**Journal of Data Analysis and Information Processing, 2016, 4, 13-20**
Published Online February 2016 in SciRes. http://www.scirp.org/journal/jdaip
http://dx.doi.org/10.4236/jdaip.2016.41002

# High Dimensionality Effects on the Efficient Frontier: A Tri-Nation Study

**Rituparna Sen[1], Pulkit Gupta[2], Debanjana Dey[3]**

[1]Indian Statistical Institute, Chennai, India
[2]Indian Institute of Technology, Kharagpur, India
[3]Indian Institute of Management, Calcutta, India
Email: rsen@isichennai.res.in

Received 5 December 2015; accepted 12 February 2016; published 15 February 2016



## Abstract

**Markowitz Portfolio theory under-estimates the risk associated with the return of a portfolio in case of high dimensional data. El Karoui mathematically proved this in [1] and suggested improved estimators for unbiased estimation of this risk under specific model assumptions. Norm constrained portfolios have recently been studied to keep the effective dimension low. In this paper we consider three sets of high dimensional data, the stock market prices for three countries, namely US, UK and India. We compare the Markowitz efficient frontier to those obtained by unbiasedness corrections and imposing norm-constraints in these real data scenarios. We also study the out-of-sample performance of the different procedures. We find that the 2-norm constrained portfolio has best overall performance.**

## Keywords

**High Dimensional Covariance Matrix Estimation, Minimum-Variance Portfolio, Norm Con-Strained Portfolio**

## 1. Introduction

The need for solutions to optimization problems in a high dimensional setting is increasing in the finance industry with huge amount of data being generated every day. Many empirical studies indicate that minimum variance portfolios in general lead to a better out-of-sample performance than stock index portfolios [2] [3]. Markowitz Portfolio theory, the most popular method for portfolio optimization, develops a serious drawback namely risk underestimation. When implementing portfolio optimization according to [4], one needs to estimate the expected asset returns as well as the corresponding variances and covariances. El Karoui studied the Markowitz problem

**How to cite this paper:** Sen, R., Gupta, P. and Dey, D. (2016) High Dimensionality Effects on the Efficient Frontier: A Tri-Nation Study. *Journal of Data Analysis and Information Processing*, **4**, 13-20.
http://dx.doi.org/10.4236/jdaip.2016.41002

as a solution to quadratic problems in [1] and [5] to establish a relationship between the two types of solution viz. one computed using population data and another estimated from sample data. This relationship is important and particularly relevant for high dimensional data where one suspects that the difference between the two may be considerable.

There is a broad literature which addresses the question of how to reduce estimation risk in portfolio optimization. De Miguel *et al.* compare portfolio strategies which differ in the treatment of estimation risk in [6] and confirm that the considered strategies perform better than the traditional plug-in implementation of Markowitz optimization. Constrained minimum-variance portfolios have been frequently advocated in the literature (see [7]-[10]).

The main aim of this paper is to compare the efficient frontier for real data based on corrected estimators of [5] and norm-constrained portfolios. One natural advantage of norm-constrained optimization is that it leads to sparse solutions, which many of the portfolio weights are zero. Such a portfolio is preferable in terms of transaction costs. On the other hand, if Gaussian assumptions are valid, then the corrected frontier is indeed the most efficient. Another advantage is that one can obtain a confidence interval for the variance at each value of return.

We carry out our analysis for three scenarios namely the Indian stock market, London Stock market and U.S stock market to facilitate a comparative study and to conclude about the uniformity of our results. We use constituent stocks of NSE CNX 100, FTSE 100 and S&P 100 respectively for the three scenarios as our data base taking daily data from 1st Jan 2013 to 1st Jan 2014 time span. The daily returns data are publicly available from NSE India and yahoo finance. Thus we have at our disposal, 100 stocks for each country with 250 observations per stock. In other words, considering $p$ to be the number of assets and $n$ to be the number of observations per asset, we arrive at a large $p$, large $n$ setting which in modern statistical parlance can be considered to be a high dimensional setting.

The rest of the paper is organized as follows. Section 2 is committed to explaining the modern portfolio theory. Section 3 deals with identifying the underestimation factors and the bias inherent in the plug in estimators and subsequently eliminating them from the empirical optimized portfolio, to arrive at the final error-free optimized weights. Section 4 deals with norm constrained models. In section 5, we present the empirical results of comparing the efficient frontiers obtained from Markowitz portfolio to error-free efficient frontier and norm constrained portfolio efficient frontiers. We present our conclusions in section 6.

## 2. Markowitz Portfolio Theory

Markowitz portfolio theory [4] is a classic portfolio optimization problem in finance where investors choose to invest according to the following framework: one picks the assets in such a way that the portfolio guarantees a certain level of expected returns but minimizes the "risk" associated with it. In standard framework this risk is measured by the variance of the portfolio whereas the expectation by the mean of the portfolio. The set-up is as follows:

- There is an opportunity to invest in p assets $A_1, A_2, \cdots, A_p$.
- The mean returns are represented by a p-dimensional vector $\mu = \mu_1, \mu_2, \cdots, \mu_p$.
- The covariance matrix of the returns is denoted by $\Sigma$.
- The aim is to create a portfolio with guaranteed mean return $\mu_k$ and minimize the risk as measured by the variance.
- The problem is to find the weights or amount allocated to various assets of the portfolio.

Note that $\Sigma$ is positive semi definite and symmetric. In ideal situation the means, variances and covariance are known and the problem is the following quadratic programming problem:

$$\text{Min } w^{\mathrm{T}}\Sigma w \quad \text{subject to } w^{\mathrm{T}}1_p = 1 \text{ and } w^{\mathrm{T}}\mu = \mu_k \tag{1}$$

Here $1_p$ is a p-dimensional vector with one in every entry.

In practice, $\Sigma$ and $\mu$ are unknown. The most common procedure known as plug-in implementation replaces them with their sample estimators as follows to obtain the optimal weights.

$$\hat{\Sigma} = \frac{1}{n-1}\left(X - \hat{\mu}1_n^{\mathrm{T}}\right)\left(X - \hat{\mu}1_n^{\mathrm{T}}\right)^{\mathrm{T}} \text{ and } \hat{\mu} = \frac{1}{n}X1_n \tag{2}$$

With $X = \left(X_1, \cdots, X_n\right)$ is a $p \times n$ matrix of the returns of the assets. It is assumed that the columns of $X$ are

independent multivariate Normal vectors

If $\hat{\Sigma}$ is invertible with $w_{opt}$ is representing the solution of the above quadratic problem then,

$$w_{opt} = \hat{\Sigma}^{-1}\hat{V}\hat{M}^{-1}U \quad \text{and} \quad \hat{M} = \hat{V}^{\mathrm{T}}\hat{\Sigma}^{-1}\hat{V} \tag{3}$$

where $\hat{V}$ a $p \times 2$ matrix is whose first column are all unity and second column are the estimated means. Also U is the 2 dimensional vector with first entry being 1 and the second entry being $\mu_k$.

The curve $w_{opt}^{\mathrm{T}}\Sigma w_{opt}$ seen as a function of $\mu_k$ is called the *efficient frontier*.

## 3. Corrected Frontier Using Gaussian Assumption

In the Markowitz setting, let us assume that the returns have normal distribution. We shall assume $n$ and $p$ both go to infinity and each $X_i \sim \mathrm{N}_p(\mu, \Sigma)$ independently and identically. The parameters of the distribution are estimated using sample estimators defined in (2).

We have from Corollary 3.3 of [1],

$$w_{emp}^{\mathrm{T}}\hat{\Sigma}w_{emp} = \left(1-\frac{p-k}{n-1}\right)\left(w_{theo}^{\mathrm{T}}\Sigma w_{theo} - \frac{p}{n}\frac{\left(U^{\mathrm{T}}M^{-1}e_k\right)^2}{\left(1+\frac{p}{n}e_k^{\mathrm{T}}M^{-1}e_k\right)}\right) + o_P\left(w_{theo}^{\mathrm{T}}\Sigma w_{theo}\sqrt{n^{-1/2}}\right) \tag{4}$$

where $M = \hat{V}^{\mathrm{T}}\Sigma^{-1}\hat{V}$ is the population quantity, $k$ being the number of constraints in the quadratic problem we are solving which in our case will be equal to 2, $w_{emp}$ represents the weights obtained from the empirical data at hand while $w_{theo}$ is its population counterpart. $e_i$ denotes the canonical basis vectors in $\mathbb{R}^k$. The corollary shows that the effects of both covariance and mean estimation are to underestimate the risk and the empirical frontier is asymptotically deterministic. The cost of not knowing the covariance matrix and estimating it is captured in the factor $\left(1-\frac{p-k}{n-1}\right)$. In other words using plug in procedures leads to over optimistic conclusions in this situation.

Also when $\frac{p}{n} \to \rho \in (0,1)$ and $\alpha = \frac{p}{n} + o\left(n^{-1/2}\right)$ and we denote $\delta_n = \frac{\left(U^{\mathrm{T}}M^{-1}e_k\right)^2}{\left(1+\frac{p}{n}e_k^{\mathrm{T}}M^{-1}e_k\right)}$ the impact of the estimation of $\mu$ by $\hat{\mu}$ will be risk underestimation by the amount $\alpha\delta_n$. Hence rearranging (4) and subtracting the bias associated with mean and covariance estimation, from our variances obtained from sample data we get the error-free actual quantities of interest. In other words,

$$w_{theo}^{\mathrm{T}}\Sigma w_{theo} = \frac{w_{emp}^{\mathrm{T}}\hat{\Sigma}w_{emp}}{\left(1-\frac{p-k}{n-1}\right)} + \frac{p}{n}\frac{\left(U^{\mathrm{T}}M^{-1}e_k\right)^2}{\left(1+\frac{p}{n}e_k^{\mathrm{T}}M^{-1}e_k\right)} \tag{5}$$

The estimator $w_{theo}$ for proposed in [1] is a modified version of the optimal solution in equation (3). The modification is to replace $M$ by $\tilde{M} = \hat{M} - ke_ke_k^{\mathrm{T}}$.

It is also shown in Theorem 5.1 of [1] that the risk is indeed underestimated by the empirical frontier. Specifically,

$$f_{emp} \le f_{theo}$$

where $f_{emp}$ and $f_{theo}$ are respectively the empirical frontier with Gaussian distributed data and the theoretical efficient frontier.

We use the 95% confidence intervals for the variance of a single Normal variable with unknown mean μ and standard deviation $\sigma$ given by:

$$\left(\frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}}\right)$$

where $s^2$ is the sample variance and $(n-1)s^2/\sigma^2$ follows a $\chi^2$ distribution with $n-1$ degrees of freedom, the confidence coefficient being equal to 0.05.

## 4. Constraining the Portfolio

The short sale constrained minimum-variance portfolio, $w_{MINC}$ is introduced in [7]. This is the solution to problem (1) with the additional constraint that the portfolio weights be nonnegative.

### 4.1. 1-Norm Constrained Portfolio

The 1-norm-constrained portfolio, $w_{NC1}$, is the solution to the traditional minimum-variance portfolio problem (1) subject to the additional constraint that the $L_1$-norm of the portfolio-weight vector be smaller than or equal to a certain threshold $c$; that is,

$$\|w\|_1 = \sum_{i=1}^{p} |w_i| \le c \tag{6}$$

1-norm constrained portfolio problem can be summarized as

$$\min_{w^{\mathrm{T}} 1_p = 1, \|w\|_1 \le c} w^{\mathrm{T}} \Sigma w \tag{7}$$

Markowitz risk minimization problem can be recast as a regression problem.

$$\mathrm{var}\left(w^{\mathrm{T}} R\right) = \min_b E\left[\left(w^{\mathrm{T}} R - b\right)^2\right] \tag{8}$$

By using the fact that the sum of total weights is one, we have

$$\mathrm{var}\left(w^{\mathrm{T}} R\right) = \min_b E\left[\left(Y - \omega_1 Z_1 - \cdots - \omega_{p-1} Z_{p-1} - b\right)^2\right] \tag{9}$$

where $R$ = Return vector, $Y = R_p$ and $Z_j = R_p - R_j$ where $j = (1, \cdots, p-1)$.

Finding the optimal weight $w$ is the same as finding the regression coefficient $w^* = \left(\omega_1, \omega_2, \cdots, \omega_{p-1}\right)^{\mathrm{T}}$. The gross-exposure constraint $\|w\|_1 \le c$ can now be expressed as $\|w^*\|_1 \le c - \left|1 - 1^{\mathrm{T}} w^*\right| = \delta\,(\text{say})$. Thus the problem (7) is similar to

$$\min_{b, \|w^*\|_1 \le \delta} E\left[Y - \left(w^*\right)^{\mathrm{T}} Z - b\right] \tag{10}$$

where $Z = \left(Z_1, \cdots, Z_{p-1}\right)^{\mathrm{T}}$ but they are not equivalent. The latter depends on choice of $Y$, while the former does not. Efron *et al.* developed an efficient algorithm in [11] by using the least-angle regression (LARS), called the LARS-LASSO algorithm, to efficiently find the whole solution path $w_{NC1}(c)$, for all $c \ge 0$, to (10). The number of non-vanishing weights varies as c ranges from 0 to ∞. It recruits successively more assets and gradually all assets. The algorithm works iteratively as follows:

$$\begin{aligned} w_{NC1_i} &= w_i^* + \left(1 - \sum_{j=1}^{p-1} w_j^*\right) \times \left(w_i\right) \quad \text{where } i \in (1, 2, \cdots, p-1) \\ w_{NC1_p} &= \left(1 - \sum_{j=1}^{p-1} w_j^*\right) \times \left(w_p\right) \end{aligned} \tag{11}$$

Here our objective is to minimize the out-of-sample portfolio variance. To choose $c$ we use leave-one-out-cross validation (see [12]).

### 4.2. 2-Norm Constrained Portfolio

The 2-norm-constrained portfolio, $w_{NC2}$, is the solution to the traditional minimum-variance portfolio problem

(1) subject to the additional constraint that the $L_2$-norm of the portfolio-weight vector is smaller than or equal to a certain threshold $c$; that is,

$$\|w\|_2 = \left(\sum_{i=1}^{p} w_i^2\right)^{1/2} \le c \tag{12}$$

2-norm constrained portfolio problem can be summarized as

$$\min_{w^{\mathrm{T}} 1_p = 1, \|w\|_2 \le c} w^{\mathrm{T}} \Sigma w \tag{13}$$

Similar to the 1-norm constrained portfolio finding the optimal weight $w$ in this case is the same as finding the regression coefficient $w^* = \left(\omega_1, \omega_2, \cdots, \omega_{p-1}\right)^{\mathrm{T}}$.

The gross-exposure constraint $\|w\|_2 \le c$ can now be expressed as $\|w^*\|_2 \le \left(c^2 - \left(1 - 1^{\mathrm{T}} w^*\right)^2\right)^{\frac{1}{2}} = \delta(\text{say})$. Thus the problem (13) is similar to

$$\min_{b, \|w^*\|_2 \le \delta} E\left[Y - \left(w^*\right)^{\mathrm{T}} Z - b\right] \tag{14}$$

where $Z = \left(Z_1, \cdots, Z_{p-1}\right)^{\mathrm{T}}$. But they are not equivalent. The latter depends on the choice of asset $Y$, while the former does not.

The whole solution pat $w_{NC2}(c)$ to (14), for all $c \ge 0$, can be efficiently obtained by the regularization algorithm of Ridge regression (see [13]). The number of non-vanishing weights varies as c ranges from 0 to $\infty$. It recruits successively more assets and gradually all assets. The algorithm works iteratively as follows:

$$\begin{aligned} w_{NC2_i} &= w_i^* + \left(1 - \sum_{j=1}^{p-1} w_j^*\right) \times \left(w_i\right) \quad \text{where } i \in (1, 2, \cdots, p-1) \\ w_{NC2_p} &= \left(1 - \sum_{j=1}^{p-1} w_j^*\right) \times \left(w_p\right) \end{aligned} \tag{15}$$

To choose $c$ we use cross validation, as in the case of 1-norm constrained portfolio.

## 5. Practical Results

Below we provide an overview of our results of Markowitz efficient frontier, corrected frontier using Gaussian assumption, 1-norm and 2-norm constrained efficient frontiers for the 3 countries.

In **Figures 1-3**, we present the efficient frontiers using the different methods. The dashed lines represent the empirical 95% confidence intervals computed for a fixed expected return. The x-axis is variance and y-axis is

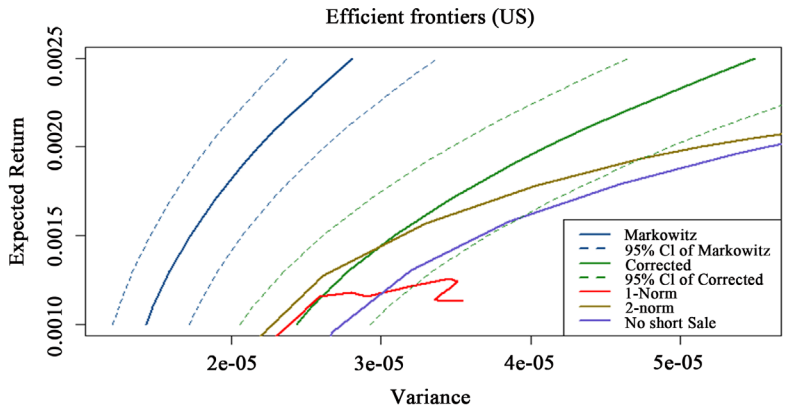


**Figure 1.** Efficient frontier of US data.

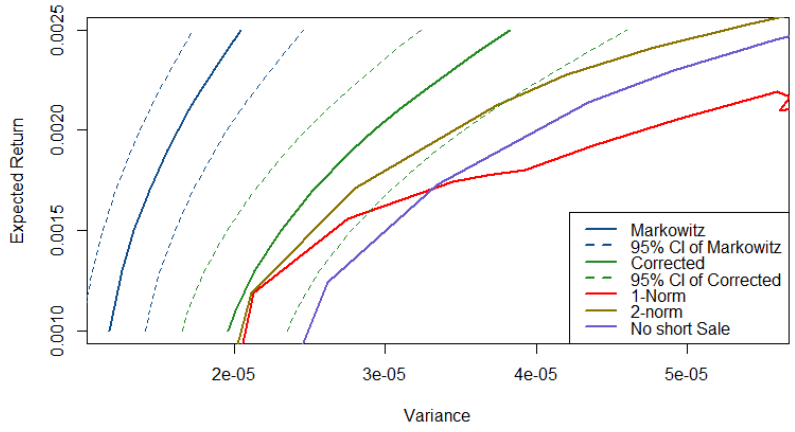


**Figure 2.** Efficient frontier of UK data.

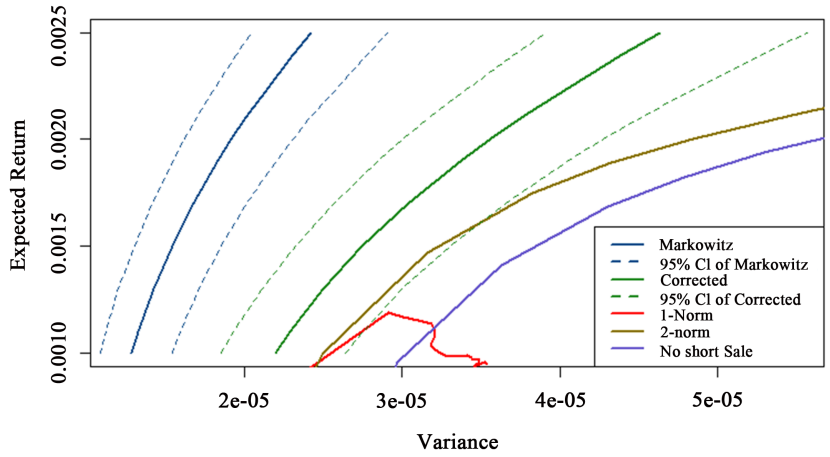


**Figure 3.** Efficient frontier of Indian data.

expected returns. We have considered the same set of $\mu$'s and $\Sigma$'s, for each individual country to keep the results comparable. It can be concluded from the relative positions of the corrected and uncorrected efficient frontiers that the risk is indeed underestimated in case of high dimensional data. But comparing to 2-norm and 1-norm constrained portfolios as they outperform the corrected frontiers. The constrained portfolios are, in general, less efficient than the corrected portfolio, in the sense that they have higher variance for each fixed level of return. Of course constrained portfolios have their own advantages due to sparsity that might out-weigh the loss in efficiency. For the 1-norm and 2-norm portfolios, the choice of the asset Y is important. We have chosen Y to be the no short sale portfolio in all our computations. For each country, the 2-norm portfolio is most efficient among the constrained portfolios and the 1-norm is not monotone.

The amount of shrinkage or regularization is directly related to the number of stocks included in the optimal portfolio. In **Figure 4** we present this for the 1-norm constrained portfolio. As expected, this is an increasing function of $c$, the bound on the $L_1$ norm. For almost all values of $c$, the number of stocks in the portfolio is highest for the Indian market and lowest for the US market. Results for the $L_2$norm are similar.

For out of sample performance we first created portfolios for all the three datasets using the return data for the first 230 trading days. These portfolios are then held for one month and rebalanced at the end next month. The summary statistics of these portfolios are presented for the three datasets as box-plots in **Figures 5-7**. 1-norm constrained portfolios were created for $c = 2$ and $c = 3$ for all the three nations. 2-norm constrained portfolios were created for the optimal c chosen by cross validation, as mentioned in Section 4. This value equals 1.2544, 1.14 and 1.0739 respectively for US, UK and Indian data.

The out-of-sample performance is very different for the three markets. For the US data, the 2-normcon-

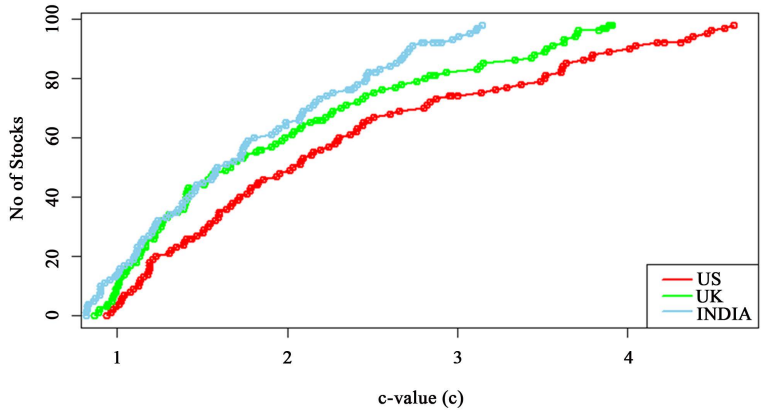


**Figure 4.** Number of stocks with respect to *c* with Y = "No short sale".

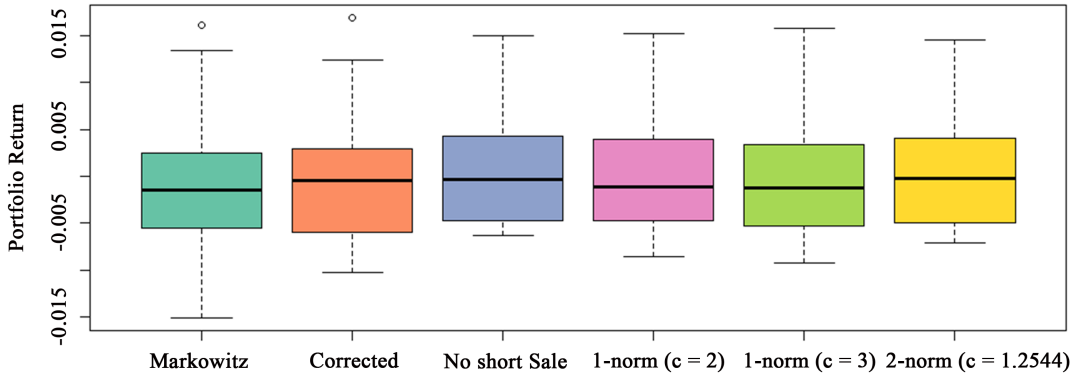


**Figure 5.** Out of sample performance of different portfolios for US data.

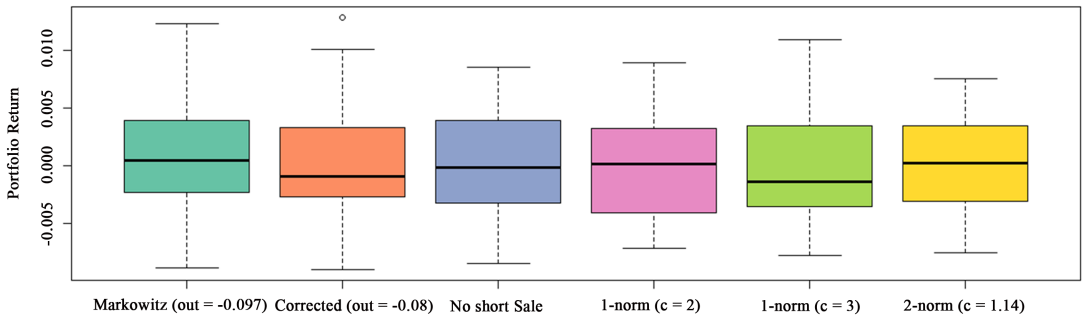


**Figure 6.** Out of sample performance of different portfolios for UK data.

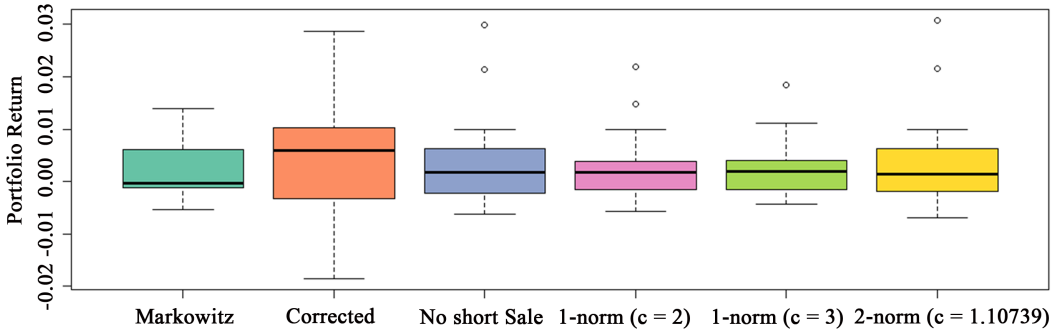


**Figure 7.** Out of sample performance of different portfolios for Indian data.

strained, corrected and no-short-sale portfolios have close to zero average returns while the other methods yield negative average returns. The variances are almost same for all methods except the Markowitz, which has a lower variance. For UK data, the 1-norm with $c = 3$ and corrected portfolios have significantly negative average return while others have small positive or zero average returns. The variances are almost all the same. For the Indian data, all portfolios except the Markowitz have high positive average returns. In particular, the corrected portfolio has very high average returns, but the variance is also quite high. Overall, from the out-of-sample results, the 2-norm constrained portfolio has higher average and comparable variance to the Markowitz portfolio in all the markets.

## 6. Conclusion

In this paper we study the effect of high dimension on the efficient frontier with real data on three markets. In particular we study how the recently suggested methods of corrected frontier based on normality assumptions and norm-constrained methods perform relative to Markowitz portfolio optimization. We observe that the Markowitz solution indeed leads to biased estimates of risk that can be improved with the corrected estimates. The norm-constrained methods are comparable and need less model assumptions. Alternative methods of improving the covariance matrix estimation are Bayesian shrinkage approach [8] or random matrix theory and principal component analysis [14]. We have ignored the time component of the data and treated the observations as i.i.d. A further improvement will be to take into account this aspect and model the high dimensional time series as in [15].

## References


[1] El Karoui, N. (2010) High-Dimensionality Effects in the Markowitz Problem and Other Quadratic Programs with Linear Constraints: Risk Underestimation. *The Annals of Statistics*, **38**, 3487-3566. http://dx.doi.org/10.1214/10-AOS795

[2] Winston, K. (1993) The Efficient Index and Prediction of Portfolio Variance. *Journal of Portfolio Management*, **19**, 27-34. http://dx.doi.org/10.3905/jpm.1993.409446

[3] Haugen, R. and Baker, N. (1991) The Efficient Market Inefficiency of Capitalization-Weighted Stock Portfolios. *Journal of Portfolio Management*, **17**, 35-40. http://dx.doi.org/10.3905/jpm.1991.409335

[4] Markowitz, H. (1952) Portfolio Selection. *The Journal of Finance*, **7**, 77-91. http://dx.doi.org/10.1111/j.1540-6261.1952.tb01525.x

[5] El Karoui, N. (2013) On the Realized Risk of High-Dimensional Markowitz portfolios. *SIAM Journal on Financial Mathematics*, **4**, 737-783. http://dx.doi.org/10.1137/090774926

[6] DeMiguel, V., Garlappi, L. and Uppal, R. (2009) Optimal versus Naive Diversification: How inefficient Is the 1/*n* Portfolio Strategy? *Review of Financial Studies*, **22**, 1915-1953. http://dx.doi.org/10.1093/rfs/hhm075

[7] Jagannathan, R. and Ma, T. (2003) Risk Reduction in Large Portfolios: Why Imposing the Wrong Constraints Helps. *Journal of Finance*, **58**, 1651-1684. http://dx.doi.org/10.1111/1540-6261.00580

[8] Ledoit, O. and Wolf, M. (2004) A Well-Conditioned Estimator for Large-Dimensional Covariance Matrices. *Journal of Multivariate Analysis*, **88**, 365-411. http://dx.doi.org/10.1016/S0047-259X(03)00096-4

[9] Fan, J., Zhang, J. and Yu, K. (2012) Vast Portfolio Selection with Gross-Exposure Constraints. *Journal of the American Statistical Association*, **55**, 798-812. http://dx.doi.org/10.1080/01621459.2012.682825

[10] DeMiguel, V., Garlappi, L., Nogales, F.J. and Uppal, R. (2009) A Generalized Approach to Portfolio Optimization: Improving Performance by Constraining Portfolio Norms. *Journal of Management Science*, **107**, 592-606.

[11] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004) Least Angle Regression (with Discussions). *The Annals of Statistics*, **32**, 409-499.

[12] Efron, B. and Gong, G. (1983) A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation. *American Statistician*, **1**, 36-48.

[13] Hoerl, A.E. and Kennard, R.W. (1970) Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, **12**, 55-67. http://dx.doi.org/10.1080/00401706.1970.10488634

[14] Bai, Z., Liu, H. and Wong, W. (2009) Enhancement of the Applicability of Markowitz's Portfolio Optimization by Utilizing Random Matrix Theory. *Mathematical Finance*, **19**, 639-667. http://dx.doi.org/10.1111/j.1467-9965.2009.00383.x

[15] Liu, H., Aue, A. and Paul, D. (2015) On the Marcenko-Pastur Law for Linear Time Series. *The Annals of Statistics*, **43**, 675-712. http://dx.doi.org/10.1214/14-AOS1294