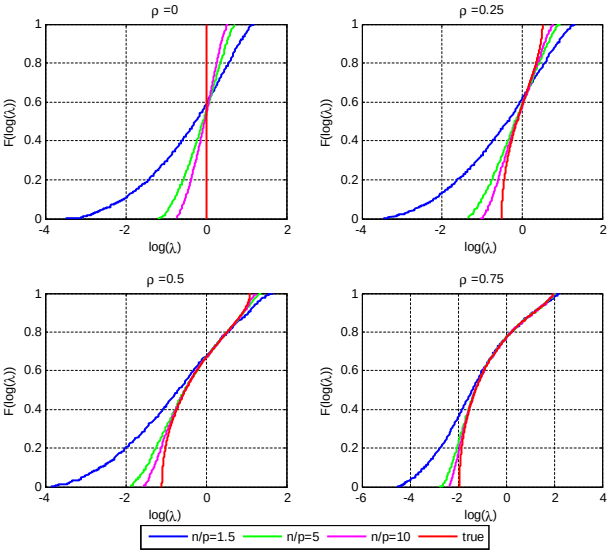


Figure 5.4: Empirical CDF of logarithm of eigenvalues of AR(1) covariance with different ρ's

5.2 Literature review on general estimation methods

There is a vast literature about covariance estimation using different methods from different angles. We try to classify them into four categories: shrinkage estimates, structured models, RMT approaches and Bayesian approaches.

5.2.1 Shrinkage estimates

Shrinkage approach generally refers to the method by weighing the sample covariance and a target (usually identity matrix):

$$\hat{\Sigma} = \rho\mu I + (1-\rho)S$$

where $\mu = tr(\Sigma)/p$, and in many discussions Σ is assumed to be trace-normalized with $tr(\Sigma) = p$.

By shrinking the sample covariance towards the identity matrix, the arrived covariance estimate is well-conditioned. In addition, since the spectral density of identity matrix puts all mass at $\lambda = 1$, the spectral distribution of shrinkage covariance estimate become less dispersed. The central problem for shrinkage estimates is to find an optimal shrinkage coefficient under ρ. The definition of optimality relies on the error metric between two covariance (or loss function). Usually, the Frobenius norm error $\left|\left|\Sigma - \hat{\Sigma}\right|\right|_F^2 = tr(\Sigma - \hat{\Sigma})(\Sigma - \hat{\Sigma})^T)$ is used. In some literature, the Frobenius norm is normalized by the dimension p.

The "shrinkage" idea dates back to Stein (1956)'s seminal work by shrinking the sample mean towards a constant to reduce estimation error. This also has an interpretation of taking a trade-off between the bias and estimation error. Efron and Morris (1977) provides a general introduction to shrinkage estimates. Haff (1980) considered similar linear combination from an empirical Bayes point of view and considered using Stein's loss $L_1(\hat{\Sigma}, \Sigma) = tr(\hat{\Sigma}\Sigma^{-1}) - \log\det(\hat{\Sigma}\Sigma^{-1}) - p$ and quadratic loss $L_2(\hat{\Sigma}, \Sigma) = tr(\hat{\Sigma}\Sigma^{-1} - I)^2$ as loss functions. However, these loss functions requires finding the inverse of covariance matrix.

Ledoit & Wolf (2003) introduced a well-conditioned covariance matrix as a weighted combination between the sample covariance and the identity matrix by considering minimizing the Frobenius norm error $E[\left|\left|\Sigma - \hat{\Sigma}\right|\right|^2]$ where $\|A\| = tr(AA^T)/p$.

Theorem 5.4 *(Ledoit & Wolf)* Consider the optimization problem:

$$\Sigma^* = \arg\min_{\hat{\Sigma}} E[\left|\left|\hat{\Sigma} - \Sigma\right|\right|^2]$$

$$\text{s.t } \hat{\Sigma} = \rho\mu I + (1-\rho)S$$

where $\mu = tr(S)/p$.

Then its solution is:

$$\rho^* = \frac{E[\|S - \Sigma\|^2]}{E[\|S - \mu I\|^2]}$$

and $(1-\rho)$ can also be written as:

$$1 - \rho^* = \frac{E[\|\Sigma - \mu I\|^2]}{E[\|S - \mu I\|^2]}.$$

since $E[\|S - \mu I\|^2] = E[\|S - \Sigma\|^2] + E[\|\Sigma - \mu I\|^2]$.

Proof): see Ledoit and Wolf (2003).

However, the optimal shrinkage coefficient ρ^* is dependent on the true covariance Σ.

Chen, Wiesel and Hero (2009) rewrite the formula of ρ^* (called the covariance estimate as oracle estimator $\hat{\Sigma}_O$) as:

$$\rho^* = \min\left(\frac{(1-2/p)tr(\Sigma^2) + tr^2(\Sigma)}{(n+1-2/p)tr(\Sigma^2) + (1-n/p)tr^2(\Sigma)}, 1\right)$$

using the identities (Letac and Massam, 2004):

$$\begin{aligned} E[tr(S_n)] &= tr(\Sigma) \\ E[tr(S_n^2)] &= \frac{n+1}{n} tr(\Sigma^2) + \frac{1}{n} tr^2(\Sigma). \end{aligned}$$

Ledoit and Wolf (2003) proposed to approximate the above quantities $E[\|S-\Sigma\|^2]$ and $E[\|\Sigma-\mu I\|^2]$ using sample information. Let $X_{p\times n} = (X_1, ..., X_n)$ where $\{X_i\}_{i=1}^n$ are the observed samples. Let $S_n = \frac{1}{n} XX^T$, it can be written as $S_n = \frac{1}{n}\Sigma_{i=1}^n X_i X_i^T$. Denote:

$$m_n \approx tr(S_n)/p$$

$$\begin{aligned} d_n^2 &= \|S_n - m_n I\|^2 \\ b_n^2 &= \min(\frac{1}{n^2}\Sigma_{i=1}^n \left\|X_i X_i^T - S_n\right\|^2, d_n^2) \end{aligned}$$

Then,

$$\hat{\rho}_{LW} = \min\left(\frac{\frac{1}{n^2}\Sigma_{i=1}^n \left\|X_i X_i^T - S_n\right\|^2}{\|S_n - m_n I\|^2}, 1\right)$$

and the shrinkage covariance estimate for Σ^* is:

$$S_n^* = \hat{\rho}_{LW} m_n I + (1-\hat{\rho}_{LW}) S_n.$$

Chen, Wiesel and Hero (2009) rewrite the formula of $\hat{\rho}_{LW}$ as:

$$\hat{\rho}_{LW} = \min\left(\frac{\frac{1}{n^2}\Sigma_{i=1}^n \left\|X_i X_i^T - S_n\right\|_F^2}{tr(S_n^2) - \frac{1}{p} tr^2(S_n)}, 1\right)$$

where $||\cdot||_F^2$ is the more general Frobenius norm without normalizing by $1/p$:

$$||A||_F^2 = tr(AA^T)$$

The LW estimator is well conditioned under small sample sizes and is distribution free without restriction to Gaussian assumption only. However, Chen, Wiesel and Hero (2009) showed that the LW estimator can be significantly improved when the distribution is Gaussian. They first applied the Rao-Blackwell theorem to the LW method, and obtained an estimator, called RBLW estimator, which dominates the LW estimator under mean-squared errors. The shrinkage coefficient of RBLW is:

$$\hat{\rho}_{RBLW} = \frac{(n-2)/n \cdot tr(S_n^2) + tr^2(S_n)}{(n+2)\left[tr(S_n^2) - 1/p \cdot tr^2(S_n)\right]}$$

The LW estimator asymptotically achieves the minimum MSE with respect to shrinkage estimates. However, for small n, there is no such guarantee that such optimality still holds. Chen, Wiesel and Hero (2009) developed an iterative method to approximate the oracle estimator under Gaussian assumptions, called OAS estimator. The proposed iteration is,

$$\hat{\rho}_{j+1} = \frac{(1-2/p)tr(\hat{\Sigma}_j S_n) + tr^2(\hat{\Sigma}_j)}{(n+1-2/p)tr(\hat{\Sigma}_j S_n) + (1-n/p)tr^2(\hat{\Sigma}_j)}$$

$$\hat{\Sigma}_{j+1} = \hat{\rho}_{j+1} m_n I + (1 - \hat{\rho}_{j+1}) S_n$$

A closed-form limit of the above iteration is then obtained as follows:

$$\hat{\rho}_{OAS} = \min\left(\frac{(1 - 2/p)tr(S_n^2) + tr^2(S_n)}{(n + 1 - 2/p)\left(tr(S_n^2) - 1/p \cdot tr^2(S_n)\right)}, 1 \right)$$

and the OAS shrinkage estimate is:

$$\hat{\Sigma}_{OAS} = \hat{\rho}_{OAS} m_n I + (1 - \hat{\rho}_{OAS}) S_n.$$

By a series of numerical simulation experiments on auto-covariance of AR(1) process and incremental fractional Brownian motion (FBM) process, the OAS estimator remarkably approximates the oracle estimator very well, and significantly outperforms the LW estimator when n is small.

In addition to the "shrinking to identity" models, there are also other shrinkage estimates by shrinking the covariance towards a factor model in empirical finance when there are known factor structure of the covariance matrix, see Ledoit and Wolf (2003).

The shrinkage covariance estimates can also be interpreted in the empirical Bayes framework. The shrinkage target μI is considered as prior information, and sample covariance contains the sample information. In a full Bayesian approach, the prior distribution is specified and a distribution of Σ would be obtained. In empirical Bayes approach, a kernel of the prior distribution is used such as the MLE or moment estimator. The weighted combination of prior and sample covariance can be understood as a posterior estimate for Σ.

5.2.2 RMT approaches

From random matrix theory and Ledoit and Wolf (2003)'s theorem, the eigenvalues of sample covariance matrix are more dispersed than those of true covariance matrix. Therefore, a natural idea is to adjust the eigenvalues of the sample covariance matrix.

Stein (1975) keeps eigenvectors intact and replaces eigenvalues of the sample covariance matrix $l_1, ..., l_p$ by:

$$nl_i / \left(n - p + 1 + 2l_i \sum_{j=1, j \neq i}^{p} \frac{1}{l_i - l_j} \right), i = 1, ..., p$$

A simple attempt to apply the random matrix theory is to only retain the eigenvalues greater than the λ_+ appeared in the Marcenko-Pastur density while remove the other eigenvalues located in (λ_-, λ_+) based on an intuition that those eigenvalues are random noises. Plerou et al (2002) truncated all the eigenvalues below λ_+ and then set the diagonal elements be the same as sample covariance matrix. Instead of truncating all the small eigenvalues in (λ_-, λ_+), Laroux et al (1999) replaced them with a constant so that the trace is the same as sample covariance. Bengtsson and Holst (2002) introduced a shrinkage approach which basically decreases the small eigenvalues at a single rate. Park and O'leary (2010) put these estimators of adjusting eigenvalues into a common framework using a Tikhonov filtering function, and shrink the eigenvalues at different rates.

However, these approaches only empirically adjust the eigenvalue's size, but do not consider from the true eigenvalues distribution of a given covariance matrix. Karoui (2008) applied the Random matrix theory, and estimated the true spectral distribution from the

sample spectral distribution, which is an inverse problem to the Marcenko-Pastur law, by finding a measure to minimize the distance between it and the Stieltjes transform of sample spectrum. In his paper, he recovers the true spectrum well, though later Ledoit and Wolf (2012) mentioned that they could not replicate the results.

Ledoit and Wolf (2012) proposed a new nonlinear shrinkage estimator similar to Karoui (2008) but by minimizing the distance between the spectral distributions directly in real space, which they stated is better than Karoui's approach, but the price is that this approach is a non-convex problem. Won (2009) estimated the covariance matrix under condition number constraints, which Ledoit mentioned as a nonlinear shrinkage as well.

5.2.3 Structured models

Factor model In high-dimensional covariance estimation, the parameters to be estimated are often comparable or much more than the number of observations. To reduce the number of parameters, a structure is often applied on the covariance. Factor models are one of the typical examples in structured covariance estimation. Broadly speaking, they include observable factor models and latent factor models (Bai and Shi (2011)). The two models can also be combined in practice.

The observable factors usually include market index factor (CAPM), macroeconomic factors (Chen, Roll and Ross (1986)), and firm-specific fundamental factors (Fama and French (1993)). The general representation of the model is:

$$r_t = Bf_t + \varepsilon_t$$

where $r_t = (r_{1t}, ..., r_{pt})^T$ is a $p \times 1$ vector of excess returns, $f_t = (f_{1t}, ..., f_{kt})^T$ is a $k \times 1$ vector of factor returns with zero mean, $B = (\beta_1, ..., \beta_p)^T$ is $(p \times k)$ factor loadings matrix where $\beta_i = (\beta_{i1}, ..., \beta_{ik})$ are the factor loadings of the i-th return, and $\varepsilon_t = (\varepsilon_{1t}, ..., \varepsilon_{pt})^T$ is the a $p \times 1$ vector of residuals. It is assumed that the ε_t are cross-sectionally independent.

Under this model, the covariance matrix of r_t is then represented as:

$$\Sigma = B\Omega_f B^T + \Omega_\varepsilon$$

where Ω_f is a $(k \times k)$ covariance matrix of f_t, and Ω_ε is a diagonal matrix including the variance of ε_t. The factor loadings matrix are obtained by linear regression, and Ω_f and Ω_ε are estimated from the factors sample and the residuals.

The latent factor models differ from observable factors in that both the factor loadings and the factors are unobservable. In this case, they can not be identified separately without restrictions. Thus, the factors are often assumed to be independent and have variance 1, so that $\Omega_f = I$. Then the covariance model is:

$$\Sigma = VV^T + D$$

This is usually referred to as factor analysis (FA).

The V and D are estimated by maximum likelihood method. Under Gaussian assumption for the distribution of the residuals ε_t, then $x_t \sim N(0, \Sigma)$ where $\Sigma = D + VV^T$.Assume we have a sample $X_{p\times n}$ the log-likelihood can be written as:

$$l(\theta|r_t) = -\frac{n}{2}\{\ln|\Sigma| + tr(\Sigma^{-1}S)\}$$

where $\theta = (D, V), S = 1/nXX^T$ is the sample covariance matrix.

The ML estimate of θ is equivalent to solving following system (Joreskog, 1967; Zhao, Yu and Jiang, 2008):

$$\begin{aligned} V &= S\Sigma^{-1}V \\ D &= diag\{S - VV^T\} \end{aligned}$$

However, this system is difficult to solve explicitly and it is necessary to use iterative procedures. Zhao, Yu and Jiang (2008) reviewed a series of iterative algorithms including Lawley (1940)'s algorithm, Rubin and Thayer (1982)'s EM algorithm, Liu and Rubin (1998)'s ECME algorithm, and proposed a new ECME algorithm and a conditional maximization (CM) algorithm in their paper.

We do not intend to go into details of the derivation of ML estimation of factor analysis. Readers can refer to the above research for details. Briefly the EM algorithm is follows:

1. Set $V^{(0)}, D^{(0)}$.
2. Let $\Sigma^{(t)} = D^{(t)} + V^{(t)}(V^{(t)})^T$.
3. $V^{(t+1)} = S(D^{(t)})^{-1}V^{(t)}(I + (V^{(t)})^T(\Sigma^{(t)})^{-1}S(D^{(t)})^{-1}V^{(t)})^{-1}$;
4. $D^{(t+1)} = diag(S - S(\Sigma^{(t)})^{-1}V^{(t)}(V^{(t+1)})^T)$.

The computation of $(\Sigma^{(t)})^{-1}V^{(t)}$ can be reduced using

$$\Sigma^{-1}V = D^{-1}V(I + V^TDV)^{-1}$$

Using factor models has great advantages in reducing the number of parameters for estimating a covariance matrix and reducing estimation error. Fan et al (2008) showed that factor model greatly outperforms the sample covariance when evaluating the estimation error using entropy loss. They also showed that factor models greatly improves the estimation of inverse covariance and thus the optimal portfolio allocation. From computational point of view, using factor model also reduces the computation of calculating the inverse of a big matrix by factorizing the matrix.

Principal component analysis (PCA) can also be used to estimating a covariance matrix. Assume the eigenvalue decomposition of sample covariance is:

$$S = UDU^T$$

where $D = diag\{d_{11}, d_{22}, ..., d_{pp}\}$ is a diagonal matrix containing the eigenvalues with $d_{11} \geq d_{22} \geq ... \geq d_{pp}$, and $U = (u_1, ..., u_p)$ include the corresponding eigenvectors. Choosing a number of principal components k, by keeping only the first k components and cutting off the small eigenvalues, we can obtain a covariance matrix estimate (Elton and Gruber, 1973), but it is rank-deficient. To obtain a full rank estimate, the diagonal components can be replaced using the sample variances. Then an estimate of the covariance can be (Bai and Shi, 2011):

$$\hat{\Sigma} = U_{1:k}D_{1:k}U_{1:k}^T + diag\{S - U_{1:k}D_{1:k}U_{1:k}^T\}$$

Bengtsson and Holst (2002) instead shrink the truncated covariance to the sample covariance:

$$\hat{\Sigma} = \rho(U_{1:k}D_{1:k}U_{1:k}^T) + (1 - \rho)S$$

A typical difference between PCA and FA is that PCA do not have the communalities term as in the diagonal matrix of D in FA. When there is heavy cross-sectional heteroskedasticity, it is better to apply PCA on the correlation matrix (see Bai (2010) and Jones (2001)).

Sparse Inverse Covariance The sparse inverse covariance is interesting because of its close relation to conditional independence. Suppose $X = (X_1, ..., X_p)^T$ is a p-dimensional random vector from a joint Gaussian distribution $N(\mu, \Sigma)$. Without loss of generality, consider the conditional covariance of X_1 and X_2 given $X_3, ..., X_p$. Let $x = (X_1, X_2)^T$, and $y = (X_3, ..., X_p)^T$, then

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim N(\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix})$$

We have

$$x|y \sim N(\mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y - \mu_y), \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx})$$

If X_1 and X_2 are conditionally independent given $y = (X_3, ..., X_p)^T$, then the off-diagonal components of $\Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}$ equal to zero. Let $\Omega = \Sigma^{-1}$, we have

$$\Omega = \begin{pmatrix} \Omega_{xx} & -\Omega_{xx}\Sigma_{xy}\Sigma_{yy}^{-1} \\ -\Sigma_{yy}^{-1}\Sigma_{yx}\Omega_{xx} & \Sigma_{yy}^{-1} + \Sigma_{yy}^{-1}\Sigma_{yx}\Omega_{11}\Sigma_{xy}\Sigma_{yy}^{-1} \end{pmatrix}$$

where $\Omega_{xx} = \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}$ is right the covariance matrix of x given y. Therefore,

$$(\Omega_{xx})_{1,2} = (\Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx})_{1,2}$$

$$= \left[\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} - \begin{pmatrix} \Sigma_{1y} \\ \Sigma_{2y} \end{pmatrix} \Sigma_{yy}^{-1} \begin{pmatrix} \Sigma_{y1} & \Sigma_{y2} \end{pmatrix} \right]_{1,2}$$

$$= \Sigma_{12} - \Sigma_{1y}\Sigma_{yy}^{-1}\Sigma_{y2} = 0 = \Omega_{12}.$$

This means if X_i and X_j are conditionally independent given $X_{-(i,j)}$, then $\Omega_{i,j} = 0$. In fact, the conditional covariance between X_i and X_j given $X_{-(i,j)}$ is

$$\Omega_{ij} = \Sigma_{ij} - \Sigma_{i,(-(i,j))}\Sigma_{(-(i,j)),(-(i,j))}^{-1}\Sigma_{(-(i,j)),j}.$$

The conditional independence assumption is reasonable and often practical. For example, the stocks are usually highly correlated with the market index, and may exhibit high correlation among themselves. However, when conditional on the market index, the correlation may be negligible. While the conditional independence is implied in the inverse covariance matrix, it is better to estimate the inverse covariance matrix instead. The inverse covariance matrix is also called precision matrix. Imposing a sparsity on the inverse covariance is also called a covariance selection model (Dempster, 1972) . It is also called Gaussian Markov graph model or Gaussian Markov Random Field (GMRF) model in machine learning, and has broad application in gene expression, speech recognition and finance.

In addition to the connection to conditional independence of a sparse inverse covariance and its use in reduction of number of parameters, there are two more reasons that why directly estimating inverse covariance matrix is interesting.

(1) Maximum likelihood is a convex problem for $J = P^{-1}$, but not for P. The negative log-likelihood of $X_{p\times n}$ from Gaussian distribution with zero mean and covariance P is:

$$l(P; X) = \log\det(P) + tr(P^{-1}S)$$

is not convex for P. However, let $J = P^{-1}$, then

$$l(J; X) = -\log\det(J) + tr(JS)$$

becomes a convex problem for J. If S is positive definite, it has a unique solution $J = S^{-1}$.

(2) The inverse covariance matrix is directly related to portfolio selection rather than covariance matrix itself. When there are often inevitable estimation error in estimating covariance matrix, this error may be amplified in the inverse covariance matrix.

The sparsity of inverse covariance matrix exhibits in many common examples such as the autocovariance of AR(1) processes. The inverse of auto-covariance of an AR(1) process is a tri-diagonal matrix (see Figure 5.5).

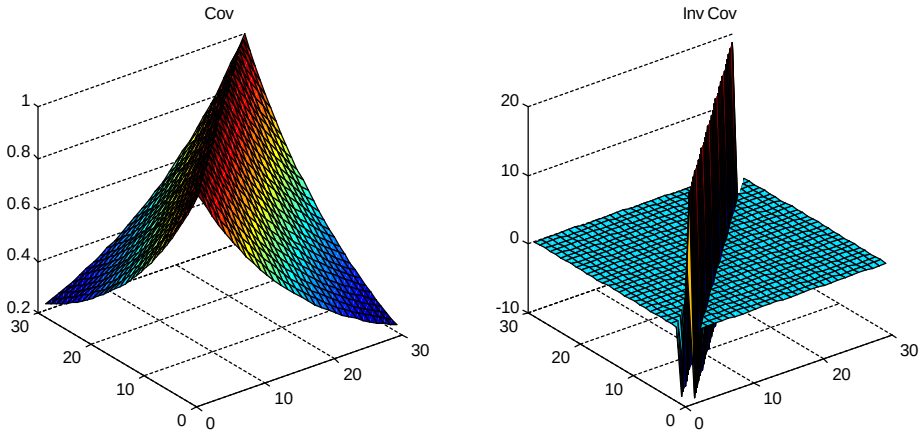


Figure 5.5: AR(1) covariance and its inverse covariance

A general representation of the sparse inverse covariance model is to add a penalty term onto the normal log-likelihood:

$$\max_{J \in S^p} \log\det J - tr(SJ) - \rho \left|\left|D \circ J\right|\right|_1$$

where $||X||_1 = \Sigma_{ij}|X_{ij}|$ is the element-wise l_1 norm of $p \times p$ matrix X, D is a 0-1 matrix include the sparsity constraint and " $\circ$ "is the element-wise product.

There is a vast literature regarding covariance selection models and algorithms such as, to name a few, Huang et al (2006), Friedman, Hastie and Tibshirani (2008), d'Aspremont, Banerjee and Ghaoui (2008), Yuan (2010), and Hsieh et al (2011).

There are also other sparse estimators when the covariance itself is assumed to be sparse such as thresholding method (Karoui, 2008b; Bickel and Lavina, 2008a), we do not discuss here.

5.2.4 Bayesian methods

For covariance matrix estimation of Gaussian distribution, there are several Bayesian estimators including full Bayesian estimators and empirical Bayesian estimators. The first problem is to choose a prior distribution for the covariance matrix. Without loss of generality, we assume $X_{p\times n} = (X_1, ..., X_n)$ is a sample with size n from a p-dimensional Gaussian distribution with zero mean and covariance Σ.

A conjugate family for Σ is the inverse Wishart distribution $IW(\Psi, v)$. The density function of inverse Wishart distribution is

$$f(\Sigma; \Psi, v) = \frac{|\Psi|^{v/2}}{2^{vp/2}\Gamma_p(v/2)} |\Sigma|^{-(v+p+1)/2} \exp\left\{-\frac{1}{2} tr(\Psi\Sigma^{-1})\right\}$$

where $\Gamma_p(\cdot)$ is the multivariate gamma function.

i) The mean and mode are respectively $\Psi/(v-p-1)$ and $\Psi/(v+p+1)$.

ii) If $\Sigma \sim IW(\Psi, v)$, then Σ^{-1} has a Wishart distribution $W(\Psi^{-1}, v)$.

iii) If $X_{p\times n} = (X_1, ..., X_n) \sim N(0, \Sigma)$, the joint distribution of $(X_1, ..., X_n)$:

$$p(X; \Sigma) = \frac{1}{(2\pi)^{np/2} |\Sigma|^{-n/2}} \exp\left\{-\frac{n}{2} tr(\Sigma^{-1} S)\right\}$$

where $S = \frac{1}{n} XX^T$.

If $\Sigma \sim IW(\Psi, v)$, then the posterior distribution of $\Sigma|X$ is $\Sigma|X \sim IW(nS + \Psi, n + v)$. This is based on an important fact that the sampling distribution of the scaled sample covariance matrix $nS = XX^T$ is a Wishart distribution $W_p(\Sigma, n)$.

Therefore, the MMSE and MAP estimate of covariance matrix is:

$$(nS + \Psi)/(n + v - p - 1)$$

and

$$(nS + \Psi)/(n + v + p + 1)$$

These estimates are in fact a weighted average of the sample covariance matrix S and the mean or mode of the prior distribution.

Applying non-informative priors for covariance estimation has been an interesting attempt. The Jeffreys' prior for covariance matrix Σ is (see Press (1982), page 79):

$$\pi_J(\Sigma) \propto |\Sigma|^{-(p+1)/2}$$

However, using Jeffreys' prior fails to achieve appropriate shrinkage of the eigenvalue. Under Stein's loss function, it simply reproduces the sample covariance estimate. Yang and Berger (1994) study the covariance estimation using reference prior. Using a reparametrization (Γ, Λ) where $\Sigma = \Gamma^T \Lambda \Gamma$ and $\Lambda = diag\{\lambda_1, ...\lambda_p)$ with $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$, they show that the reference prior for Σ is:

$$\pi_R(\Sigma) \propto |\Sigma|^{-1} \Pi_{i<j}(\lambda_i - \lambda_j)^{-1}$$

and the posterior distribution is

$$\pi_R(\Sigma|S) \propto \frac{\exp\left\{-n/2 tr(\Sigma^{-1} S)\right\}}{|\Sigma|^{n/2+1} \Pi_{i<j}(\lambda_i - \lambda_j)}$$

They also present a Metropolis algorithm to obtain the posterior estimate of Σ.

There have been also empirical Bayesian approaches to estimation of covariance, which include Efron and Morris (1976) and Haff (1980). However, these do not achieve most desirable shrinkage of eigenvalues. Daniels and Kass (1999, 2001) apply a set of hierarchical priors for the covariance matrix to produce shrinkage of eigenvalues or shrinkage toward certain structure.

5.3 Smooth monotone covariance

When there exists a natural ordering among variables, a particular structure of the covariance is often assumed such as that variables far apart in the ordering have smaller correlation. Furrer and Bengtsson (2007) consider tapering the covariance matrix which is gradually shrinking the off-diagonal entries within the band to 0. Bickel and Levina (2008b) consider banding the covariance matrix (by setting the entries far away from the diagonal to be zeros and keeping the entries within the band unchanged) or banding the inverse covariance matrix, and tapering the covariance matrix. Wu and Pourahmadi (2009) study banded sample autocovariance matrix for AR and MA processes. These estimators can be viewed as thresholding estimators where banding is a hard thresholding and tapering is a soft-thresholding with some known ordering of the random variables.

Malioutov (2011) introduces a nonparametric approach to regularize the covariance by imposing a monotone and smooth structure. The motivation for this approach is to model the covariance structure among the interest rate changes along an interest rate term structure curve, whose correlation exhibits a monotone and smooth effect. However, this approach has a variety of potential applications in empirical finance, such as time series autocorrelation and joint movements of bond futures prices and credit default rates.

Compared to the banded or tapered covariances, the smooth monotone covariance imposes a less strict structure. Under the assumption that random variables are indexed by some natural ordering, one can assume that the covariances (or correlations) between the nearby random variables are higher than those between the distant random variables. Furthermore, smoothness structure is assumed to avoid the stairs-effect of the covariance which are rarely reasonable in practice.

The mathematical formulation of smooth monotone covariance is as follows. Without loss of generality, suppose we have a p dimensional zero-mean random vector $\mathbf{x}(t) = (x_1(t), ..., x_p(t))^T$. Here we ignore the temporal dependence but assume the samples are independent and identically distributed, and we consider the cross-sectional covariance matrix $P^* = E[\mathbf{x}\mathbf{x}^T]$. Assuming that the available length of sample data is n, and that the sample data is $X_{p\times n} = (\mathbf{x}(t_1), ..., \mathbf{x}(t_n))$,then the sample covariance matrix is:

$$Q = \frac{1}{n} X X^T$$

However, when n is comparable to or even smaller than p, this estimate is highly inaccurate. We want to find an estimate P with minimal distance under a prespecified error metric $D(P, Q)$ under monotonicity constraints on the covariance such that $P \in \mathcal{M}$, where

$$\mathcal{M} = \{P | P \succeq 0, P_{ij} \geq P_{ik} \text{ for } i < j < k\}.$$

In Malioutov (2011), the Frobenius norm error $\|P - Q\|_F^2$ was used for the distance measure. To avoid the "stairs-effect" of P, a smoothness penalty on the curvature is added to the objective function, and a discrete version of the Laplacian operator is used:

$$L(P) = \sum_v \nabla_v^2(P),$$

where

$$\nabla_v^2(f) = \sum_{u \in N(v)} (f(x_u) - f(x_v)).$$

Here, $N(v)$ is the set of neighbors of point v. For covariance P_{ij} the neighbors of vertex $v = (i, j)$ can be set to $(i \pm 1, j)$ and $(i, j \pm 1)$. By penalizing the square of the Laplacian operator, we formulate the optimization problem as following:

$$\min_P D(P, Q) + \lambda \sum_v \left(\nabla_v^2(P)\right)^2, \text{ such that } P \in \mathcal{M}$$

where λ is a parameter adjusting the smoothness. For simplicity, in the simulation experiment in this paper, we set λ equal to 1 which implies the equal weight between the accuracy and smoothness. In practice, the parameter is often chosen by cross validation techniques.

When the distance $D(P, Q)$ is convex with respect to P, the problem can be represented as a semidefinite optimization problem:

$$\min_P D(P, Q) + \lambda \left\| D_2 vec(P) \right\|_2, \text{ such that } P \succeq 0,\ Dvec(P) \geq 0.$$

where the operation $vec(P)$ denotes stacking the columns of P into a vector, and the matrices D_2 and D compute the differences of relevant entries of P encoding smoothness and monotonicity constraints respectively. The resulting problem for small p can be readily solved via an interior point method using one of the convex optimization toolboxes for MATLAB such as cvx (Grant and Boyd, 2012). For large p, Malioutov, Corum and Cetin (2012) studied the fast first-order methods for smooth monotone covariances.

Note that it is straightforward to add additional constraints such as positivity of correlations, or $P_{ii} = 1$ for correlation matrices. Without loss of generality, in this paper, we all focus on the correlation matrix

$$R = D^{-1} \Sigma D^{-1}$$

where $D = diag\{\sigma_i\}$, σ_i^2 are the variances (the diagonal elements of the covariance matrix Σ). In practice, the covariance can be decomposed into a correlation matrix and the diagonal elements. This will relieve the difficulty of estimating the covariance matrix when it is highly ill-conditioned because of heteroskedasticity on the diagonals. In addition, to better forecast the time-varying variances, GARCH models are often used to model the volatilities. In this paper, we do not step into those issues but only focus on the correlation matrix.

5.4 Elliptical distributions

5.4.1 Introduction

Elliptical distributions have been an important extension of the Gaussian distributions because they provide a better fit to observed financial returns according to many empirical studies (see Rachev, Menn and Fabozzi (2005)). It is more flexible than the Gaussian distribution to allow for heavier tails and higher peaks.

If the characteristic function of a random vector X has the representation $\phi_X(t) = e^{it^T \mu} \psi(t^T \Sigma t)$, where μ is location parameter, Σ is the dispersion matrix, and ψ is the characteristic generator, we say that X follows elliptical distribution $E_d(\mu, \Sigma, \psi)$. With a specified characteristic generator, the type of distribution is then decided, and a relation between the dispersion matrix and the covariance matrix is known if the covariance exists.

In financial markets, in addition to the observed heavy-tailedness and leptokurtosis, the distribution of financial returns have also been often observed to exhibit negative skewness. The normal variance mean mixture distributions include a lot of examples such as the skewed Student t distribution, generalized hyperbolic, and normal tempered stable distributions,

which provide much more realistic fittings to observed financial returns than the Gaussian distribution. However, to focus on the covariance estimation, in this paper we ignore the skewness part, considering only the normal variance mixture cases which have zero skewness and include typical examples for elliptical distributions such as the multivariate Student t distribution, normal inverse Gaussian distribution, and normal tempered stable distributions (with zero skewness).

Normal variance mixture distributed random variables can be represented as

$$X = \mu + \sqrt{W}AZ \tag{3}$$

where $Z \sim N(0, I)$, $W > 0$ is a scalar random variable independent of Z, μ is the location vector, and A is a lower triangular matrix that determines the dispersion matrix $\Sigma = AA^T$.

The typical elliptical distributions used in empirical finance can be formulated by choosing different distributions for the mixture random variable W. For example:

1. when $W \sim$ Inverse Gamma $(v/2, v/2)$, we say that X follows multivariate Student t distribution $t(\mu, \Sigma, v)$;
2. when $W \sim$ Inverse Gaussian (χ, ψ), X is said to follow a normal inverse Gaussian distribution $NIG(\mu, \Sigma, \chi, \psi)$;
3. when W is a classical tempered stable subordinator $CTS(\theta, \alpha)$, X is said to have a normal tempered stable distribution $NTS(\mu, \Sigma, \theta, \alpha)$ (Rachev et al (2011)).

Therefore, we will discuss the general methods of estimating covariance for elliptical distributions, which can then be applied in fitting the above distributions that are useful in finance.

5.4.2 Multivariate t-distribution

The density of a p-dimensional multivariate t-distribution (or multivariate Student distribution) is given by:

$$f(x; \mu, \Sigma, v) = \frac{\Gamma(\frac{v+p}{2})}{\Gamma(\frac{v}{2})(v\pi)^{\frac{p}{2}}} |\Sigma|^{-\frac{1}{2}} \left(1 + \frac{1}{v}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)^{-\frac{v+p}{2}}$$

It belongs to the normal variance mixture family. Let X be a $p \times 1$ random vector, X has a multivariate t-distribution if

$$X = \mu + \sqrt{W}AZ$$

where (i) $Z \sim N_k(0, I_k)$; (ii) $A \in R^{p\times k}$ such that $\Sigma = AA^T$; (iii) $\mu \in R^p$ is the location parameter; (iv) $W \geq 0$ is a scalar r.v. independent of Z and has a inverse gamma distribution $IG(v/2, v/2)$ where ν is the degrees of freedom.

Kotz and Nadarajah (2004) has a detailed introduction of multivariate t-distributions and their applications. For fixed degrees of freedom and no constraints on Σ, an explicit iteration formula of μ and Σ through EM algorithm is given by:

$$\mu^{(t+1)} = \textstyle\sum_{i=1}^n (w_i^{(t)} x_i) / \sum_{i=1}^n w_i^{(t)}$$

$$\Sigma^{(t+1)} = 1/n \textstyle\sum_{i=1}^n (w_i^{(t)} (x_i - \mu^{(t+1)})((x_i - \mu^{(t+1)})^T)$$

where $w_i^{(t)} = (v+p)/(v+\delta_i^{(t)})$, $\delta_i^{(t)} = (x_i - \mu^{(t)})^T(\Sigma^{(t)})^{-1}(x_i - \mu^{(t)})$.

Particularly, when μ is known to be 0,the iteration of covariance estimate is:

$$\Sigma^{(t+1)} = 1/n \sum_{i=1}^{n}(w_i^{(t)} x_i x_i^T)$$

The detailed derivation of EM algorithm can be found in Liu and Rubin (1995) (see Appendix F).

5.4.3 Multivariate generalized hyperbolic (GH) distribution

A random vector $X_{p\times 1}$ has a multivariate GH distribution if

$$X := \mu + W\gamma + \sqrt{W}AZ$$

where (i) $Z \sim N_k(0, I_k)$; (ii) $A \in R^{p\times k}$; (iii)$\mu, \gamma \in R^p$; (iv) $W \geq 0$ is a scalar r.v. independent of Z and has a Generalized Inverse Gaussian (GIG) distribution $GIG(\lambda, \chi, \psi)$, μ is the location parameter, $\Sigma = AA^T$ is the dispersion matrix, γ is the skewness parameter. (λ, χ, ψ) determines the shape of the distribution. The density function and properties of GH and GIG distribution can be found in McNeil et al (2005).

Note that

$$X|W = w \sim N(\mu + w\gamma, w\Sigma)$$

The GH variables has following properties:

(1) Expectation and covariance: $E[X] = \mu + E[W]\gamma; Cov(X) = E[Cov[X|W]] + Cov[E[X|W]] = var(W)\gamma\gamma^T + E[W]\Sigma$.

(2) Linear transformations: If $X \sim GH_p(\lambda, \chi, \psi, \mu, \Sigma, \gamma)$ and $Y = B_{k\times p}X + b_{k\times 1}$, then $Y \sim GH_p(\lambda, \chi, \psi, B\mu + b, B\Sigma B^T, B\gamma)$.

The GH distribution has various parametrization which can be transformed to each other (Breyman and Luthi, 2013):

(1) $(\lambda, \chi, \psi, \mu, \Sigma, \gamma)-$Parametrization:

$W \sim GIG(\lambda, \chi, \psi)$.It has a draw back because $GH_p(\lambda, \chi, \psi, \mu, \Sigma, \gamma)$ and $GH_p(\lambda, \chi/k, k\psi, \mu, k\Sigma, k\gamma)$ are identical.

(2) $(\lambda, \bar{\alpha}, \mu, \Sigma, \gamma)-$Parametrization:

There is a way to eliminate the degree of freedom by constraining the determinant of the dispersion matrix Σ to 1. Here, we simply require $E[W] = 1$.

For $W \sim GIG(\lambda, \chi, \psi)$, $E[W]$ exists if $\psi > 0$. If $\psi \to 0$, the GIG distribution approaches the Inverse Gamma distribution, then $E[W]$ only exists if $\gamma < -1$.

Define

$$E[W] = \sqrt{\frac{\chi}{\psi}}\frac{K_{\lambda+1}(\sqrt{\chi\psi})}{K_\lambda(\sqrt{\chi\psi})} = 1$$

and set

$$\bar{\alpha} = \sqrt{\chi\psi}.$$

Switching from $(\lambda, \bar{\alpha}, \mu, \Sigma, \gamma)$ to $(\lambda, \chi, \psi, \mu, \Sigma, \gamma)$:

$$\psi = \bar{\alpha}\frac{K_{\lambda+1}(\bar{\alpha})}{K_\lambda(\bar{\alpha})},\ \chi = \frac{\bar{\alpha}^2}{\psi} = \bar{\alpha}\frac{K_\lambda(\bar{\alpha})}{K_{\lambda+1}(\bar{\alpha})}, \Sigma = \Sigma, \gamma = \gamma$$

When $\lambda = -\frac{1}{2}, X$ is said to have the NIG distribution, in which case $\psi = \chi = \bar{\alpha}$.

Switching from $(\lambda, \chi, \psi, \mu, \Sigma, \gamma)$ to $(\lambda, \bar{\alpha}, \mu, \Sigma, \gamma)$:

Set $k = \sqrt{\frac{\chi}{\psi}} \frac{K_{\lambda+1}(\sqrt{\chi\psi})}{K_\lambda(\sqrt{\chi\psi})}$.

$$\bar{\alpha} = \sqrt{\chi\psi},\ \Sigma \equiv k\Sigma,\ \gamma \equiv k\gamma$$

Multivariate GH distributions can be estimated using EM type algorithms by making use of its normal mean-variance mixture representation. (see McNeil, Frey and Embrechts, 2005; Protassov (2004), Hu (2005), and Breymann and Luthi (2008)). In particular, for the NIG distribution where the mixture variable follows an inverse Gaussian distribution, see Karlis (2002).

5.4.4 Multivariate normal tempered stable (MNTS) distribution

The multivariate normal tempered stable (MNTS) distribution is a normal variance-mean mixture distribution with a classical tempered stable (CTS) subordinator and has been applied in finance (see Kim et al, 2012). Let $\alpha \in (0, 2)$ and $\theta > 0$, if the characteristic function of a non-Gaussian infinitely divisible random variable W is given by

$$\phi_W(u) = \exp\left(-\frac{2\theta^{1-\alpha/2}}{\alpha}\left((\theta - iu)^{\alpha/2} - \theta^{\alpha/2}\right)\right)$$

and let

$$X = \mu + \beta(W - 1) + \sqrt{W}AZ$$

$$X|W = w \sim N(\mu + \beta(w - 1), w\Sigma)$$

where $\Sigma = AA^T$. X is said to follow a multivariate normal tempered stable distribution $MNTS(\mu, \Sigma, \beta, \theta, \alpha)$.

The covariance of X is related to Σ :

$$\begin{aligned} cov(X) &= Cov[E[X|W]] + E[Cov[X|W]] \\ &= E[(W - 1)^2]\beta\beta^T + E[W]\Sigma \end{aligned}$$

Through simple calculation, we have $E[W] = 1$, and $E[(W - 1)^2] = \frac{2-\alpha}{2\theta}$, then

$$cov(X) = \Sigma + \frac{2 - \alpha}{2\theta}\beta\beta^T$$

When β equals to 0, the MNTS distribution is an elliptical distribution and $cov(X) = \Sigma$. Since the density for MNTS distribution is intractable, sample covariance is often plugged into $cov(X)$. However, sample covariance is known to be a poor estimate when n is comparable to p or $n > p$. In addition, sample covariance is not robust for heavy-tailed distributions where there are likely more outliers.

5.5 Covariance estimation methods for elliptical distributions

Since the sample covariance (or Pearson's correlation) is not robust to the outliers, it is generally even worse for the heavy-tailed elliptical distributions. Therefore, alternative estimators are needed. In this subsection, we review some alternative estimators of the covariance (or dispersion) matrix for elliptical distributions.

Under the Gaussian assumption, the sample covariance coincides with the ML estimator. However, for elliptical distributions, these two estimates are generally different. For the

normal variance mixture distributions, the EM algorithm is very useful. Note that when W is known, then $X|W = w$ follows a Gaussian distribution $N(\mu; w\Sigma)$,and the density of X can be written as $p_N(X|W; \mu, w\Sigma)p(W; \theta_W)$. In each iteration, the EM algorithm first treats W as a known hidden variable, calculates the expectation of the log-likelihood with the known W, and then finds the next W by maximizing the expected log-likelihood. For the estimation of multivariate student t distribution, see Liu and Rubin (1995), Kotz and Nadarajah (2004). For the estimation of generalized hyperbolic distribution, see McNeil, Frey, and Embrechts (2005), Protassov (2004), Hu (2005), and Breymann and Luthi (2008). In particular, for the NIG distribution, see Karlis (2002). For MNTS distribution, the multivariate density function is highly intractable, and parameters are usually estimated separately for each distribution, and sample covariance is used for solving the dispersion matrix.

Another solution is to separate the covariance matrix into variance components and the correlation matrix. The marginal variances are obtained by sample variance or time-varying volatility models in the non i.i.d. cases. For the normal distribution, there is a well-known relation between the linear correlation and Kendall's tau $\rho = \sin(\tau\pi/2)$. Lindskog, Mcneil, and Schmock (2003) show that this relation also holds for all elliptical distributions. Given that the Kendall's tau, as a rank correlation coefficient, is less sensitive to the outliers, we can use it to obtain the linear correlation and thus the covariance.

Another covariance estimation method for elliptical distributions is Tyler's method (1987), whose idea is estimating with the normalized samples instead of the original samples. Let $\{\mathbf{x}_1, ..., \mathbf{x}_n\}$ be n random vectors drawn from a p dimensional elliptical distribution and, without loss of generality, assume the mean is zero. Then, $\mathbf{x}_i$ has the representation $\mathbf{x}_i = \sqrt{W_i}\mathbf{u}_i$, where $\sqrt{W_i}$ is the mixture random variable, and $\mathbf{u}_i$ is a $p \times 1$ zero-mean jointly Gaussian random vector with covariance $\Sigma = AA^T$.

Denote the normalized sample by $\mathbf{s}_i = \mathbf{x}_i / ||\mathbf{x}_i||_2$, then the density of $\mathbf{s}_i$ is given by (Frahm (2004)):

$$f(\mathbf{s}_i; \Sigma) = \frac{\Gamma(p/2)}{2\pi^{p/2}} \sqrt{\det(\Sigma^{-1})} \left(\mathbf{s}_i^T \Sigma^{-1} \mathbf{s}_i\right)^{-p/2},$$

and a fixed-point representation of the maximum likelihood solution of Σ when $n > p$ is (Tyler (1987)):

$$\Sigma = \frac{p}{n} \sum_{i=1}^{n} \frac{\mathbf{s}_i \mathbf{s}_i^T}{\mathbf{s}_i^T \Sigma^{-1} \mathbf{s}_i}.$$

The solution can be found through fixed point iterations (Chen, Wiesel, and Hero (2011)):

$$\widehat{\Sigma}_{j+1} = \frac{p}{n} \sum_{i=1}^{n} \frac{\mathbf{s}_i \mathbf{s}_i^T}{\mathbf{s}_i^T \widehat{\Sigma}_j^{-1} \mathbf{s}_i}.$$

Chen, Wiesel, and Hero (2011) proposed the shrinkage covariance estimates for elliptical distributions based on the Tyler's method. Hero (2012) further proposed a unified framework for regularized covariance estimation in scaled Gaussian models.

6 Smooth monotone covariance for elliptical distributions and applications in finance

6.1 Introduction

Covariance estimation is an important problem in risk management, portfolio selection, and asset pricing. Sample covariance has been widely known as a poor estimate when the sample size n is close to or less than the dimension p. This problem is frequently faced by investors because the investment universe is very large relative to the historical return data available to estimate the covariance. Therefore, in the asset management community, the covariance estimation problem for "large p small n" situations has attracted considerable attention.

The classical results in random matrix theory have shown that the eigenvalue spectrum is biased with p/n being fixed as $n \to \infty$, which asymptotically follows the Marcenko-Pastur law (Marcenko and Pastur, 1967). The theory states that the eigenvalues for the sample covariance are more dispersed than those of the true covariance. In the portfolio selection problem, researchers have dealt with this by adjusting the eigenvalues using different methods such as truncating the smallest eigenvalues (Elton and Gruber, 1973), adjusting the principal components in some interval (Plerou et al 1999, Laloux et al 1999, Conlon et al 2007, Kwapien et al 2006), or imposing constraints on the condition number of the covariance matrix such that it is well-conditioned (Won, 2006). Park and O'Leary (2010) apply Tikhonov regularization to filter the eigenvalues and place various estimators for adjusting the eigenvalues in a common framework which differs only in the choice of the filtering function. Shrinkage estimates are another way to adjust the eigenvalues such as weighting the sample covariance matrix with a structured covariance matrix (Ledoit and Wolf, 2003, 2004) such as the identity matrix or constant correlation model, or shrinking toward a truncated covariance matrix (Bengtsson and Holst, 2002). Because these shrinkage estimates are adjusting the eigenvalues by a linear rate, Karoui (2008) and Ledoit and Wolf (2012) propose nonlinear shrinkage estimators by estimating the spectral distribution from the sample spectral distribution, which is an inverse problem to the Marcenko-Pastur law.

However, estimators that only adjust the eigenvalues still fail to provide consistent covariance estimates because the eigenvectors are also inconsistent estimates in "large p, large n" asymptotics (Silverstein, 1995). To reduce the large number of parameters relative to the sample size, there are many solutions proposed by imposing a sparse or a low-dimensional structure on the covariance such as factor models (Rubin 1982, Fan et al 2008), sparse inverse covariance (d'Aspremont et al 2008, Rothman et al 2008, Yuan and Lin 2007, Friedman et al 2008), and thresholding methods (Bickel and Levina 2008a).

Another class of regularization methods relies on a natural ordering among the random variables. This is common in finance, examples being autocorrelated time series, bonds with different maturities, bonds with different ratings, options with respect to different strikes or maturities. Furrer and Bengtsson (2007) consider "tapering" the sample covariance matrix by gradually shrinking off-diagonal elements toward zero. Wu and Pourahmadi (2003) and Huang et al (2006) use the Cholesky decomposition of the covariance matrix to "band" the inverse covariance matrix, which assumes the conditional independence between faraway components. Bickel and Levina (2008b) considered banding the sample covariance matrix or estimating a banded version of the inverse population covariance matrix. The models for banding the inverse covariance matrix are also known in the field of statistics as covariance selection and in the field of machine learning as Gaussian graphical model or Gaussian Markov Random Field (MRF). Parametric models assume a functional form for

the covariance such as an exponential or power-law decay; Rasmussen and Williams (2006) provide a general framework for such models. However, despite the successful applications of these banded or parametric models, in general they still impose very strict structures on the covariance. Malioutov (2011) proposes a nonparametric regularization method by imposing monotone constraints and a smoothness penalty when estimating the covariance matrix, which he calls a smooth monotone covariance. He shows the outperformance of this method in terms of Frobenius norm error to the sample covariance or principal component analysis for the term-rate covariances by using an interpolated smooth and monotone variant of the sample covariance as the true covariance.

In this paper, we first discuss the smooth monotone covariance by using various covariance distance measures under Gaussian assumptions, which is usually assumed in the vast literature about covariance estimation. In the literature, the Frobenius norm error (or mean-squared error (MSE)) has been mainly used for both the objective function in covariance estimation and evaluating the performance against the true covariance. However, Fan et al (2008) showed that Frobenius norm error may not be a good measure to tell the difference between two covariance estimates. They also used a scaled quadratic loss or Frobenius norm error of inverse covariance to measure distances. We will study various loss functions including Frobenius norm error, quadratic loss, entropy loss (or, Stein's loss), Kullback-Leibler divergence and Hellinger distance together for evaluating the performance of our smooth monotone covariance estimates. We will also explore whether improved estimates can be obtained by replacing the Frobenius norm error in the objective function with a more statistically meaningful distance measure.

In finance, non-Gaussian distributions are more interesting because of the heavy-tailedness of financial returns. We will review the usual solutions for estimating covariance matrix under elliptical distributions assumptions which serve as a broad family of particular distributions such as the Student t distribution, normal inverse Gaussian distribution, and the like. The typical estimation methods include sample covariance, Kendall's tau transformation method, maximum likelihood estimation, and Tyler's method (Tyler, 1987). For heavy-tailed elliptical distributions, the sample covariance and linear correlation become very sensitive to the presence of outliers, and rank correlations such as Kendall's tau are more robust estimates. Therefore, an intuitive way is to decompose the covariance matrix into variance components and correlation matrix, and estimate them separately where the latter can be obtained by transforming from rank correlations. Lindskog, Mcneil, and Schmock (2003) show that the well-known relation between the linear correlation and Kendall's tau under Gaussian assumptions exists for all elliptical distributions as well. However, the problem of using the Kendall's tau transformation to form a correlation matrix is that the resulting matrix might not be positive definite. A more classical way of estimating covariance matrix is to use the maximum likelihood estimation (MLE) method. Unlike Gaussian distributions, for elliptical distributions the ML estimates are usually different from the sample covariance. For a broad class of elliptical distributions - the normal variance mixture distributions, the ML estimate can be obtained readily using an expectation-maximization (EM) algorithm. However, the MLE method requires an assumption for the type of distribution and the existence of its probability density function. An easier but also robust way to estimate the covariance is using Tyler's method (Tyler, 1987; Chen, Wiesel and Hero, 2011) by working on the normalized samples. We will investigate these four types of covariance estimates using simulation studies, and incorporate them into the smooth monotone covariance estimation by using them as initial estimates in the optimization.

The paper is structured as follows. Section 2 reviews a variety of covariance distances,

and investigates the performance of smooth monotone covariance for the Gaussian distribution under various distances using simulation studies. In addition, the improvement of smooth monotone covariance in portfolio risk measurement and the generalization of the smooth monotone covariance by minimizing alternative distances are studied. In Section 3, we compare various covariance estimates for elliptical distributions and study the performance of their corresponding smooth monotone covariance estimates. In Section 4, we apply our methodology to Eurodollar futures and corporate bonds portfolios, and compare the performance of the smooth monotone covariance and the sample covariance.

6.2 Covariance estimation for Gaussian distributions

6.2.1 Distance measures

In the objective function of smooth monotone covariance, the distance between the covariance estimate and the sample covariance (or, any other initial covariance) is needed. When evaluating the performance of the covariance estimate, its distance against the true covariance is also needed. In the following, we denote the estimated covariance by P, the sample covariance by Q, and the true covariance by P^*.Note that the sample covariance Q may not be full rank (e.g., $p > n$) so some well-known distances such as Kullback Leibler divergence from an estimate to the sample covariance might not be well-defined. In addition, we would like to have $D(P, Q)$ as a convex function with respect to P so that we can formulate the problem as a convex optimization problem, which is much easier to solve. So special attention is needed when selecting a good measure for $D(P, Q)$.The simplest and easiest distance is the Frobenius norm error (or, MSE):

$$||P - Q||_F^2 = tr((P - Q)^T(P - Q))$$

This measure has been mainly used as an objective function for covariance estimation and also for evaluating the performance of a covariance estimate in the literature (see Ledoit and Wolf (2003, 2004, 2012), Chen et al (2009, 2011)). Despite the intuition of minimizing mean-squared error, to the best of our knowledge, there is no direct connection to classical likelihood-based statistical inference, and no justification in the literature for preferring this error metric over alternative ones in applications such as portfolio risk measurement.

In portfolio selection, the variance of the portfolio returns is frequently used as a proxy for portfolio risk. For elliptical distributions, the value-at-risk (VaR) can also be represented as a function of the variance. Therefore, the error of the portfolio variance estimate is more interesting than the Frobneius norm error of the covariance estimate. Fan et al (2008) reported that their proposed covariance estimate performs almost as well as the sample covariance when computing the portfolio risk for an equal-weighted portfolio, but greatly outperforms the sample covariance for a Markowitz optimal weighted portfolio risk.

Under Gaussian assumptions, the maximum absolute error of the variance between two

portfolios with the two covariances P and Q (normalized by $||w||_2 = w^T w$) is:

$$\begin{aligned} & \max_w \left| \frac{w^T P w - w^T Q w}{w^T w} \right| \\ = & \max_w \left(\frac{w^T (P-Q) w}{w^T w}, \frac{w^T (Q-P) w}{w^T w} \right) \\ = & \max(\lambda_{\max}(P-Q), \lambda_{\max}(Q-P)) \\ = & \max(\lambda_{\max}(P-Q), -\lambda_{\min}(P-Q)) \\ = & \max |\lambda_i(P-Q)| = \sigma_1(P-Q) \\ = & ||P-Q||_2 \end{aligned}$$

where $||\cdot||_2$ is the the matrix 2-norm. The maximum is reached at $w = v_1(P-Q)$ or $w = -v_p(P-Q)$. Note the connection between the squared eigenvalues of $(P-Q)$ and the squared Frobenius norm of $(P-Q)$:

$$\begin{aligned} \|P-Q\|_F^2 &= tr((P-Q)^2) = \sum_i \lambda_i((P-Q)^2) \\ &= \sum_i \lambda_i^2(P-Q) \end{aligned}$$

Therefore, minimizing the Frobenius norm of $P-Q$ constrains the squared eigenvalues of $(P-Q)$, thereby constraining the maximal absolute value of the eigenvalues of $(P-Q)$. Assuming P is the estimated covariance and Q is the true covariance, it then constrains the error between the estimated portfolio risk and the true portfolio risk. Therefore, the Frobenius norm error is a reasonable measure for covariance estimation when one needs to estimate the portfolio risk.

Apart from the Frobenius norm error, there are also several other measures related to statistical inference. To simplify the discussion, we first focus on the Gaussian distributions. Without loss of generality, we assume that the mean is known to be zero. Suppose $\mathbf{x}$ is a p-dimensional zero-mean random vector from $N(0, P)$:

$$f(x; P) = (2\pi)^{-p/2} |P|^{-1/2} \exp\left\{ -\frac{1}{2} x^T P^{-1} x \right\}$$

If $X = (\mathbf{x}_1, ..., \mathbf{x}_n)$ are n identical and independently distributed (i.i.d.) samples from this distribution, the log-likelihood is (omitting the constants):

$$\begin{aligned} L(X; P) &= -\frac{n}{2} \log |P| - \frac{1}{2} \sum_{i=1}^{n} x_i^T P^{-1} x_i \\ &= -\frac{n}{2} \log |P| - \frac{n}{2} tr(P^{-1} Q) \end{aligned}$$

where $Q = \frac{1}{n} X X^T$ is the sample covariance matrix.

The maximum likelihood estimates can also be understood as minimizing Kullback-Leibler divergence, which is a non-symmetric information theoretic measure of the difference between two probability distributions. Assuming the density functions of two distributions F and G are $f(x)$ and $g(x)$, then the Kullback-Leibler divergence from F to G is defined as:

$$D_{KL}(F||G) = \int f(x) \log \frac{f(x)}{g(x)} dx$$

Let $N(0,Q)$ be the empirical distribution, and $N(0,P)$ be the estimated distribution, to simplify the notations, we write the Kullback-Leibler divergence between two zero-mean Gaussian distributions using their covariances. Then, the Kullback-Leibler divergence from P to Q is:

$$D_{KL}(P||Q) = \frac{1}{2}(tr(PQ^{-1}) - \log|PQ^{-1}| - p).$$

which is also called Stein's loss (Stein (1956), Yang and Berger (1994)) or Entropy loss (omit the constant 1/2), which is convex with respect to P.

However, maximum likelihood estimate of P is obtained by minimizing the Kullback-Leibler divergence from Q to P:

$$D_{KL}(Q||P) = \frac{1}{2}(tr(P^{-1}Q) - \log|P^{-1}Q| - p)$$

which is convex with respect to P^{-1} but not for P, and is not appropriate for our smooth monotone regularization because the objective is not convex.

Another loss function that appeared in the literature (Yang and Berger (1994), Haff (1980), and Fan et al (2008)) is the so-called quadratic loss (QL):

$$D_{QL}(P||Q) = tr((PQ^{-1} - I)^2)$$

By some matrix calculation, we can show that the QL is proportional to a second order approximation of KL in a neighborhood of the true covariance (see Haff (1980), Appendix G):

$$D_{KL}(P||Q) \approx cD_{QL}(P||Q)$$

The above distances are in fact not a metric, which needs to be symmetric. Therefore, we also consider another important statistical distance: Hellinger distance. Hellinger distance is used to quantify the similarity between two probability distributions. The Hellinger distance between two probability distributions F and G with densities $f(x)$ and $g(x)$ is defined as

$$\begin{aligned} H^2(F,G) &= \frac{1}{2}\int((f(x)^{1/2} - g(x)^{1/2})^2 dx \\ &= 1 - \int f(x)^{1/2} g(x)^{1/2} dx. \end{aligned}$$

The Hellinger distance has several nice statistical properties. First, it is a true metric satisfying symmetry and triangular inequality. Second, its value is bounded between 0 and 1. This provides more intuitive meaning of how close two distributions are when one only looks at the value of the distance. The closer the Hellinger distance is to one, the less likely that two samples come from the same distribution. In contrast, Kullback-Leibler divergence is not symmetric, and only has relative meaning in comparing two estimates.

By some calculations (Appendix H), the Hellinger distance between two Gaussian distributions with zero mean and covariances P and Q is (similarly we use the covariance P and Q to :

$$H^2(P,Q) = 1 - |P|^{1/4}|Q|^{1/4}\left|\frac{P+Q}{2}\right|^{-1/2}$$

In most of our experiments later in this paper, we will use the Frobenius norm error (MSE), Kullback-Leibler divergence from P to Q (KL), quadratic loss (QL), and Hellinger distance (Hel) to evaluate the performance of the smooth monotone covariance over sample covariance against the true covariance.

6.2.2 Comparing sample and smooth monotone covariance using various distances

In this subsection, we compare the estimation errors of sample covariance and smooth monotone covariance against the true covariance using various distances including the MSE, QL, KL, and Hellinger distance. Without loss of generality, for all the simulations, we focus on the correlation estimation, for which the diagonal elements are normalized to be one. For our simulated-data experiments we focus on the autocovarance matrix of an autoregressive process with order 1 (AR(1) process):

$$P^*(i,j) = \rho^{|i-j|}.$$

As this covariance satisfies the smooth and monotone assumption and is frequently used as a simple example in the literature (e.g., Bickel and Levina (2008b), Chen et al (2011)), we use it as our example throughout the simulation studies in this paper to demonstrate the benefit of smooth monotone covariance. However, we do not aim to use this estimate for estimating the covariance of an AR(1) process because with known parametric structure of the covariance one can use other methods to estimate ρ more accurately. In general, the true structure of a covariance is often unknown and it might be too strict for one to impose an assumed parametric structure on the covariance.

In the following experiment, we set the dimension $p = 30$, and the true covariance be the autocovariance matrix of an AR(1) process with autocorrelation $\rho = 0.9$, which is a Toeplitz matrix with the first column being $(1, \rho, \rho^2, ..., \rho^{p-1})^T$. We conduct $M = 200$ trials at each sample size n varying from 40 to 300. Using the sample correlation as an input for the smooth and monotone estimator, we obtain the smooth monotone correlation. Then, we compute the average distance measure between the estimated covariance P and the true covariance P^* using the four types of distances defined earlier: MSE, QL, KL, and Hellinger distance.

The results are shown in Figure 6.1. It shows that the smooth monotone covariance uniformly outperforms the sample covariance for all four measures. This is important because the usual performance measures used in the literature for covariance comparison are mostly the MSE, but the other distances with direct connection to basic statistical estimation principles are not investigated. From this experiment, we see that the smooth monotone covariance not only reduces the MSE, but also reduces the other errors. In fact, the reduction of error is more significant when using other measures than using MSE.

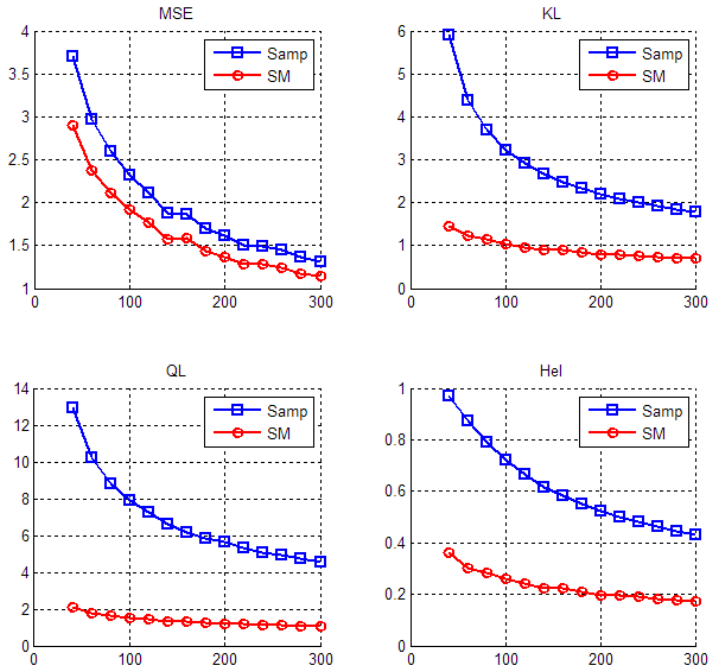


Figure 6.1: Estimation errors of sample and smooth monotone covariance

Fan et al (2008) point out that the MSE may not be a good measure to evaluate the performance of their factor-model based estimation of the covariance matrix. They show that under QL and MSE of the inverse covariance, the performance is much better than the sample covariance. Here we see similar results for our smooth monotone covariance that the reduced errors are much more significant when using the QL, KL, and Hellinger as distance measures. However, we do not claim MSE is not a good measure for our smooth monotone covariance but recommend using various distance measures to evaluate the covariance when the errors under MSE is not significant. When ρ is smaller (e.g., $\rho = 0.7$), we find our smooth monotone covariance greatly outperforms the sample covariance under MSE as well.

6.2.3 Comparison of sample and smooth monotone covariance on portfolio risk measurement

Although there are numerous distance measures for comparing the covariances, in practice what matters more is whether it leads to a more accurate estimate for the portfolio risk, or better performance in portfolio selection. The importance of covariance estimation also varies under different objectives. For an equally weighted or random portfolio, the sample covariance or a factor risk model often provides good estimates for portfolio risk (Saxena and Stubbs, 2010). Fan et al (2008) show that the factor model and sample covariance have almost the same performance for portfolio risk when applied to an equally weighted portfolio, but the factor model has much less error for an optimal weighted portfolio.

In the following, we conduct a similar experiment as Fan et al (2008). We simulate $M = 500$ samples, with sample size varying from $n = 40$ to $n = 200$, from a $p = 30$

dimensional Gaussian distribution where covariance is an AR(1) autocovariance matrix with $\rho = 0.7$. We note that for $\rho = 0.7$, the sample covariance performs even worse than with $\rho = 0.9$, and we choose 0.7 to show that even in such a difficult case with equally-weighted portfolios we do not suffer much when estimating portfolio variances. However for optimized portfolios, the sample covariance performs poorly even for $\rho = 0.9$. We estimate the portfolio risk by sample covariance and smooth monotone covariance, and compare the estimated risk and true risk, where we use the portfolio standard deviation as a proxy for portfolio risk.

Figure 6.2 presents the mean relative estimation error of portfolio risk with respect to varying sample size. In the left plot, we calculate the portfolio for an equally-weighted portfolio. It shows that sample covariance and smooth monotone covariance have roughly the same performance. Both of them estimate the true portfolio risk very well (relative error is less than 8% for all sample size from 40 to 200). In the right plot, we calculate the error for an optimal weighted portfolio. The weights are determined by $w = P^{-1}\mu$, where we assume μ be a vector with all equal elements 0.05.We find that the smooth monotone covariance leads to much smaller error than the sample covariance.

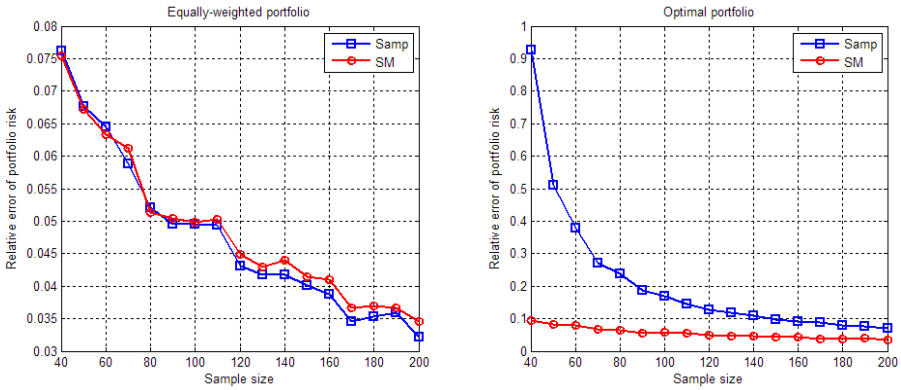


Figure 6.2: Estimation error of sample and smooth monotone covariance for equally-weighted portfolio risk and optimal portfolio risk

In fact, it is known that the sample covariance actually underestimates the risk of optimal portfolios and is a poor input for the Markowitz portfolio selection (see Muller 1993, Karoui 2008). Michaud (1989) argues that optimization in these cases is in fact an "error-maximization" process that will overestimate the expected returns and underestimate the variances. In the following experiment, we artificially create a portfolio consisting of $p = 30$ financial instruments with expected returns ranging from -0.05 to 0.05. The covariance is AR(1) autocovariance with $\rho = 0.7$. We generate the returns from the joint Gaussian distribution with sample size $n = 50$, and calculate the mean-variance optimal portfolio with target portfolio return μ^* ranging from 0 to 0.05.

The mean-variance optimal portfolio problem in the example is:

$$\min w^T \Sigma w$$

$$s.t. w^T \mu = \mu^*, w^T 1 = 1, w > 0$$

For each simulated sample, we calculate the mean-variance efficient frontier using the sample covariance and smooth monotone covariance and repeat the experiment for $M = 200$

times. We then compare the difference between the optimal portfolio risk and the true portfolio risk $\widehat{\sigma} - \sigma^*$ using sample covariance and smooth monotone covariance. In Figure 6.3, we plot the 10%, 25%, 50%, 75%, and 90% quantiles of the difference between the estimated portfolio risk and true portfolio risk for 20 portfolio target return levels ranging from 0 to 0.05. As can be seen from Figure 6.3, using the sample covariance will underestimate the portfolio risk, while the smooth monotone covariance has a much smaller error.

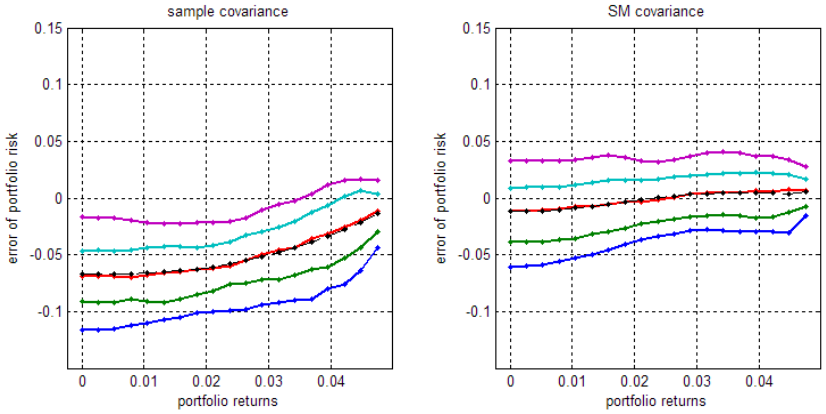


Figure 6.3: Estimation errors of portolio risk using sample and smooth monotone covariance in Markowitz portfolio selection

In the literature, there are also a variety of portfolio optimization methods using alternative risk measures or utilities such as mean VaR, and mean-CVaR. VaR is defined as the possible loss for a trading position or portfolio under a given confidence level α, and has become a popular risk measure for setting bank capital requirements. CVaR (conditional value-at-risk), also called expected-tail loss (ETL), average value-at-risk (AVaR) or expected shortfall (ES), is the expectation of losses given that the losses exceed the VaR: $CVaR_\alpha = E[L|L > VaR_\alpha]$. For elliptical distributions, since the smooth monotone covariance has less estimation error for portfolio variance, it naturally leads to less estimation error for these alternative risk measures as well because VaR and CVaR are in fact functions of the variance for elliptical distributions.

6.2.4 Generalization of smooth monotone covariance using alternative distances

In this subsection, we consider generalizing the smooth and monotone covariance by using various alternative distances as the objective function such as Kullback-Leibler divergence. However, Hellinger distance $H(P, Q)$ and the Kullback-Leibler divergence from Q to P (i.e., $D_{KL}(Q||P)$) is not convex with respect to P. The Quadratic loss $D_{QL}(P||Q)$ and the Kullback-Leibler divergence from P to Q (i.e., $D_{KL}(P||Q)$, or the Stein's loss) are two convex functions with respect to P, and can be used as alternative distances in our smooth monotone estimates.

In Section 3.1, we showed that the matrix 2 norm $||P - Q||_2$ is the maximum absolute error of the normalized portfolio variances. Minimizing the 2 norm of $P - Q$ can also be transformed to a semidefinite programming problem as follows:

$$\min_{P,s} s$$

$$\text{s.t. } -sI \prec (P-Q) \prec sI$$

When there are no monotone or smoothness constraints, we know that the estimates under all norms recover the sample covariance exactly. By imposing the monotone constraints and smoothness penalty, we can generalize the smooth monotone covariance by minimizing the alternative distances. However, the smoothness penalty parameter may be difficult to choose when minimizing KL or QL. From optimization theory, instead of minimizing the joint objective, we can consider the constrained form, where we move the smoothness from the objective function to the constraint. We can see that the smoothness parameter in the joint form serves as a Lagrangian multiplier for the smoothness constraint in the second form. To compare the performance of these different estimates, we formulate the smoothness penalty as a constraint for the three different problems.

Let Q be the sample covariance, and P be the covariance estimate, then the optimization problems of smooth monotone correlation under the Frobenius norm error, 2 norm error, KL and QL are respectively formulated as follows:

- Minimizing Frobenius norm:

$$\min_P \ ||P-Q||_F^2,$$
$$\text{s.t. } P \succ 0, diag(P) = 1, P \in \mathcal{M}, \sum(\nabla^2 P)^2 \leq \alpha$$

- Minimizing 2 norm:

$$\min_{P,s} s$$
$$\text{s.t. } -sI \prec (P-Q) \prec sI$$
$$P \succ 0, diag(P) = 1, P \in \mathcal{M}, \sum(\nabla^2 P)^2 \leq \alpha$$

- Minimizing KL:

$$\min_P D_{KL}(P||Q) = tr(PQ^{-1}) - \log\left|PQ^{-1}\right| - p$$
$$\text{s.t. } P \succ 0, diag(P) = 1, P \in \mathcal{M}, \sum(\nabla^2 P)^2 \leq \alpha$$

- Minimizing QL:

$$\min_P D_{QL}(P||Q) = tr((PQ^{-1} - I)^2)$$
$$\text{s.t. } P \succ 0, diag(P) = 1, P \in \mathcal{M}, \sum(\nabla^2 P)^2 \leq \alpha$$

For minimizing KL and QL, the inverse of the sample covariance matrix is needed, which is undefined when $p > n$. Therefore, for the following simulations, we focus on the well-defined cases ($n > p$).

In the following, for all cases we set the smoothness constraint parameter $\alpha = 1$ to make sure the constraints are the same. Note that α and λ are different, and it is not trivial to find a direct relation between α and λ. The true covariance in this experiment is AR(1) covariance with $\rho = 0.9$ and dimension $p = 30$. The sample size varies from 40 to 200. Figure 6.4 shows the distance from the optimum covariance estimates to the sample covariance using MSE, 2-norm, QL, and KL as the objective function, respectively. This is used to confirm that the optimization is correct — for training data, when the distance measure used for optimization is the same as the one used for evaluation, it should achieve the minimal distance.

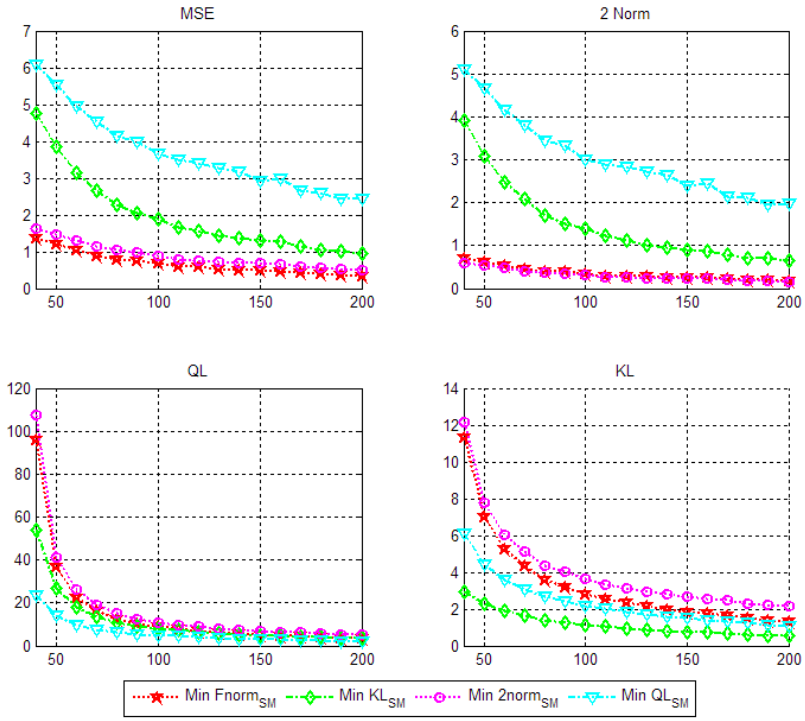


Figure 6.4: Distances of various smooth monotone covariance estimates against the sample covariance

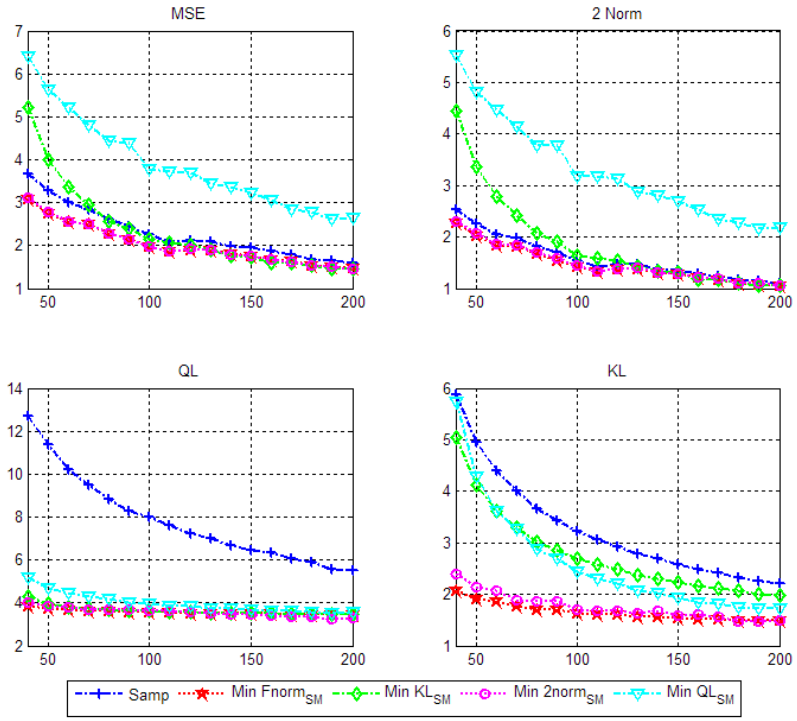


Figure 6.5: Distances of various smooth monotone covariance estimates against the true covariance

However, when making comparisons to the true covariance, this may not be true due to the finite sample effect. Figure 6.5 shows the distances of optimal covariance estimates against the true covariance. We can see that, when comparing the obtained covariance estimates with the true covariance, minimizing KL or QL does not lead to minimal distance even when it is evaluated by KL or QL. Instead, when minimizing MSE or 2-norm, the errors against the true covariance under any of the four distances are much smaller.

The results in Figures 4 and 5 also show that when minimizing the QL or KL to the sample covariance under smooth and monotone constraints, it leads to much larger MSE and 2 norm errors. In fact, the performance can be even much worse than the sample covariance! This is likely caused by the fact that computing KL and QL requires inverting the sample covariance matrix, which is very sensitive for small samples. In addition, the estimates that are close under KL and QL might be very far away under MSE and 2 norm (and vice versa). Therefore, minimizing KL or QL against the sample covariance matrix is not a good choice.

Comparing the results of minimizing MSE and 2 norm, we see that they are very close to the true covariance for any of the four distances, suggesting both of them are more robust estimates. The similarity of the results between minimizing MSE and 2 norm is natural because the squared Frobenius norm is the sum of singular values of $(P-Q)^2$, while the squared 2 norm is the largest singular value of $(P-Q)^2$. In some extreme cases (e.g., AR(1) with $\rho = 0.9999$), the difference between minimizing Frobenius norm and 2-norm may become substantial, but in typical cases they lead to very similar answers. Therefore, we will use Frobenius norm in the studies performed in the rest of this paper.

6.3 Covariance estimation for elliptical distributions

Under the classical Markowitz portfolio selection framework, a representation of the portfolio returns and portfolio variances are needed. The elliptical distributions fit well within this framework because portfolio risk can be represented in terms of the weights in each asset and their covariance matrix. Mathematically speaking, if $X \sim E_d(\mu, \Sigma, \psi)$ is a d-dimensional random vector, and $Y = AX + b \in R^k$ is a linear transformation of X where $A \in R^{k\times d}$ and $b \in R^k$, then we will have $Y \sim E_k(A\mu + b, A\Sigma A^T, \psi)$. For elliptically distributed returns X with dispersion matrix Σ, the dispersion matrix for a portfolio with weights $w = (w_1, ..., w_d)^T$ is then represented by $w^T\Sigma w$.

6.3.1 Comparison of various covariance estimates for elliptical distributions

In this subsection, we compare the four covariance estimators: sample covariance, Kendall's tau transformation, EM algorithm, and Tyler's method. As an example, we choose the most typical elliptical distribution — Student t distribution in our simulations. We focus on comparing the correlation estimation by converting the covariances to correlations. For our experiment, we focus on the cases where $p < n$ such that the MLE and Tyler's estimate both exist. For the Kendall's tau transform method, the obtained correlation may not be positive definite, so we adjust it by setting the non-positive eigenvalues to be a small number ε (here we set $\varepsilon = 10^{-6}$). Suppose the raw correlation transformed from Kendall's tau is P, if $P = VDV^T$, then we let $\widetilde{D}$ be a diagonal matrix with $\widetilde{D}_{ii} = D_{ii}1_{\{D_{ii}>0\}} + \varepsilon 1_{\{D_{ii}\leq 0\}}$, and adjust the estimate to be $\widetilde{P} = V\widetilde{D}V^T$.

In the following simulation, we compare the correlation estimates obtained by the above four methods with the prespecified true correlation under four types of distances – MSE, KL, QL, and Hellinger distance. The settings for the experiment are as follows: dimension $p = 30$, sample size ranging n from 40 to 200 with a step size 10, true correlation is autocovariance of AR(1) process with $\rho = 0.9$, and the number of trials M at each sample size is 1000. The estimation errors of the four types of covariance estimates against the true covariance under the four types of distance measures are shown in Figure 6.6. Note that the KL and Hellinger distance are derived from Gaussian distributions and we use them here as two derived loss functions between two covariance matrices rather than the true KL and Hellinger distance for Student t distributions.

We see that the EM and Tyler's method lead to very close estimates and have the least errors under any of the four distances. The Kendall's tau method leads to a smaller MSE and QL than the sample covariance, suggesting it provides a more robust estimate than sample covariance under these two distances. However, due to the non-positive definiteness of the correlation transformed from Kendall's tau, we have to adjust the non-positive eigenvalues to be a small positive value. Therefore, the errors of the covariance estimates by Kendall's tau transform method evaluated under the KL and Hellinger distance are very large. To reduce the error under KL and Hellinger distance as well requires a better adjustment method, which might be mainly related to shrinkage methods to make the eigenvalues less dispersed. We do not discuss those issues here. The most interesting fact is that Tyler's method provides very robust estimates in heavy-tailed elliptical distributions and the result is very close to the EM method. However, the EM method is often more complex to implement, and furthermore, it may not even be available if there is no explicit analytical form for the density function.

Given that the EM and Tyler's method provide more robust covariance estimates, when

we use these estimates as inputs for the smooth and monotone regularization, we should obtain more robust smooth monotone covariance estimates as well. In the following, we use the four correlation – sample covariance, Kendall's tau, EM, and Tyler's estimate — as inputs to get the smooth monotone covariance for each. The settings of the experiment are the same as above, and the results are shown in Figure 6.7.

Comparing the estimation errors of the smooth monotone covariance with the errors of raw estimates, we can see that the smooth monotone covariance greatly reduces the error for any of the four distances. Similar to the raw estimates, with smooth monotone regularization the EM and Tyler's method also provide the closest estimates to the true covariance. Note that the EM and Tyler's method cannot be applied when the covariance is not invertible if $p < n$. In that case, one may use the Kendall's tau transformation method or modify the initial covariance and the iteration steps of the EM and Tyler's method such that the obtained covariance in each iteration is positive definite.

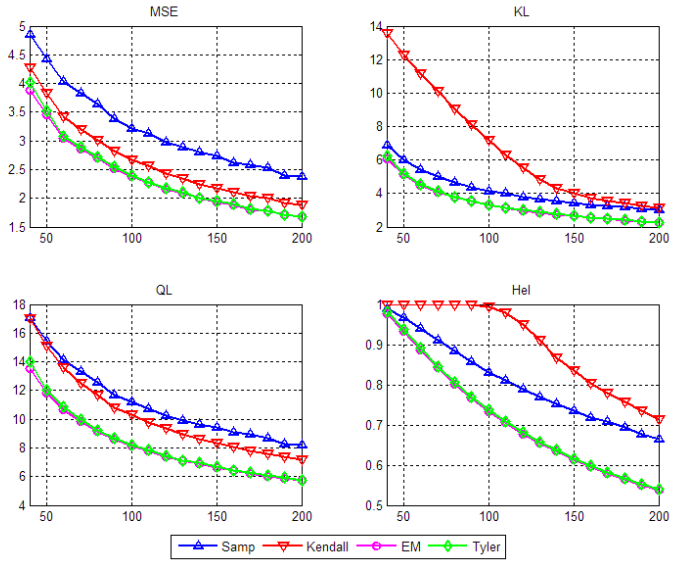


Figure 6.6: Errors of various covariance estimates for $t(5)$ distribution

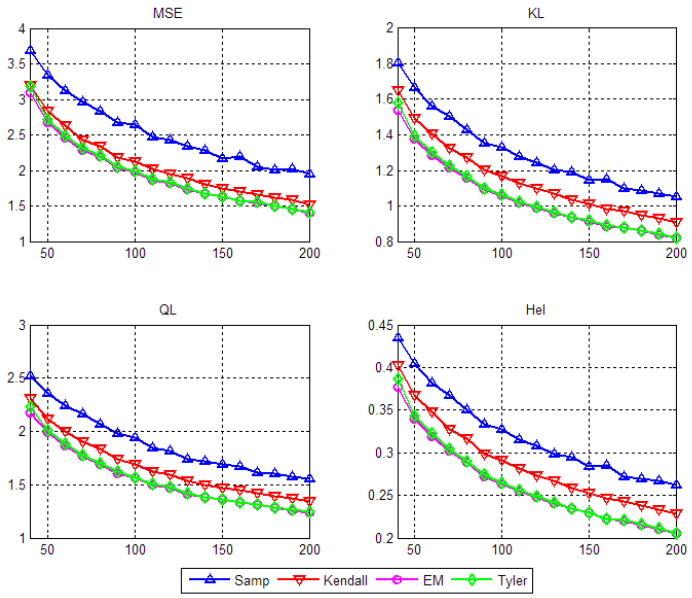


Figure 6.7: Errors of various smooth monotone covariance estimates for $t(5)$ distribution

6.3.2 Smooth monotone covariance estimates of portfolio risk for elliptical distributions

We repeat the experiment of portfolio optimization for elliptical distributions by calculating the estimation error of portfolio risk for Markowitz mean-variance portfolios. Because the results for mean-CVaR portfolios are very similar, we do not report here.

The settings of the experiment are exactly the same as for the Gaussian distributions, except the distribution here is changed to the multivariate t distribution with 5 degrees of freedom. Figure 6.8 shows the estimation error for the optimal portfolio risk using mean-variance optimization where the distribution is $t(5)$, sample size n is 50, dimension p is 30, and the covariance is AR(1) covariance with $\rho = 0.7$. We see that for Student t distribution, the sample covariance underestimates more risk than for Gaussian distributions, and smooth monotone covariance has much smaller error.

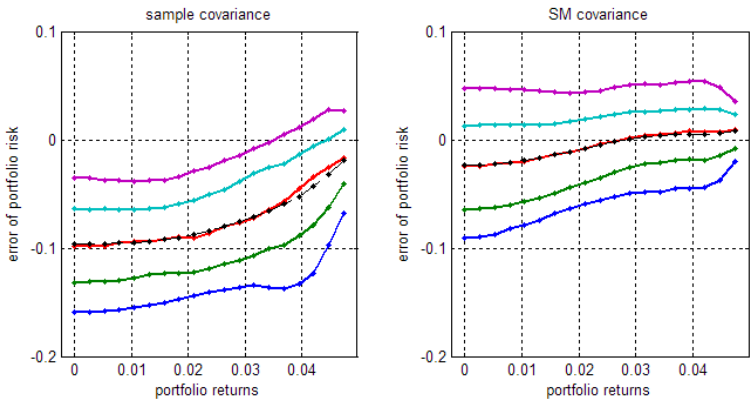


Figure 6.8: Estimation errors of optimal portfolio risk for $t(5)$ distribution using sample and smooth monotone covariance

In our simulation study for elliptical distributions, we used the multivariate t distribution as our examples. However, the Kendall's tau transformation method and Tyler's method for covariance estimation apply to all the elliptical distributions such as the normal inverse Gaussian, generalized hyperbolic distributions, and normal tempered stable distributions. Therefore, smooth monotone covariance using various inputs for these distributions can be obtained easily as well.

6.4 Empirical application

In this section, we are trying to apply the smooth and monotone covariance estimation to real problems in the finance area. In practice, the entities or instruments that we considered might be indexed by some natural order. For example, in yield curve modeling, the yield of different bonds can be ordered by their maturities such as 1M, 3M, 6M, 1Y, 2Y, 5Y, 20Y etc. When studying joint movement over time of the yields with different maturities, an accurate estimate for their covariance plays a central role. A monotone assumption is natural such that the covariance between adjacent yields are greater than faraway yields, and to prevent the stairs effect a smoothness penalty is applied.

6.4.1 Eurodollar futures portfolio optimization

To make the empirical study transparent and easily replicable, we use the most liquid bond futures: the Eurodollar (ED) futures. The ED futures are frequently used for hedging interest rate risk or constructing bond portfolio similarly as equities. A Eurodollar future is similar to a forward rate agreement to borrow or lend US$1,000,000 for three months starting on the contract settlement date. For historical reasons ED contracts are priced as $100 - x$ where x is the forward rate. A combination of ED rates with different expiries at a specific date reflects a snapshot of the forward curve. The covariance of yield changes are often well approximated by sample covariance or PCA method due to the well-known dominant three factors (level, slope, and curvature). However, if a portfolio has significant exposure to these futures, then the first principal component will be so dominant that it is hard to measure the relative merit of different covariance models. Also, in practice, the exposure to level-shift is usually tightly controlled so the portfolios can be viewed as mostly spread portfolios. Therefore, we focus on the covariance of spread changes rather than the yield changes: if $y_i(t)$ is the price of the i-th ED contract at time t, the i-th spread is $x_i(t) = y_i(t) - y_{i+1}(t)$. We use the spreads as our synthetic instruments, for which better covariance modelling can have dramatic improvements over simple models.

The data we used for our study are the historical monthly last prices of 32 generic ED tickers namely the 1-st contract, 2nd contract, etc. The data are downloaded from Bloomberg terminal and the generic tickers are made by rolling each contract to the next contract using a rolling method in Bloomberg called "with active futures". The period we choose is from 1995/01 to 2012/12 with 216 months in total and there are 32 tickers with full data in this period. Figure 6.9 shows the price data for the 32 tickers and the 31 spread monthly changes data.

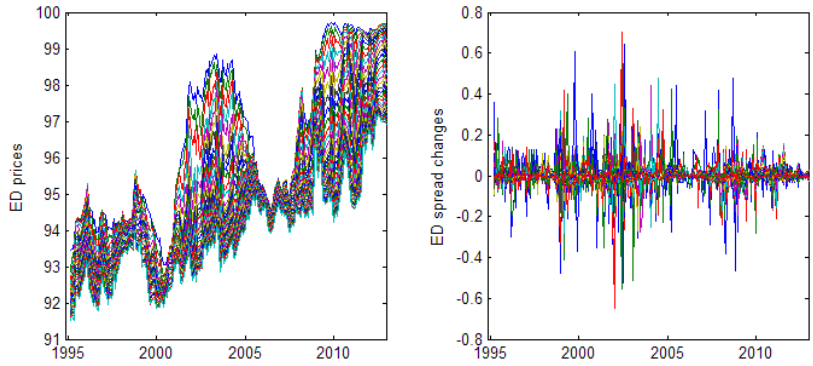


Figure 6.9: The daily close prices and spread changes of Eurodollar futures

Comparison of portfolio selection using sample and smooth monotone covariance We compare the performance of portfolio selection based on the spreads using sample covariance and smooth monotone covariance. The dimension of portfolio is 31, and the sample size we use for covariance estimate is 40. To focus on the performance comparison of the two covariances, we specify the variances are the same as the sample variances, and plug in the sample correlation and smooth monotone correlation to get the two covariances. Note that better modelling of the variances can be done by GARCH models, but it's outside the scope of this paper.

We use following mean-variance portfolio selection without considering the transaction costs:

$$\max_{w} \; w^T\mu - \frac{1}{2}\lambda w^T\Sigma w$$

The explicit solution of this portfolio is:

$$w = \frac{1}{\lambda}\Sigma^{-1}\mu.$$

In our simulation, we use constant total capital at the beginning of each period, so we normalize the weights:

$$w^* = \frac{w}{|w|^T 1}.$$

We conduct the single period portfolio optimization according to the above selection algorithm using sample correlation and smooth monotone correlation respectively from the 101-th month June 2003 to December 2012. We will examine the portfolios' performance when the covariances are calculated based on different sample sizes ranging from 30 to 100. To facilitate analyzing the impact of different sample sizes on the covariance estimation, we should fix the forecasting of expected returns so that it does not interfere with our judgement about covariance. Therefore, we first forecast the expected returns using an enlarging window approach and will use them for all the cases. Note that one can also select a window size and use a moving window approach for expected returns model, but that is beyond the focus of our discussion: covariance modelling.

For the expected returns model, we apply a p-th order forward linear predictor which predicts the current values based on past samples:

$$\widehat{r}(t) = -a(2)r(t-1) - a(3)r(t-2) - ... - a(p+1)r(t-p)$$

where the coefficients $a = (1, a(2), ..., a(p+1))$ are obtained by Levison-Durbin recursion ("lpc" function in Matlab Signal Processing Toolbox). The reason why this model is used is to make the expected return prediction model simple enough so that one can replicate the results easily, and also we find it provide satisfactory prediction of the returns and lead to positive portfolio returns in our case. Here we choose $p = 5$, and calibrate the coefficients at each period using all available historical data.

The weights of the portfolio are rebalanced every month, and we calculate the realized gains of the two portfolios every month. Figure 6.10 shows the cumulative gains of the two portfolios using sample covariance and smooth monotone covariance with sample size n equal to 40. By using the almost 10 years back-testing performance, we can see that the smooth monotone covariance improves the gross gains significantly, and also improves the Sharpe ratio. Here, the Sharpe ratio is normalized to an annual base by multiplying $\sqrt{12}$.

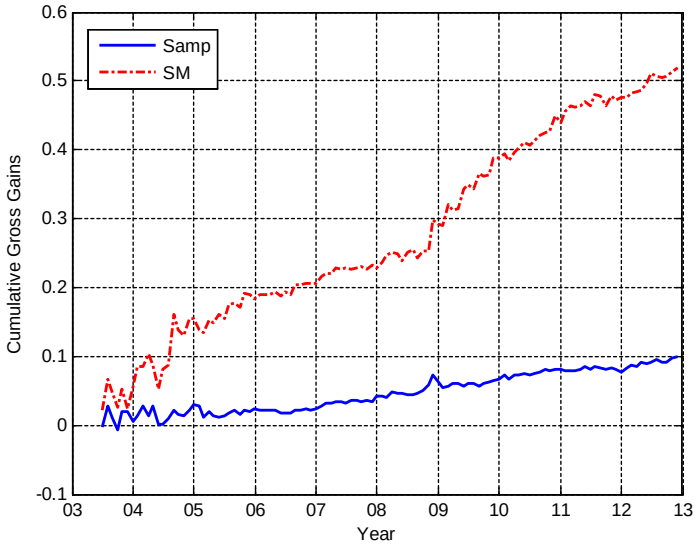


Figure 6.10: Realized portfolio return using sample and smooth monotone covariance

Sensitivity of portfolio performance to the length of sample data To investigate the performance of smooth monotone covariance and sample covariance to varying sample size, we conduct the above back-testing for sample size from 30 to 100 with a step size 5, and plot the cumulative returns and Sharpe ratio at each sample size for the two covariances in Figure 6.11.

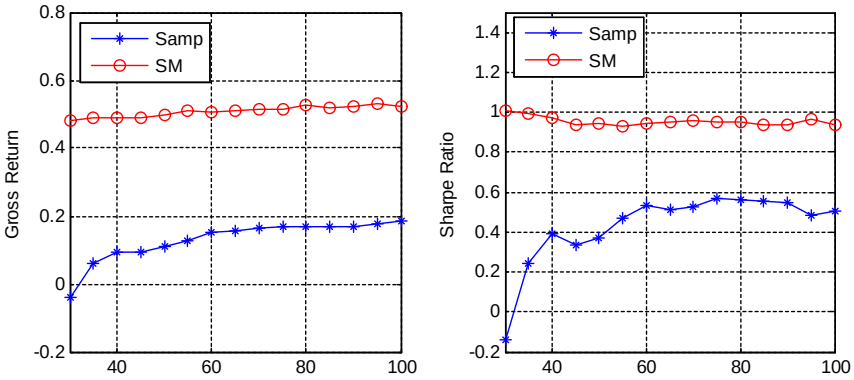


Figure 6.11: Sensitivity of portfolio return and sharpe ratio with respect to sample size

From Figure 6.11, we see that the portfolio returns and sharpe ratio is more sensitive to the change of sample size when using the sample covariance. Overall, we see that the smooth monotone covariance outperforms the sample covariance greatly at either small or large sample sizes in both portfolio returns and Sharpe ratio.

6.4.2 Corporate bonds application

Block Smooth monotone covariance In this subsection, we provide another example of application of smooth monotone covariance for fixed income securities - corporate bonds. Corporate bonds of different companies usually have different coupon rates, different maturities and some might have callable options. To make our example simpler, we use the corporate bond indices composed by Bloomberg. Specifically, we download the daily price of United States Composite BVal AA, A, and BBB Indices (tickers "IGUUDC, IGUUAC, IGUUBC"), which are populated with USD denominated fixed-rate bonds issued by domestic companies in United States. For each rating of these curves, the maturities include 1y, 2y, 3y, 5y, 7y, 10y, 15y, 20y, and 25y. So, in total, there are 27 indices. Given that for each rating the correlation of the yields exhibit a smooth and monotone effect, the correlation of all indices together is then a block matrix with smooth and monotone effect on the diagonal blocks if we arrange the indices of the same ratings together. For the non-diagonal components, we can also add a smoothness penalty.

Suppose the diagonal blocks are $P_{11}, P_{22}, ...P_{BB}$, the total covariance is

$$P = \begin{bmatrix} P_{11} & P_{12} & ... & P_{1B} \\ P_{12}^T & P_{22} & ... & P_{2B} \\ ... & ... & ... & ... \\ P_{1B}^T & P_{2B}^T & ... & P_{BB} \end{bmatrix}$$

We impose monotone constraints on the diagonal blocks P_{ii}, and smoothness penalty on both the diagonal and off-diagonal blocks:

$$\min_P D(P,Q) + \lambda_1 \left\| D_2^{(ii)} vec(P_{ii}) \right\|_2 + \lambda_2 \left\| D_2^{(ij)} vec(P_{ij}) \right\|_2,$$

$$\text{such that } P \succeq 0,\ D^{(i)} vec(P_{ii}) \geq 0.$$

In Figure 6.12, the left two plots show the sample correlation of yield returns ($r_{i,t} = \log \frac{y_{i,t}}{y_{i,t-1}}$) and the spread changes $x_{it} - x_{i,t-1}$, where $x_{it} = y_{it} - y_{i-1,t}$. Similarly as the Eurodollar futures data, both of them exhibit a smooth and monotone effect but the first principle component of the correlation of yield returns might be very dominant so that the outperformance of smooth monotone correlation over sample correlation may be less significant. Therefore, we estimate the block smooth monotone correlation for both the yield returns and the spread changes. The smooth monotone covariance for both of them are shown in the right two plots. It can be found that, by using the monotone and smooth regularization, the noise effect in the covariance has been reduced.

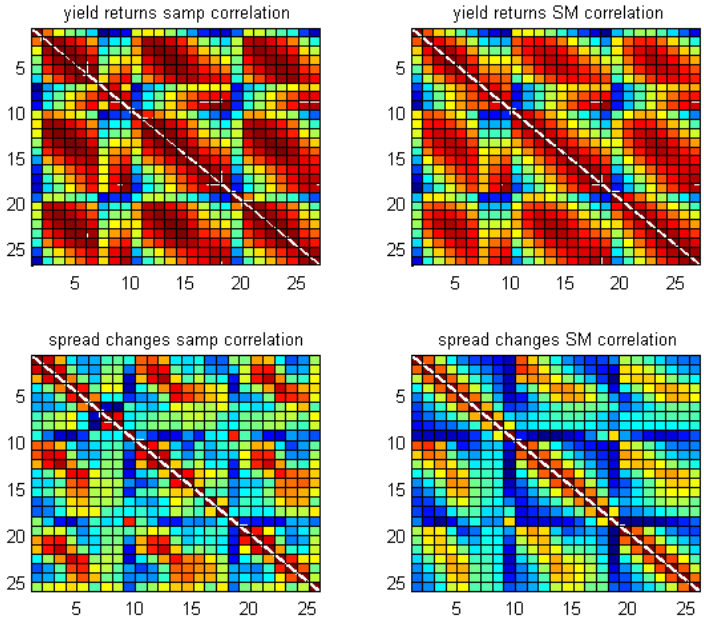


Figure 6.12: Sample and SM correlation of yield returns and spread changes

Correlation Forecasting To compare the sample covariance and smooth monotone covariance, we conduct a study of forecasting future correlation coefficient matrices over several years of corporate bond indices data. The approach is as follows: use previous N days data as a training window, calculate the sample correlation and also the block smooth monotone correlations. We compare the correlation with the realized future correlations, which we use 50 days out-of-sample sample correlation. We have the 770 daily yields of the 27 corporate bond indices from 01/01/2010 to 12/31/2012, and calculate the log-returns of the yields and also the spread changes. For both the two data sets, we estimate the block smooth monotone correlations. To choose the smoothness penalty parameter λ, we use an approach similar to cross validation.

Suppose we have a p dimensional sample data $X_{p\times M}$, and we divide it into m groups so that in each group there are $n = M/m$ data points and denote the i-th group data as $X^{(i)}$. For each group data $X^{(i)}$ $(i = 1, 2, ..., m)$, estimate the smooth monotone correlation using different λ, and calculate the Frobenius norm error against the sample correlation using the

rest data $X\backslash X^{(i)}$. For each group i, find out the optimal λ_i which leads to the smallest error. Then, calculate the average optimal λ over all the m groups: $\lambda = \frac{1}{m}\sum_{i=1}^{m}\lambda_i$.

In the following, we use the first 300 data points as our in-sample data, and the rest 469 data points for out-of-sample testing. We divide the 300 data points into 10 groups each with 30 days data, and use the above cross validation technique to choose the optimal λ, and find the approximate optimal λ for the yield returns correlation and spread changes correlation to be 0.6 and 0.2 respectively. For simplicity, we let the smoothness penalty on the diagonal blocks and off-diagonal blocks be the same.

Figure 6.13 shows the realized Frobenius norm error of sample correlation and block smooth monotone correlation with respect to varying training window size from 10 to 50. We use running windows with shifts by 5 business days over the 469 daily data for both the yield returns and the spread changes. It shows that the block smooth monotone correlation has better out-of-sample forecasting performance than the sample correlation. For the spread changes correlation, the reduced error is more significant. The smooth and monotone regularization appears especially valuable for a small training window size, demonstrating robustness in forecasting risk in scenarios with severely limited data.

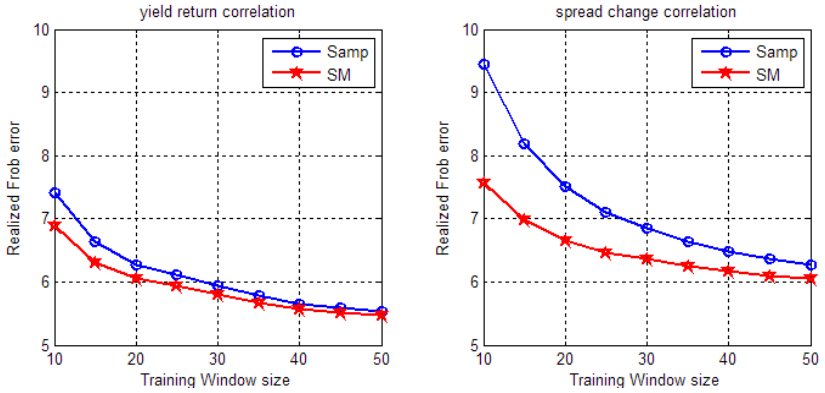


Figure 6.13: The errors of out-of-sample forecast of correlation using sample and SM covariance

6.5 Conclusion

In this paper, we study high-dimensional covariance estimation with smooth and monotone regularization, which is useful when the instruments have an ordering. Using a series of simulation studies, we first analyze its performance under Gaussian distributions by using various distances to measure the error against true covariance. It shows that the smooth monotone covariance works uniformly better than sample covariance under not only Frobenius norm error, but also KL, QL and Hellinger distances. We also find its improvement in the optimal portfolio risk measurement by comparing the Markowitz efficient frontiers. We also study the generalization of the smooth monotone covariance by using alternative distances as objective functions such as 2-norm, KL and QL. However, we find it is not advisable to minimizing the KL or QL although they are good for evaluating covariance estimates.

To fit with heavy-tailed distributions which are more common in finance, we study the smooth monotone covariance for elliptical distributions using various inputs such as

sample covariance, Kendall's tau transform method, ML estimates (EM algorithm), Tyler's estimate. It shows that, by using alternative robust covariance inputs, the performance of smooth monotone covariance can be improved. Finally, we provide two real application examples: one is the Eurodollar futures portfolio optimization, in which we show the smooth monotone covariance outperforms the sample covariance by generating a larger return and Sharpe ratio; the other one is corporate bond portfolio, in which we extend the smooth monotone covariance into block smooth monotone covariance, and demonstrate that it is more robust than the sample covariance for out-of-sample covariance prediction. More applications can also be found in credit portfolio risk, volatility surface modeling, and time series autocovariance where the smooth and monotone assumptions are valid.

References

[1] Anderson, T. W., & Darling, D. A. (1952). Asymptotic theory of certain" goodness of fit" criteria based on stochastic processes. The Annals of Mathematical Statistics, 23(2), 193-212.

[2] Anderson, T. W., & Darling, D. A. (1954). A test of goodness of fit. Journal of the American Statistical Association, 49(268), 765-769.

[3] Atkinson, C., & Mitchell, A. F. (1981). Rao's distance measure. Sankhyā: The Indian Journal of Statistics, Series A, 345-365.

[4] Aue, F., & Kalkbrener, M. (2006). LDA at work: Deutsche Bank's approach to quantifying operational risk. Journal of Operational Risk, 1(4), 49-93.

[5] Bai, J. and S. Ng, 2010. Instrumental variable estimation in a data rich environment. Econometric Theory 26, 1577-1606.

[6] Bai, J., & Shi, S. (2011). Estimating High Dimensional Covariance Matrices and its Applications. Annals of Economics and Finance, 12(2), 199-215.

[7] Baksh, M. F., Bohning, D., & Lerdsuwansri, R. (2011). An extension of an over-dispersion test for count data. Computational Statistics & Data Analysis, 55(1), 466-474.

[8] Barndorff-Nielsen, O. (1978). Hyperbolic distributions and distributions on hyperbolae. Scandinavian journal of statistics, 151-157.

[9] Basel, I. (2001). Operational Risk — Consultative Document, Supporting document to the New Basel Capital Accord. Basel Committee on Banking Supervision, Bank for International Settlements, Basel.

[10] Basel, I. (2004). Revised international capital framework. Basel, Switzerland: Basel Committee on Banking Supervision.

[11] Basel, I. (2006). International Convergence of Capital Measurement and Capital Standards, A Revised Framework. Comprehensive Version, Basel Committee on Banking Supervision, Bank for International Settlements, Basel, June 2006.

[12] Basel, I. (2011). Operational Risk - Supervisory Guidelines for the Advanced Measurement Approaches - final document. Basel Committee on Banking Supervision, Bank for International Settlements, Basel, June 2011.

[13] Baud, N., Frachot, A., Roncalli, T. (2002). Internal data, external data and consortium data for operational risk measurement: How to pool data properly? Groupe de Recherche Operationnelle, Credit Lyonnais, France, 1-18.

[14] Bee, M. (2005). On maximum likelihood estimation of operational loss distributions. University of Trento Department of Economics Working Paper No. 2005-03.

[15] Bengtsson, C., & Holst, J. (2002). On portfolio selection: Improved covariance matrix estimation for Swedish asset returns. Paper presented at the 31st Meeting, Euro Working Group on Financial Modeling.

[16] Berger, J. O., & Bernardo, J. M. (1989). Estimating a product of means: Bayesian analysis with reference priors. Journal of the American Statistical Association, 84(405), 200-207.

[17] Berger, J. O., & Bernardo, J. M. (1992). On the development of reference priors. Bayesian statistics, 4(4), 35-60.

[18] Berger, J. O., Bernardo, J. M., & Sun, D. (2009). The formal definition of reference priors. The Annals of Statistics, 905-938.

[19] Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference. Journal of the Royal Statistical Society. Series B (Methodological), 113-147.

[20] Best, D., Rayner, J., & Thas, O. (2007). Goodness of fit for the zero-truncated Poisson distribution. Journal of Statistical Computation and Simulation, 77(7), 585-591.

[21] Bickel, P. J., & Levina, E. (2008a). Covariance regularization by thresholding. The Annals of Statistics, 36(6), 2577-2604.

[22] Bickel, P. J., & Levina, E. (2008b). Regularized estimation of large covariance matrices. The Annals of Statistics, 36(1), 199-227.

[23] Birnbaum, Z. W. (1952). Numerical tabulation of the distribution of Kolmogorov's statistic for finite sample size. Journal of the American Statistical Association, 47(259), 425-441.

[24] Bocker, K., & Kluppelberg, C. (2005). Operational VaR: a closed-form approximation. RISK-LONDON-RISK MAGAZINE LIMITED-, 18(12), 90.

[25] Bocker, K., & Kluppelberg, C. (2008). Modeling and measuring multivariate operational risk with Levy copulas. The Journal of Operational Risk 3(2): 3-27.

[26] Bocker, K., & Sprittulla, J. (2006). Operational VAR: meaningful means. Risk Magazine, 12, 96-98.

[27] Bohning, D. (1994). A Note on a Test for Poisson Overdispersion. Biometrika, 418-419.

[28] Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics, 31(3), 307-327.

[29] Breiman, L. (1995). Better subset regression using the nonnegative garrote. Technometrics, 37(4), 373-384.

[30] Breymann, W., Luthi, D.(2008). ghyp: A package on generalized hyperbolic distributions. Tech. rep. Institute of Data Analysis and Process Design. http://cran.r-project.org/

[31] Cambanis, S., Huang, S., & Simons, G. (1981). On the theory of elliptically contoured distributions. Journal of Multivariate Analysis, 11(3), 368-385.

[32] Carlin, B. P., & Louis, T. A. (2000). Empirical Bayes: Past, present and future. Journal of the American Statistical Association, 95(452), 1286-1289.

[33] Casella, G. (1992). Illustrating empirical Bayes methods. Chemometrics and intelligent laboratory systems, 16(2), 107-125.

[34] Cavallo, A., Rosenthal, B., Wang, X., & Yan, J. (2012). Treatment of the data collection threshold in operational risk: a case study using the lognormal distribution. The Journal of Operational Risk, 7(1), 3-38.

[35] Chen, N.-F., Roll, R., & Ross, S. A. (1986). Economic forces and the stock market. Journal of business, 383-403.

[36] Chen, Y. (2011). Regularized Estimation of High-dimensional Covariance Matrices. Dissertation. University of Michigan.

[37] Chen, Y., Wiesel, A., & Hero, A. O. (2009). Shrinkage estimation of high dimensional covariance matrices. Paper presented at the Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on.

[38] Chen, Y., Wiesel, A., & Hero, A. O. (2011). Robust shrinkage estimation of high-dimensional covariance matrices. Signal Processing, IEEE Transactions on, 59(9), 4097-4107.

[39] Chernobai, A. and Rachev, S.T. (2004). Stable Modelling of Operational Risk, in M. G. Cruz, ed., "Operational Risk Modelling and Analysis. Theory and Practice," RISK Book,s London, pp. 139-169.

[40] Chernobai, A., Menn, C., Rachev, S. T., Truck, C., & Moscadelli, M. (2006). Treatment of incomplete data in the field of operational risk: The effects on parameter estimates, EL and UL figures. The Advanced Measurement Approach to Operational Risk, 145-468.

[41] Chernobai, A., Menn, C., Rachev, S., Truck, S. (2005b). Estimation of operational value-at-risk in the presence of minimum collection thresholds. Tech. rep., University of California, Santa Barbara.

[42] Chernobai, A., Menn, C., Truck, S., Rachev, S. (2005a). A note on the estimation of the frequency and severity distribution of operational losses. Mathematical Scientist, 30(2).

[43] Chernobai, A., Rachev, S., & Fabozzi, F. (2005). Composite goodness-of-fit tests for left-truncated loss samples. Department of Statistics and Applied Probability, University of California Santa Barbara.

[44] Chernobai, A., Rachev, S., Fabozzi, F. (2007). Operational risk: a guide to Basel II capital requirements, models, and analysis. John Wiley & Sons Inc.

[45] Clarke, B., & Ghosal, S. (2010). Reference priors for exponential families with increasing dimension. Electronic Journal of Statistics, 4, 737-780.

[46] Conlon, T., Ruskin, H. J., & Crane, M. (2007). Random matrix theory and fund of funds portfolio optimisation. Physica A: Statistical Mechanics and its applications, 382(2), 565-576.

[47] Cope, E. W. (2011). Penalized likelihood estimators for truncated data. Journal of Statistical Planning and Inference, 141(1), 345-358.

[48] Cope, E. W. (2012). Combining scenario analysis with loss data in operational risk quantification. The Journal of Operational Risk, 7(1), 39-56.

[49] Cope, E. W., Mignola, G., Antonini, G., & Ugoccioni, R. (2009). Challenges and pitfalls in measuring operational risk from loss data. Journal of Operational Risk, 4(4), 10.

[50] Cope, E., Labbi, A. (2008). Operational loss scaling by exposure indicators: Evidence from the ORX database. The Journal of Operational Risk, 3(4), 25-46.

[51] Cruz, M. (2002). Modeling, measuring and hedging operational risk. John Wiley & Sons New York.

[52] Dahen, H., Dionne, G. (2007). Scaling models for the severity and frequency of external operational loss data. Available at SSRN: http://ssrn.com/abstract=958759

[53] Dalla Valle, L., Giudici, P. (2008). A Bayesian approach to estimate the marginal loss distributions in operational risk management. Computational Statistics & Data Analysis, 52(6), 3107-3127.

[54] Daniels, M. J., & Kass, R. E. (1999). Nonconjugate Bayesian estimation of covariance matrices and its use in hierarchical models. Journal of the American Statistical Association, 94(448), 1254-1263.

[55] Daniels, M. J., & Kass, R. E. (2001). Shrinkage estimators for covariance matrices. Biometrics, 57(4), 1173-1184.

[56] d'Aspremont, A., Banerjee, O., & El Ghaoui, L. (2008). First-order methods for sparse covariance selection. SIAM Journal on Matrix Analysis and Applications, 30(1), 56-66.

[57] Datta, G., Ghosh, M. (1996). On the invariance of noninformative priors. The Annals of Statistics, 24(1), 141-159.

[58] De Fontnouvelle, P., Rosengren, E., Jordan, J. (2007). Implications of alternative operational risk modeling techniques. University of Chicago Press.

[59] Degen, M., Embrechts, P., & Lambrigger, D. D. (2007). The quantitative modeling of operational risk: between g-and-h and EVT. Astin Bulletin, 37(2), 265.

[60] Dempster, A., Laird, N., Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological), 1-38.

[61] Doob, J. L. (1949). Heuristic approach to the Kolmogorov-Smirnov theorems. The Annals of Mathematical Statistics, 20(3), 393-403.

[62] Dutta, K. K., & Babbel, D. F. (2013). Scenario analysis in the measurement of operational risk capital: a change of measure approach. Journal of Risk and Insurance.

[63] Dutta, K., Perry, J. (2006). A tale of tails: An empirical analysis of loss distribution models for estimating operational risk capital. Federal Reserve Bank of Boston, 06–13.

[64] Efron, B. & Morris, C. (1977). Stein's paradox in statistics. Scientific American, 237:119-127

[65] Efron, B., & Morris, C. (1976). Multivariate empirical Bayes and estimation of covariance matrices. The Annals of Statistics, 22-32.

[66] Eggermont, P. P. B., & LaRiccia, V. N. (2001). Maximum penalized likelihood estimation (Vol. 2): Springer.

[67] Elton, E. J., & Gruber, M. J. (1973). Estimating the Dependence Structure of Share Prices-Implications for Portfolio Selection. The Journal of Finance, 28(5), 1203-1232.

[68] Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. Econometrica: Journal of the Econometric Society, 987-1007.

[69] Fabozzi, F. J., Kolm, P. N., Pachamanova, D., & Focardi, S. M. (2007). Robust portfolio optimization and management: Wiley.

[70] Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. Journal of Financial Economics, 33(1), 3-56.

[71] Fan, J., Fan, Y., & Lv, J. (2008). High dimensional covariance matrix estimation using a factor model. Journal of Econometrics, 147(1), 186-197.

[72] Feller, W. (1948). On the Kolmogorov-Smirnov limit theorems for empirical distributions. The Annals of Mathematical Statistics, 19(2), 177-189.

[73] Firth, D. (1993). Bias reduction of maximum likelihood estimates. Biometrika, 80(1), 27.

[74] Frachot, A., Georges, P., Roncalli, T. (2001). Loss distribution approach for operational risk. Manuscript. Groupe de Recherche Operationnelle, Credit Lyonnais, France. April. http://gro.creditlyonnais.fr/content/wp/lda.pdf.

[75] Frachot, A., Roncalli, T., & Salomon, E. (2004). The correlation problem in operational risk. OperationalRisk Risk's Newsletter.

[76] Frahm, G. (2004). Generalized elliptical distributions: theory and applications. Dissertation. Universitat zu Koln.

[77] Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. Biostatistics, 9(3), 432-441.

[78] Furrer, R., & Bengtsson, T. (2007). Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants. Journal of Multivariate Analysis, 98(2), 227-255.

[79] Giacometti, R., Rachev S., Chernobai A., Bertocchi M. (2008). Aggregation issues in operational risk. The Journal of Operational Risk, 3(3), 3-23.

[80] Goodd, I., & Gaskins, R. (1971). Nonparametric roughness penalties for probability densities. Biometrika, 58(2), 255-277.

[81] Grant, M. C., & Boyd, S. P. (2012). The CVX Users' Guide. CVX Research, Inc.

[82] Greene, W. H. (1994). Accounting for Excess Zeros and Sample Selection in Poisson and Negative Binomial Regression Models: New York University, Leonard N. Stern School of Business, Department of Economics.

[83] Haff, L. (1980). Empirical Bayes estimation of the multivariate normal covariance matrix. The Annals of Statistics, 8(3), 586-597.

[84] Haff, L. (1982). Solutions of the Euler-Lagrange equations for certain multivariate normal estimation problems. Unpublished manuscript.

[85] Hampel, F.R. (1968). Contributions to the theory of robust estimation. Doctoral Thesis, University of California, Berkeley.

[86] Hero (2012). unified framework for regularized covariance estimation in scaled Gaussian models.

[87] Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. The Annals of Statistics, 1163-1174.

[88] Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1), 55-67.

[89] Hsieh, C.-J., Sustik, M. A., Dhillon, I. S., & Ravikumar, P. (2011). Sparse inverse covariance matrix estimation using quadratic approximation. Advances in Neural Information Processing Systems, 24, 2330-2338.

[90] Hu, W. (2005). Calibration Of Multivariate Generalized Hyperbolic Distributions Using The EM Algorithm, With Applications In Risk Management, Portfolio Optimization And Portfolio Credit Risk.

[91] Huang, J. Z., Liu, N., Pourahmadi, M., & Liu, L. (2006). Covariance matrix selection and estimation via penalised normal likelihood. Biometrika, 93(1), 85-98.

[92] Huber, P.J. (1981). Robust Statistics. Wiley.

[93] Huber, S. (2010). (Non-) robustness of maximum likelihood estimators for operational risk severity distributions. Quantitative Finance, 10(8), 871-882.

[94] Jacquier E., Polson N. G., & Rossi P. E. (1994). Bayesian Analysis of Stochastic Volatility Models. Journal of Business & Economic Statistics. Vol 2, No. 4.

[95] James, W., & Stein, C. (1961). Estimation with quadratic loss. Paper presented at the Proceedings of the fourth Berkeley symposium on mathematical statistics and probability.

[96] Jeffreys, H. (1967). Theory of Probability. Oxford University Press, London

[97] Johnson, D. (2013). Private communication.

[98] Jones, C. S., 2001. Extracting Factors from Heteroskedastic Asset Returns. Journal of Financial Economics 62, 293-325.

[99] Joreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. Psychometrika, 34(2), 183-202.

[100] Karlis, D. (2002). An EM type algorithm for maximum likelihood estimation of the normal–inverse Gaussian distribution. Statistics & Probability Letters, 57(1), 43-52.

[101] Karoui, N. E. (2008a). Spectrum estimation for large dimensional covariance matrices using random matrix theory. The Annals of Statistics, 36(6), 2757-2790.

[102] Karoui, N. E. (2008b). Operator norm consistent estimation of large-dimensional sparse covariance matrices. The Annals of Statistics, 2717-2756.

[103] Kelker, D. (1970). Distribution theory of spherical distributions and a location-scale parameter generalization. Sankhyā: The Indian Journal of Statistics, Series A, 419-430.

[104] Kim, Y. S., Giacometti, R., Rachev, S. T., Fabozzi, F. J., & Mignacca, D. (2012). Measuring financial risk and portfolio optimization with a non-Gaussian multivariate model. Annals of Operations Research, 201(1), 325-343.

[105] Kolmogorov A (1933). Sulla determinazione empirica di una legge di distribuzione. G. Inst. Ital. Attuari 4: 83.

[106] Komaki, F. (2006). Shrinkage priors for Bayesian prediction. The Annals of Statistics, 34(2), 808-819.

[107] Kotz, S., & Nadarajah, S. (2004). Multivariate t-Distributions and their Applications: Cambridge University Press.

[108] Kuiper, N. H. (1960). Tests concerning random points on a circle. Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen, Series A 63: 38–47.

[109] Kwapień, J., & Drożdż, S. (2006). The bulk of the stock market correlation matrix is not pure noise. Physica A: Statistical Mechanics and its applications, 359, 589-606.

[110] Laloux, L., Cizeau, P., Bouchaud, J. P., & Potters, M. (1999). Noise dressing of financial correlation matrices. Physical Review Letters, 83(7), 1467-1470.

[111] Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics, 34(1), 1-14.

[112] Lawley, D. N. (1940). The estimation of factor loadings by the method of maximum likelihood. Proceedings of the Royal Society of Edinburgh, 60(2), 64-82.

[113] Ledoit, O., & Wolf, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. Journal of Empirical Finance, 10(5), 603-621.

[114] Ledoit, O., & Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. Journal of Multivariate Analysis, 88(2), 365-411.

[115] Ledoit, O., & Wolf, M. (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. The Annals of Statistics, 40(2), 1024-1060.

[116] Lehmann, E. E. L., & Romano, J. P. (2005). Testing statistical hypotheses: Springer Science+ Business Media.

[117] Letac, G., & Massam, H. (2004). All invariant moments of the Wishart distribution. Scandinavian journal of statistics, 31(2), 295-318.

[118] Lindskog, F., Mcneil, A., & Schmock, U. (2003). Kendall's tau for elliptical distributions. Credit risk: Measurement, evaluation and management, 149-156.

[119] Liu, C., & Rubin, D. B. (1995). ML estimation of the t distribution using EM and its extensions, ECM and ECME. Statistica Sinica, 5(1), 19-39.

[120] Liu, C., & Rubin, D. B. (1998). Maximum likelihood estimation of factor analysis using the ECME algorithm with complete and incomplete data. Statistica Sinica, 8, 729-748.

[121] Malioutov, D. (2011). Smooth isotonic covariances. Paper presented at the Statistical Signal Processing Workshop (SSP), 2011 IEEE.

[122] Malioutov, D. , A. Corum, and M. Cetin. (2012). Smooth and monotone covariance regularization: fast first-order methods. Working paper.

[123] Marčenko, V. A., & Pastur, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. Sbornik: Mathematics, 1(4), 457-483.

[124] Marsaglia G, Tsang W., Wang J (2003). Evaluating Kolmogorov's Distribution. Journal of Statistical Software 8 (18): 1–4

[125] Marsaglia, G., & Marsaglia, J. (2004). Evaluating the anderson-darling distribution. Journal of Statistical Software, 9(2), 1-5.

[126] Mason, D. M. (1982). Laws of large numbers for sums of extreme values. The Annals of Probability, 754-764.

[127] Massey, F.J., (1951) The Kolmogorov-Smirnov Test for Goodness of Fit. Journal of the American Statistical Association, 46(253):68-78.

[128] McNeil, A. J., Frey, R., & Embrechts, P. (2005). Quantitative risk management: concepts, techniques, and tools: Princeton university press.

[129] Miller, L.H., (1956) Table of Percentage Points of Kolmogorov Statistics. Journal of the American Statistical Association, 51(273):111-121.

[130] Moscadelli, M. (2004). The modelling of operational risk: experience with the analysis of the data collected by the Basel Committee. Available at SSRN 557214.

[131] Muller, P. (1993). Empirical tests of biases in equity portfolio optimization: Cambridge University Press Cambridge, UK.

[132] Myung J., Navarro, D. (2004). Information matrix. Encyclopedia of Behavioral Statistics. Wiley. Ohio State University.

[133] Nguyen, T., & Samorodnitsky, G. (2011). Tail Inference: where does the tail begin? Extremes, 1-25.

[134] Opdyke, J., Cavallo, A. (2012). Estimating Operational Risk Capital: The Challenges of Truncation, the Hazards of MLE, and the Promise of Robust Statistics. Journal of Operational Risk Volume, 7(3), 3-90.

[135] Park, S., & O'Leary, D. P. (2010). Portfolio selection using tikhonov filtering to estimate the covariance matrix. SIAM Journal on Financial Mathematics, 1(1), 932-961.

[136] Peters, G., Sisson, S. (2006). Bayesian inference, Monte Carlo sampling and operational risk. The Journal of Operational Risk, 1(3), 27-50.

[137] Plerou, V., Gopikrishnan, P., Amaral, L. A. N., Meyer, M., & Stanley, H. E. (1999). Scaling of the distribution of price fluctuations of individual companies. Physical Review E, 60(6), 6519.

[138] Plerou, V., Gopikrishnan, P., Rosenow, B., Amaral, L. A. N., Guhr, T., & Stanley, H. E. (2002). Random matrix approach to cross correlations in financial data. Physical Review E, 65(6), 066126.

[139] Press (1982). Applied Multivariate analysis: Using Bayesian and Freuentist measures of inference. New York: Holt, Rinehart & Winston.

[140] Protassov, R. S. (2004). EM-based maximum likelihood parameter estimation for multivariate generalized hyperbolic distributions with fixed λ. Statistics and Computing, 14(1), 67-77.

[141] Rachev, S. T., Hsu, J. S., Bagasheva, B. S., & Fabozzi, F. J. (2008). Bayesian methods in finance (Vol. 153): Wiley.

[142] Rachev, S. T., Kim, Y. S., Bianchi, M. L., & Fabozzi, F. J. (2011). Financial models with Lévy processes and volatility clustering. Wiley.

[143] Rachev, S. T., Menn, C., & Fabozzi, F. J. (2005). Fat-tailed and skewed asset return distributions: Implications for risk management, portfolio selection, and option pricing. John Wiley & Sons: Hoboken, NJ.

[144] Rachev, S. T., & Mittnik, S. (2000). Stable paretian models in finance. John Wiley & Sons Inc.

[145] Rao, C. R., & Chakravarti, I. (1956). Some small sample tests of significance for a Poisson distribution. Biometrics, 12(3), 264-282.

[146] Rasmussen, C. E., & Williams, C. K. I. (2006). Gaussian processes for machine learning (Vol. 1): MIT press Cambridge, MA.

[147] Rencher, A. C., & Christensen, W. F. (2012). Methods of multivariate analysis (Vol. 709): Wiley.

[148] Roehr, A. (2002). Modelling operational losses. Algo research quarterly, 5(2), 53-64.

[149] Rothman, A. J., Bickel, P. J., Levina, E., & Zhu, J. (2008). Sparse permutation invariant covariance estimation. Electronic Journal of Statistics, 2, 494-515.

[150] Rousseeuw (1984). Least median of squares regression. Journal of the American Statistical Association. 79(388): 871-880.

[151] Rozenfeld, I. (2010). Using Shifted Distributions in Computing Operational Risk Capital. Available at SSRN 1596268.

[152] Rubin, D. B., & Thayer, D. T. (1982). EM algorithms for ML factor analysis. Psychometrika, 47(1), 69-76.

[153] Saxena, A., & Stubbs, R. A. (2010). Alpha alignment factor: A solution to the underestimation of risk for optimized active portfolios. Axioma, Inc., Research Report, 15.

[154] Samorodnitsky, G., & Taqqu, M. S. (1994). Stable non-Gaussian processes. Stochastic Models with Infinite Variance, Chapman & Hall, New York.

[155] Shevchenko, P. (2010). Implementing loss distribution approach for operational risk. Applied Stochastic Models in Business and Industry, 26(3), 277-307.

[156] Shevchenko, P. (2010). Modelling Operational Risk Using Bayesian Inference. Springer Verlag.

[157] Siegel (1982). Robust regression using repeated medians. Biometrika. 69(1):242-244.

[158] Silverstein, J. W. (1995). Strong convergence of the empirical distribution of eigenvalues of large dimensional random matrices. Journal of Multivariate Analysis, 55(2), 331-339.

[159] Smirnov NV (1948). Tables for estimating the goodness of fit of empirical distributions. Annals of Mathematical Statistics 19: 279.

[160] Smirnov, N.V. (1937). On the distribution of the ω^2-criterion of von Mises. Rec. Math. (NS), Vol. 2. pp. 973-993.

[161] Stefanski, L. A., & Boos, D. D. (2002). The calculus of M-estimation. The American Statistician, 56(1), 29-38.

[162] Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In Neyman, J., editor, Proceedings of the Third Berkeley Symposium on Mathematical and Statistical Probability, pages 197-206. University of California, Berkeley. Volume 1.

[163] Stein, C. (1956). Some problems in multivariate analysis, Part I. Statist. Dept., Stanford Univ., Stanford, CA, Tech. Rep.

[164] Stein, C. (1975). Estimation of a covariance matrix. Rietz Lecture. 39th Annual Meeting IMS, Atlanta, GA.

[165] Tanaka, F., & Komaki, F. (2008). A superharmonic prior for the autoregressive process of the second-order. Journal of Time Series Analysis, 29(3), 444-452.

[166] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), 267-288.

[167] Tracy, C. A., & Widom, H. (2002). Distribution functions for largest eigenvalues and their applications. arXiv preprint math-ph/0210034.

[168] Tsay, R. S. (2005). Analysis of financial time series (Vol. 543): Wiley-Interscience.

[169] Tyler, D. E. (1987). A distribution-free $ M $-estimator of multivariate scatter. The Annals of Statistics, 15(1), 234-251.

[170] van der Vaart, A. W. (2000). Asymptotic statistics (Cambridge series in statistical and probabilistic mathematics). Cambridge University Press,Cambridge UK.

[171] van Houwelingen, J. (2001). Shrinkage and penalized likelihood as methods to improve predictive accuracy. Statistica Neerlandica, 55(1), 17-34.

[172] Won, J.-H., Lim, J., Kim, S.-J., & Rajaratnam, B. (2009). Maximum likelihood covariance estimation with a condition number constraint. Technical Report 2009-10, Dept. Statistics, Stanford univ.

[173] Wu, W. B., & Pourahmadi, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. Biometrika, 90(4), 831-844.

[174] Wu, W. B., & Pourahmadi, M. (2009). Banding sample autocovariance matrices of stationary processes. Statistica Sinica, 19(4), 1755.

[175] Yang, R., & Berger, J. O. (1994). Estimation of a covariance matrix using the reference prior. The Annals of Statistics, 22(3), 1195-1211.

[176] Yuan, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. The Journal of Machine Learning Research, 99, 2261-2286.

[177] Yuan, M., & Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. Biometrika, 94(1), 19-35.

[178] Zhao, J.-H., Philip, L., & Jiang, Q. (2008). ML estimation for factor analysis: EM or non-EM? Statistics and Computing, 18(2), 109-123.

A Appendix: Fisher information for truncated normal distribution

The log-likelihood for truncated normal distribution is:

$$l(x;\theta) = \log f(x;\mu,\sigma^2) = -\log(\sqrt{2\pi}\sigma) - \frac{(x-\mu)^2}{2\sigma^2} - \log(1-\Phi(u^*))$$

where $u^* = \frac{u-\mu}{\sigma}$.

$$\frac{\partial l}{\partial \mu} = \frac{(x-\mu)}{\sigma^2} - \frac{\phi(u^*)}{1-\Phi(u^*)}\frac{1}{\sigma},$$

$$\frac{\partial l}{\partial \sigma} = -\frac{1}{\sigma} + \frac{(x-\mu)^2}{\sigma^3} - \frac{\phi(u^*)}{1-\Phi(u^*)}\frac{u-\mu}{\sigma^2},$$

Let $\alpha(u^*) = \frac{\phi(u^*)}{1-\Phi(u^*)}$,

$$\frac{\partial^2 l}{\partial \mu^2} = -\frac{1}{\sigma^2} + \frac{1}{\sigma^2}(\alpha^2(u^*) - u^*\alpha(u^*)),$$

$$\frac{\partial^2 l}{\partial \sigma^2} = \frac{1}{\sigma^2} + \frac{-3\sigma^2(x-\mu)^2}{\sigma^6} + (\alpha^2(u^*) - u^*\alpha(u^*)) \left(-\frac{u-\mu}{\sigma^2}\right)^2 + \alpha(u^*)\frac{2(u-\mu)}{\sigma^3},$$

$$\frac{\partial^2 l}{\partial \mu \partial \sigma} = \frac{-2(x-\mu)}{\sigma^3} + (\alpha^2(u^*) - u^*\alpha(u^*))\frac{u-\mu}{\sigma^3} + \alpha(u^*)\frac{1}{\sigma^2}.$$

For truncated normal distribution, the first and second moments are:

$$E[x-\mu] = \int_u^\infty (x-\mu)\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \frac{1}{1-\Phi(\frac{u-\mu}{\sigma})} dx = \sigma\alpha(u^*),$$

$$E[(x-\mu)^2] = \int_u^\infty (x-\mu)^2\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \frac{1}{1-\Phi(\frac{u-\mu}{\sigma})} dx = \sigma^2(1+u^*\alpha(u^*)),$$

Then, $I = \begin{bmatrix} A & B \\ B & C \end{bmatrix}$,where

$$A = \frac{1 + u^*\alpha(u^*) - \alpha^2(u^*)}{\sigma^2},$$

$$B = -E[\frac{\partial^2 l}{\partial \mu \partial \sigma}] = \frac{\alpha(u^*)(1 - \alpha(u^*)u^* + u^{*2})}{\sigma^2},$$

$$C = -E[\frac{\partial^2 l}{\partial \sigma^2}] = \frac{2 + u^*\alpha(u^*) - \alpha^2(u^*)u^{*^2} + u^{*3}\alpha(u^*)}{\sigma^2}.$$

B Appendix: Fisher information for truncated gamma distribution

Let $g(x;\kappa,\lambda) = x^{\kappa-1}\frac{e^{-x/\lambda}}{\Gamma(\kappa)\lambda^\kappa}$ be the density of gamma distribution, and $G(x;\kappa,\lambda)$ denotes the cdf, then the log-likelihood function is:

$$l(x;\kappa,\lambda) = \log f(x;\kappa,\lambda) = (\kappa-1)\ln x - \frac{x}{\lambda} - \ln\Gamma(\kappa) - \kappa\ln\lambda - \ln(1-G(u;\kappa,\lambda)).$$

Taking derivatives with respect to κ and λ,

$$\frac{\partial l}{\partial \kappa} = \ln x - \frac{\Gamma'(\kappa)}{\Gamma(k)} - \ln\lambda + \frac{\frac{\partial}{\partial\kappa}\int_0^u g(x;\kappa,\lambda)dx}{1-G(u;\kappa,\lambda)},$$

$$\frac{\partial l}{\partial \lambda} = \frac{x}{\lambda^2} - \frac{\kappa}{\lambda} + \frac{\frac{\partial}{\partial\lambda}\int_0^u g(x;\kappa,\lambda)dx}{1-G(u;\kappa,\lambda)}.$$

Let $H_\kappa = \frac{\partial}{\partial\kappa}\int_0^u g(x;\kappa,\lambda)dx, H_\lambda = \frac{\partial}{\partial\kappa}\int_0^u g(x;\kappa,\lambda)dx, H_{\kappa\kappa} = \frac{\partial H_\kappa}{\partial\kappa}, H_{\lambda\lambda} = \frac{\partial H_\lambda}{\partial\lambda}$,then

$$\frac{\partial^2 l}{\partial \kappa^2} = -\Psi'(\kappa) + \frac{H_{\kappa\kappa}(1-G(u;\kappa,\lambda)) + H_\kappa^2}{(1-G(u;\kappa,\lambda))^2} = -A,$$

$$\frac{\partial^2 l}{\partial\kappa\partial\lambda} = -\frac{1}{\lambda} + \frac{H_{\kappa\lambda}(1-G(u;\kappa,\lambda)) + H_\lambda H_\kappa}{(1-G(u;\kappa,\lambda))^2} = -B,$$

$$\frac{\partial^2 l}{\partial\lambda^2} = -\frac{2x}{\lambda^3} + \frac{\kappa}{\lambda^2} + \frac{H_{\lambda\lambda}(1-G(u;\kappa,\lambda)) + H_\lambda^2}{(1-G(u;\kappa,\lambda))^2}.$$

For truncated gamma, the first moment is:

$$\begin{aligned} E[x] &= \int_u^\infty x\frac{g(x;\kappa,\lambda)}{1-G(u;\kappa,\lambda)}dx \\ &= \frac{\kappa\lambda}{1-G(u;\kappa,\lambda)}\int_u^\infty x^\kappa \frac{e^{-x/\lambda}}{\Gamma(\kappa+1)\lambda^{\kappa+1}}dx \\ &= \frac{\kappa\lambda(1-G(u;\kappa+1,\lambda))}{1-G(u;\kappa,\lambda)}. \end{aligned}$$

Therefore,

$$C = -E[\frac{\partial^2 l}{\partial\lambda^2}] = \frac{2\kappa(1-G(u;\kappa+1,\lambda))}{\lambda^2(1-G(u;\kappa,\lambda))} - \frac{\kappa}{\lambda^2} - \frac{H_{\lambda\lambda}(1-G(u;\kappa,\lambda)) + H_\lambda^2}{(1-G(u;\kappa,\lambda))^2}.$$

Further,

$$H_\kappa = \frac{\partial}{\partial\kappa}\int_0^u g(x;\kappa,\lambda)dx = \int_0^u g_\kappa(x;\kappa,\lambda)dx = \int_0^u (\ln\frac{x}{\lambda} - \Psi(\kappa))g(x;\kappa,\lambda)dx,$$

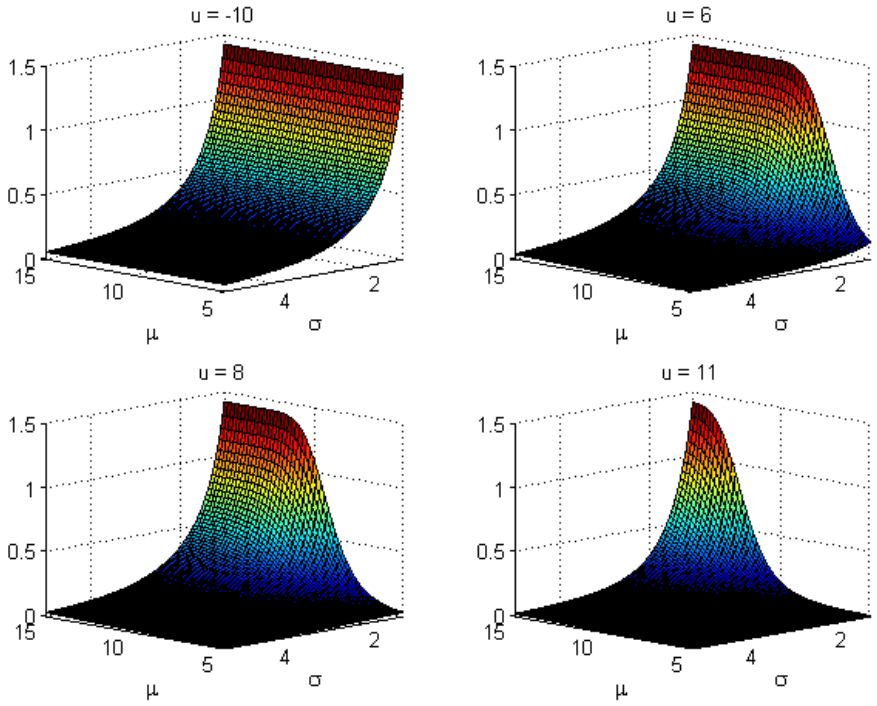
$$H_{\kappa\kappa} = \frac{\partial^2}{\partial\kappa^2}\int_0^u g(x;\kappa,\lambda)dx = -\Psi'(k)G(u;\kappa,\lambda) + \int_0^u (\ln\frac{x}{\lambda} - \Psi(\kappa))^2 g(x;\kappa,\lambda)dx,$$

$$\begin{aligned} H_\lambda &= \frac{\partial}{\partial\lambda}\int_0^u g(x;\kappa,\lambda)dx = \frac{\partial}{\partial\lambda}\int_0^u x^{\kappa-1}\frac{e^{-x/\lambda}}{\Gamma(\kappa)\lambda^\kappa}dx \\ &= \frac{\partial}{\partial\lambda}\int_0^{u/\lambda} y^{\kappa-1}\frac{e^{-y}}{\Gamma(\kappa)}dy \text{ (let } y = x/\lambda) \\ &= \frac{\partial}{\partial u^*}\left(\int_0^{u^*} y^{\kappa-1}\frac{e^{-y}}{\Gamma(\kappa)}dy\right)\frac{\partial u^*}{\partial\lambda}(\text{let } u^* = u/\lambda) \\ &= -\frac{u}{\lambda^2}g(\frac{u}{\lambda};\kappa,1) = -\frac{u}{\lambda}g(u;\kappa,\lambda), \end{aligned}$$

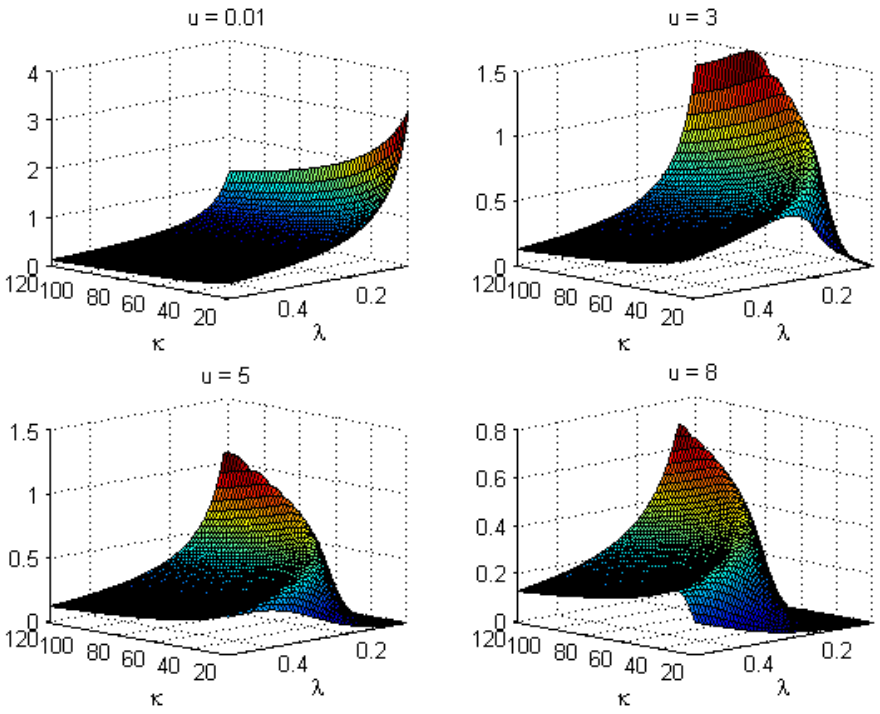
$$\begin{aligned} H_{\lambda\lambda} &= \frac{\partial H_\lambda}{\partial\lambda} = \frac{\partial}{\partial\lambda}(-\frac{u}{\lambda}\frac{u^{\kappa-1}e^{-u/\lambda}}{\Gamma(\kappa)\lambda^\kappa}) = -\frac{u^\kappa}{\Gamma(\kappa)}\frac{\partial}{\partial\lambda}(\frac{e^{-u/\lambda}}{\lambda^{\kappa+1}}) \\ &= -\frac{u^\kappa}{\Gamma(\kappa)}\left(\frac{u}{\lambda^2}\frac{e^{-u/\lambda}}{\lambda^{\kappa+1}} - e^{-u/\lambda}(\kappa+1)\lambda^{-(\kappa+2)}\right) \\ &= \frac{\kappa(\kappa+1)}{\lambda}(g(u;\kappa+1,\lambda) - g(u;\kappa+2,\lambda)), \end{aligned}$$

$$H_{\kappa\lambda} = \frac{\partial H_\lambda}{\partial\kappa} = -\frac{u}{\lambda}(\ln\frac{u}{\lambda} - \Psi(\kappa))g(u;\kappa,\lambda).$$

C Appendix: Jeffreys' prior for truncated normal distribution



D Appendix: Jeffreys' prior for truncated gamma distribution



E Appendix: Jeffreys' prior for general truncated distributions

Roehr (2002) provided a general derivation of the Fisher information matrix for truncated distributions.

Consider a left-truncated sample from a distribution with density $f(x,\theta)$, and suppose the truncation threshold is known as u,then the truncated distribution function is

$$g(x;u,\theta) = \frac{f(x,\theta)}{1-F(u,\theta)} = \frac{f(x,\theta)}{\bar{F}(u,\theta)}$$

For the complete distribution $f(x,\theta)$, the fisher information matrix

$$\begin{aligned} I(\theta)_{jk} &= E[\partial_{\theta_j} \log f(x,\theta) \partial_{\theta_k} \log f(x,\theta)] \\ &= \int_{-\infty}^{\infty} \partial_{\theta_j} \log f(x,\theta) \partial_{\theta_k} \log f(x,\theta) \cdot f(x,\theta) dx \end{aligned}$$

For the truncated distribution $g(x;u,\theta)$, the Fisher information matrix

$$\begin{aligned} I(u,\theta)_{jk} &= E[\partial_{\theta_j} \log g(x;u,\theta) \partial_{\theta_k} \log g(x;u,\theta)] \\ &= E[\partial_{\theta_j} \log \frac{f(x,\theta)}{\bar{F}(u,\theta)} \partial_{\theta_k} \log \frac{f(x,\theta)}{\bar{F}(u,\theta)}] \\ &= \int_{u}^{\infty} \partial_{\theta_j} \log \frac{f(x,\theta)}{\bar{F}(u,\theta)} \partial_{\theta_k} \log \frac{f(x,\theta)}{\bar{F}(u,\theta)} \cdot \frac{f(x,\theta)}{\bar{F}(u,\theta)} dx \end{aligned}$$

Suppose the support for $f(x,\theta)$ is R^+,then $g(x;0,\theta) = f(x,\theta)$,and $I(0,\theta) = I(\theta)$. In the following we use this notation. (If the support is R, then $I(-\infty,\theta) = I(\theta)$.)

Note the following formula holds for any function $g(x,u)$:

$$\partial_u \int_u^\infty g(x,u)dx = -g(u,u) + \int_u^\infty \partial_u g(x,u)dx$$

$$\begin{aligned}\partial_u(I(u,\theta)_{jk}) &= -\partial_{\theta_j}\log\frac{f(u,\theta)}{\bar{F}(u,\theta)}\partial_{\theta_k}\log\frac{f(u,\theta)}{\bar{F}(u,\theta)}\cdot\frac{f(u,\theta)}{\bar{F}(u,\theta)} \\ &\quad + \int_u^\infty \partial_u\left[\partial_{\theta_j}\log\frac{f(x,\theta)}{\bar{F}(u,\theta)}\partial_{\theta_k}\log\frac{f(x,\theta)}{\bar{F}(u,\theta)}\cdot\frac{f(x,\theta)}{\bar{F}(u,\theta)}\right]dx\end{aligned}$$

Let $H(u,\theta) = \log\frac{f(u,\theta)}{\bar{F}(u,\theta)}$, then we write the first term in the RHS as:

$$-H_{\theta_j}H_{\theta_k}\frac{f(u,\theta)}{\bar{F}(u,\theta)}$$

The second term in the RHS is:

$$\begin{aligned}&\int_u^\infty \partial_u\left[\partial_{\theta_j}\log\frac{f(x,\theta)}{\bar{F}(u,\theta)}\partial_{\theta_k}\log\frac{f(x,\theta)}{\bar{F}(u,\theta)}\cdot\frac{f(x,\theta)}{\bar{F}(u,\theta)}\right]dx \\ = &\int_u^\infty \partial_u\left[\partial_{\theta_j}\log\frac{f(x,\theta)}{\bar{F}(u,\theta)}\partial_{\theta_k}\log\frac{f(x,\theta)}{\bar{F}(u,\theta)}\right]\cdot\frac{f(x,\theta)}{\bar{F}(u,\theta)}dx \\ &+\int_u^\infty \partial_{\theta_j}\log\frac{f(x,\theta)}{\bar{F}(u,\theta)}\partial_{\theta_k}\log\frac{f(x,\theta)}{\bar{F}(u,\theta)}\partial_u\frac{f(x,\theta)}{\bar{F}(u,\theta)}dx\end{aligned}$$

in which the first term is:

$$\begin{aligned}&\int_u^\infty \partial_u\left[\partial_{\theta_j}\log\frac{f(x,\theta)}{\bar{F}(u,\theta)}\partial_{\theta_k}\log\frac{f(x,\theta)}{\bar{F}(u,\theta)}\right]\cdot\frac{f(x,\theta)}{\bar{F}(u,\theta)}dx \\ = &\int_u^\infty \partial_u\left[\partial_{\theta_j}\log\frac{f(x,\theta)}{\bar{F}(u,\theta)}\right]\partial_{\theta_k}\log\frac{f(x,\theta)}{\bar{F}(u,\theta)}\cdot\frac{f(x,\theta)}{\bar{F}(u,\theta)}dx \\ &+\int_u^\infty \partial_u\left[\partial_{\theta_k}\log\frac{f(x,\theta)}{\bar{F}(u,\theta)}\right]\partial_{\theta_j}\log\frac{f(x,\theta)}{\bar{F}(u,\theta)}\cdot\frac{f(x,\theta)}{\bar{F}(u,\theta)}dx\end{aligned}$$

and we can show this is equal to 0. Note that

$$\begin{aligned}\partial_u\left(\partial_{\theta_j}\log\frac{f(x,\theta)}{\bar{F}(u,\theta)}\right) &= \partial_{\theta_j}\left(\partial_u\log\frac{f(x,\theta)}{\bar{F}(u,\theta)}\right) \\ &= \partial_{\theta_j}\left(\frac{\bar{F}(u,\theta)}{f(x,\theta)}\partial_u\frac{f(x,\theta)}{\bar{F}(u,\theta)}\right)\end{aligned}$$

and use

$$\partial_u\frac{f(x,\theta)}{\bar{F}(u,\theta)} = \frac{f(x,\theta)f(u,\theta)}{\bar{F}(u,\theta)\bar{F}(u,\theta)}$$

then,

$$\partial_u \left(\partial_{\theta_j} \log \frac{f(x,\theta)}{\bar{F}(u,\theta)} \right) = \partial_{\theta_j} \left(\frac{f(u,\theta)}{\bar{F}(u,\theta)} \right)$$

So,

$$\begin{aligned}
& \int_u^\infty \partial_u \left[\partial_{\theta_j} \log \frac{f(x,\theta)}{\bar{F}(u,\theta)} \right] \partial_{\theta_k} \log \frac{f(x,\theta)}{\bar{F}(u,\theta)} \cdot \frac{f(x,\theta)}{\bar{F}(u,\theta)} dx \\
= & \int_u^\infty \partial_{\theta_j} \left(\frac{f(u,\theta)}{\bar{F}(u,\theta)} \right) \partial_{\theta_k} \log \frac{f(x,\theta)}{\bar{F}(u,\theta)} \cdot \frac{f(x,\theta)}{\bar{F}(u,\theta)} dx \\
= & \partial_{\theta_j} \left(\frac{f(u,\theta)}{\bar{F}(u,\theta)} \right) \int_u^\infty \partial_{\theta_k} \log \frac{f(x,\theta)}{\bar{F}(u,\theta)} \cdot \frac{f(x,\theta)}{\bar{F}(u,\theta)} dx \\
= & \partial_{\theta_j} \left(\frac{f(u,\theta)}{\bar{F}(u,\theta)} \right) \int_u^\infty \frac{\bar{F}(u,\theta)}{f(x,\theta)} \partial_{\theta_k} \frac{f(x,\theta)}{\bar{F}(u,\theta)} \cdot \frac{f(x,\theta)}{\bar{F}(u,\theta)} dx \\
= & \partial_{\theta_j} \left(\frac{f(u,\theta)}{\bar{F}(u,\theta)} \right) \int_u^\infty \partial_{\theta_k} \frac{f(x,\theta)}{\bar{F}(u,\theta)} dx \\
= & \partial_{\theta_j} \left(\frac{f(u,\theta)}{\bar{F}(u,\theta)} \right) \partial_{\theta_k} \left(\int_u^\infty \frac{f(x,\theta)}{\bar{F}(u,\theta)} dx \right) \\
= & 0
\end{aligned}$$

Therefore,

$$\begin{aligned}
\partial_u (I(u,\theta)_{jk}) &= -H_{\theta_j} H_{\theta_k} \frac{f(u,\theta)}{\bar{F}(u,\theta)} + \int_u^\infty \partial_{\theta_j} \log \frac{f(x,\theta)}{\bar{F}(u,\theta)} \partial_{\theta_k} \log \frac{f(x,\theta)}{\bar{F}(u,\theta)} \partial_u \frac{f(x,\theta)}{\bar{F}(u,\theta)} dx \\
&= -H_{\theta_j} H_{\theta_k} \frac{f(u,\theta)}{\bar{F}(u,\theta)} + \frac{f(u,\theta)}{\bar{F}(u,\theta)} \int_u^\infty \partial_{\theta_j} \log \frac{f(x,\theta)}{\bar{F}(u,\theta)} \partial_{\theta_k} \log \frac{f(x,\theta)}{\bar{F}(u,\theta)} \frac{f(x,\theta)}{\bar{F}(u,\theta)} dx
\end{aligned}$$

which is

$$\begin{aligned}
\partial_u (I(u,\theta)_{jk}) \bar{F}(u,\theta) &= -H_{\theta_j} H_{\theta_k} f(u,\theta) + f(u,\theta) \int_u^\infty \partial_{\theta_j} \log \frac{f(x,\theta)}{\bar{F}(u,\theta)} \partial_{\theta_k} \log \frac{f(x,\theta)}{\bar{F}(u,\theta)} \frac{f(x,\theta)}{\bar{F}(u,\theta)} dx \\
&= -H_{\theta_j} H_{\theta_k} f(u,\theta) + f(u,\theta) I(u,\theta)_{jk}
\end{aligned}$$

We also have,

$$\begin{aligned}
\partial_u \left(I(u,\theta)_{jk} \bar{F}(u,\theta) \right) &= \partial_u (I(u,\theta)_{jk}) \bar{F}(u,\theta) - I(u,\theta)_{jk} f(u,\theta) \\
&= -H_{\theta_j} H_{\theta_k} f(u,\theta)
\end{aligned}$$

Therefore,

$$\begin{aligned}
I(u,\theta)_{jk} \bar{F}(u,\theta) &= I(0,\theta)_{jk} \bar{F}(0,\theta) + \int_0^u \partial_z \left(I(z,\theta)_{jk} \bar{F}(z,\theta) \right) dz \\
&= I(0,\theta)_{jk} + \int_0^u \partial_z \left(I(z,\theta)_{jk} \bar{F}(z,\theta) \right) dz \\
&= I(0,\theta)_{jk} - \int_0^u H_{\theta_j} H_{\theta_k} f(z,\theta) dz
\end{aligned}$$

where $H_{\theta_j}(x,\theta) = \partial_{\theta_j} \log \frac{f(x;\theta)}{1-F(x;\theta)}$.

When exploring the behaviors of the Jeffreys' priors for truncated normal and truncated gamma distributions, we found that they are always below the Jeffreys' prior for the complete distributions. So it's interesting to see if this is true for all distributions.

Conjecture 1. *The Jeffreys' priors for truncated distributions are bounded by the Jeffreys' priors for complete distributions, i.e.*

$$det\left(I(u,\theta)\right) < \det(I(\theta))$$

where $I(u,\theta)$ *is the Jeffreys' prior for truncated distribution* $\frac{f(\theta)}{1-F(u)}$ *while* $I(\theta)$ *is the Jeffreys' prior for full distribution* $f(\theta)$.

F Appendix: EM algorithm for multivariate-t distribution with fixed degrees of freedom ν

The log-likelihood is (first assume the d.o.f v is known, and omit the constants):

$$L(X;\mu,\Sigma,v) = -\frac{n}{2}\log|\Sigma| - \frac{v+p}{2}\sum_{i=1}^{n}\log\left[v+(x-\mu)^T\Sigma^{-1}(x-\mu)\right]$$

The multivariate t-distribution belongs to the normal variance mixture family. The n indepenent draws from a multivariate t distribution $t_p(\mu,\Sigma,v)$ can be described as:

$$Y_i|\mu,\Sigma,\tau \sim N_p(\mu,\Sigma/\tau_i), i=1,2,...,n$$

and

$$\tau_i|v \sim \Gamma(v/2,v/2), i=1,2,...,n.$$

where $\Gamma(x;\alpha,\beta) \sim \frac{x^{\alpha-1}e^{-\beta x}\beta^{\alpha}}{\Gamma(\alpha)}$is the gamma distribution.

The EM algorithm for estimating multivariate t-distribution can be derived using this representation by treating the τ_i as a latent variable (see Liu and Rubin, 1995):

Notice that gamma distribution is a conjugate prior, if

$$\tau \sim \Gamma(v/2,v/2),$$

then,

$$\begin{aligned}
p(\tau|x) &\propto p(x|\tau)p(\tau) \\
&\propto N_p(\mu,\Sigma/\tau)\Gamma(v/2,v/2) \\
&\propto (2\pi)^{-p/2}\left|\frac{\Sigma}{\tau}\right|^{-\frac{p}{2}} e^{-\frac{1}{2}(x-\mu)^T(\frac{\Sigma}{\tau})^{-1}(x-\mu)}\frac{\tau^{\frac{v}{2}-1}e^{-\frac{v}{2}\tau}\left(\frac{v}{2}\right)^{\frac{v}{2}}}{\Gamma(v/2)} \\
&\propto \tau^{\frac{v}{2}+\frac{p}{2}-1}e^{-\frac{v+\delta_x}{2}\tau} \sim \Gamma(\frac{v+p}{2},\frac{v+\delta_x}{2})
\end{aligned}$$

where $\delta_x = (x-\mu)^T\Sigma^{-1}(x-\mu)$. Therefore,

$$E[\tau|X] = \frac{v+p}{v+\delta_x}.$$

The log-likilihood function of the t distribution can be represented as:

$$L(\mu, \Sigma, v|Y, \tau) = L_N(\mu, \Sigma|Y, \tau) + L_G(v|\tau)$$

where

$$\begin{aligned} L_N(\mu, \Sigma, v|Y, \tau) &= -\frac{n}{2}\ln|\Sigma| - \frac{1}{2}Tr(\Sigma^{-1}\sum_{i=1}^n \tau_i Y_i Y_i^T) \\ &\quad +\mu^T\Sigma^{-1}\sum_{i=1}^n \tau_i Y_i - \frac{1}{2}\mu^T\Sigma^{-1}\mu\sum_{i=1}^n \tau_i, \end{aligned}$$

and

$$L_G(v|\tau) = -n\ln\Gamma(v/2) + \frac{nv}{2}\ln(v/2) + v/2\sum_{i=1}^n(\ln\tau_i - \tau_i).$$

The complete-data sufficient statistics for μ, Σ, v are:

$$\begin{aligned} S_{\tau Y} &= \sum_{i=1}^n \tau_i Y_i, S_{\tau YY} = \sum_{i=1}^n \tau_i Y_i Y_i^T, \\ S_\tau &= \sum_{i=1}^n \tau_i, S_{\tau\tau} = \sum_{i=1}^n(\ln\tau_i - \tau_i). \end{aligned}$$

Therefore, the EM algorithm steps are:

- E-step:

  Let $\Omega^{(t)} = \{Y, \mu^{(t)}, \Sigma^{(t)}, v\}$,

  $$w_i^{(t+1)} = E[\tau_i|\Omega^{(t)}] = \frac{v+p}{v+\delta_i^{(t)}},$$

  where $\delta_i^{(t)} = (Y_i - \mu^{(t)})^T(\Sigma^{(t)})^{-1}(Y_i - \mu^{(t)})$.

  $$\begin{aligned} S_\tau^{(t+1)} &= E[\sum_{i=1}^n \tau_i|\Omega^{(t)}] = \sum_{i=1}^n w_i^{(t+1)}, \\ S_{\tau Y}^{(t+1)} &= E[\sum_{i=1}^n \tau_i Y_i|\Omega^{(t)}] = \sum_{i=1}^n w_i^{(t+1)} Y_i, \\ S_{\tau YY}^{(t+1)} &= E[\sum_{i=1}^n \tau_i Y_i Y_i^T|\Omega^{(t)}] = 1/n\sum_{i=1}^n w_i^{(t+1)} Y_i Y_i^T. \end{aligned}$$

- M-step:

  $\mu^{(t+1)}$ and $\Sigma^{(t+1)}$ can be obtained by solving:

  $$\begin{aligned} \max_{\mu,\Sigma} L_N(\mu, \Sigma, v|Y, \tau) &= -\frac{n}{2}\ln\left|\Sigma^{(t)}\right| - \frac{n}{2}Tr(\Sigma^{(t)-1}S_{\tau YY}^{(t+1)}) \\ &\quad +\mu^{(t)T}\Sigma^{(t)-1}S_{\tau Y}^{(t+1)} - \frac{1}{2}\mu^T\Sigma^{-1}\mu S_\tau^{(t+1)}, \end{aligned}$$

  Particularly, when μ is known to be 0,the iteration of covariance estimate is:

  $$\Sigma^{(t+1)} = \arg\max -\frac{n}{2}\ln\left|\Sigma^{(t)}\right| - \frac{n}{2}Tr(\Sigma^{(t)-1}S_{\tau YY}^{(t+1)})$$

  which is similar to the MLE for Gaussian density but here the sample covariance is replaced as a weighted average sum of squares of the observations. The explicit solution is: $\Sigma^{(t+1)} = S_{\tau YY}^{(t+1)}$.

G Appendix: Relation between KL and QL

Proof): Using the following formula

$$\log\det(I+\delta V)=\sum_{n=1}^{\infty}\frac{(-1)^{n-1}}{n}\delta^n tr(V^n),$$

we have

$$\begin{aligned}\log\det(PQ^{-1}) &= \log\det(I+PQ^{-1}-I)\\ &= tr(PQ^{-1}-I)-\frac{1}{2}tr((PQ^{-1}-I)^2)+...\\ &\approx tr(PQ^{-1})-p-\frac{1}{2}tr((PQ^{-1}-I)^2)\end{aligned}$$

Thus,

$$\begin{aligned}2D_{KL}(P||Q) &= tr(PQ^{-1})-\log\det(PQ^{-1})-p\\ &\approx \frac{1}{2}tr((PQ^{-1}-I)^2)=\frac{1}{2}D_{QL}(P,Q).\end{aligned}$$

H Appendix: Hellinger distance between two Gaussian

Proof):

$$\begin{aligned}H^2(P,Q) &= 1-\int\frac{1}{(2\pi)^{d/2}|P|^{1/4}|Q|^{1/4}}\exp\left\{-\frac{x^T(P^{-1}+Q^{-1})x}{4}\right\}dx\\ &= 1-\frac{\left|\left(\frac{P^{-1}+Q^{-1}}{2}\right)^{-1}\right|^{1/2}}{|P|^{1/4}|Q|^{1/4}}\int\frac{1}{(2\pi)^{d/2}\left|\left(\frac{P^{-1}+Q^{-1}}{2}\right)^{-1}\right|^{1/2}}\exp\left\{-\frac{x^T(P^{-1}+Q^{-1})x}{4}\right\}dx\\ &= 1-\frac{\left|\left(\frac{P^{-1}+Q^{-1}}{2}\right)^{-1}\right|^{1/2}}{|P|^{1/4}|Q|^{1/4}}=1-\frac{\left|\frac{P^{-1}+Q^{-1}}{2}\right|^{-1/2}}{|P|^{1/4}|Q|^{1/4}}\\ &= 1-\left|\frac{P+Q}{2PQ}\right|^{-1/2}|P|^{-1/4}|Q|^{-1/4}\\ &= 1-|P|^{1/4}|Q|^{1/4}\left|\frac{P+Q}{2}\right|^{-1/2}\end{aligned}$$