UNIVERSITY OF CALIFORNIA, SAN DIEGO

The Shrinkage Least Absolute Deviation Estimator in Large Samples and Its Application to the Treynor-Black Model

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy
in Economics

by

Tae-Hwan Kim

Committee in Charge:

Professor Halbert White, Chair
Professor Clive Granger
Professor James Hamilton
Professor Patrick Fitzsimmons
Professor Alex Kane

1998

UMI Number: 9901433

The dissertation of Tae-Hwan Kim is approved, and it is acceptable in quality and form for publication on microfilm:

____

____

____

____

______________________________ Chair

University of California, San Diego

1998

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENT

I am deeply grateful to my advisor, Halbert White, for his insightful guidance, continual encouragement, and for teaching me the right way to do Econometric Theory. He also provided me with many opportunities to work as a research assistant, which allowed me to apply Econometric Theory to real world problems. His advice and support were invaluable in my dissertation as well as in my graduate life at UCSD.

In addition, I would like to thank Alex Kane for teaching me Finance and introducing me to various exciting research areas in Finance. I would also like to thank Clive Granger and James Hamilton for insightful comments on my papers which helped me improve my dissertation substantially. Their encouragement during my job search process is greatly appreciated. I wish to express my thanks to Patrick Fitzsimmons for teaching me Probability Theory. He helped me with many technical problems I encountered in writing my dissertation. My thanks to all other faculty members, especially Rob Engle, Allan Timmerman, Bruce Lehmann, Leonard Haff, Grahm Elliot, Wouter Den Haan, John Conlisk, Joel Sobel, and Valerie Ramey.

I wish to thank all of my classmates and friends: Imho Kang, Jasun Noh, Joshua Rosenberg, Shinichi Sakata, Andreas Gottschling, Chris Stomberg, Namwon Hyung, Gi-Hong Kim, Jin-Kyu Oh, Kangoh Yi, Changho Han, Yongil Jeon, Don McCubbin, Takashi Yamashita, Yi-Chen Lin, Phillip Lau, Antonio Mayoral, Andrew Keeling Dongheon Kim, JinSeo Cho, Daniel Mckiernan, Marvin Barth, Alejandro Gaviria, Tomoo Inoue, Anthony Webb, Andreas Heinen, and Yueting Tong. My thanks to staff members who provided helpful assistance, Michael Bacci, Sue Papp, Nancy Gilmore, Danielle Bouvier, Stacy Williams, John Rowe, and James McCaw.

I am delighted to acknowledge the support from the UCSD Project in Econometric Analysis and the Department of Economics. The financial support from the Rotary International Foundation is greatly appreciated. The Rotary Ambassadorial Scholarship program also provided me with opportunities to make nice friends including Grace and Charles Guderjahn, and David Cupp. Many thanks to them.

I would like to express my sincere thanks to a special friend, Margaret Knetzer, who was an English teacher for me and my wife for many years and now is my daughter's best friend and teacher. Her help was invaluable.

I am deeply grateful to my parents, Byung-Chan Kim and Hwa-Mo Kim, who have always encouraged me to purse my academic ideals. I also like to thank my sisters, Yun-Hi Kim and Yung-Jeong Kim, for their support.

Finally I would like to thank Yung-Sook Lee who has been, and will be, the best friend and colleague in my life and Ji-Young Kim who made me the happiest father. This dissertation is dedicated to them.

VITA

July 3, 1965	Born, Seoul, Korea.
1989	B.A. in Business Administration, Yonsei University.
1991	M.A. in Economics, Yonsei University.
1993	Research Assistant, The Bank of Korea
1994-1998	Teaching Assistant, Department of Economics, University of California, San Diego.
1995-1998	Research Assistant, Department of Economics, University of California, San Diego.
1998	Doctor of Philosophy, University of California, San Diego.

FIELDS OF STUDY

Major Field: Economics

Studies in Econometrics
Professors Halbert White, Clive Granger, Robert Engle and James Hamilton

Studies in Probability Theory
Professor Patrick Fitzsimmons

Studies in Finance
Professors Alex Kane, Allan Timmerman and Bruce Lehmann

ABSTRACT OF THE DISSERTATION

The Shrinkage Least Absolute Deviation Estimator in Large Samples
and Its Application to the Treynor-Black Model

by

Tae-Hwan Kim
Doctor of Philosophy in Economics
University of California, San Diego, 1998
Professor Halbert White, Chair

The dissertation explores the extension of the James-Stein estimator in a direction that enables it to preserve its superiority when the sample size goes to infinity. The first chapter develops the theoretical foundation for the extension. Instead of shrinking a base estimator towards a fixed point, we shrink towards a data-dependent point, which makes it possible that the prior becomes more accurate as the sample size grows. We prove that the extended James-Stein estimator shrunk towards a data-dependent point has smaller asymptotic risk than the base estimator. It turns out that shrinking an estimator toward a data-dependent point is equivalent to combining two random variables using the James-Stein rule. We propose a general combination scheme which includes random combination and non-random combination as special cases. The result allows us to apply the extended James-Stein estimator to robust regression, especially to the Least Absolute Deviations Estimator. We show analytically and by simulation that if we shrink the LAD Estimator, then we have smaller risk.

The second chapter provides a way to obtain the sampling distributions and confidence intervals for the James-Stein type combination estimator using a bootstrapping method. It is well known that in order to get a better bootstrap confidence interval, one should use

pivotal or asymptotically pivotal statistics. We use Ullah (1990)'s results to derive the first moment and the second moment. We use a consistent estimator for this asymptotic variance to obtain the bootstrapping pivotal statistics.

The third applies shrinkage estimation to the construction of optimal portfolios as proposed by Treynor and Black (1973) using alpha and beta forecast data from a financial institution. The Treynor Black model provides a method to exploit security analysis; however, its success depends critically on both the predictive ability of abnormal return forecasts and the conversion of this predictive power into the portfolio construction. We use the OLS estimator, the LAD estimator and shrinkage LAD estimators to extract predictive ability from raw alpha forecasts. The estimated correlation between ex-post abnormal returns and alpha forecasts is as low as 0.04, yet out-of-sample experiments show that the use of robust estimators can yield superior portfolios for the forecast database.

Chapter 1

The James-Stein Estimator in Large Sample and its Application to Robust Regression: LAD Estimation

1.1 Introduction

Shrinkage techniques for linear regression model have been studied extensively since the seminal works by Stein (1955), James and Stein (1960) and Hoerl and Kennard (1970). Stein (1955) and James and Stein (1960) prove that the usual estimator for the mean of multivariate normal distribution is inadmissible[1] and there exists an improved estimator with smaller risk when the dimension of the multivariate normal vector is greater than two. On the other hand Hoerl and Kennard (1970) show that there exists a ridge parameter value such that the risk of the ridge estimator is smaller than the usual estimator. These two areas developed independently have a common basic idea to shrink the usual estimator to reduce its variance making the estimator biased.

Even though there are many papers in these areas, most of them have focused on the improvement over the Ordinary Least Squares Estimator (OLS) and little attention has been paid to the application to robust estimations. Askin and Montgomery (1980) apply shrinkage techniques such as ridge, Stein shrinkage and principal components to M-estimation to stabilize the estimator and residuals in the presence of multicollinearity and non-normality. In a series of papers, Saleh and Sen (1985, 1987) apply two shrinkage techniques to the M-estimator: a preliminary test shrinkage version and a special type of Stein shrinkage version. They show analytically that the classical M-estimator, the preliminary test shrinkage version of the M-estimator, and the Stein shrinkage version of the M-estimator are all asymptotically risk-equivalent in that their finite sample risks are converging to the same limit when the number of observations converges to infinity. Schmoyer and Arnold (1989) show that the shrinkage M-estimator asymptotically dominates the M-estimator as long as a special non-random guess sequence about the true parameters is available.

---

[1] Suppose we have two estimators for the parameter $\theta \in \Theta$. Let $R_{Ti}(\theta)$ be the ith estimator's risk. Then T_1 is defined to be "better than" T_2 if and only if (1) $R_{T1}(\theta) \leq R_{T2}(\theta)$ for all θ in Θ and (2) $R_{T1}(\theta) < R_{T2}(\theta)$ for at least one θ in Θ. An estimator T is defined to be admissible if and only if there is no better estimator.

We generalize the approach taken by Schmoyer and Arnold (1989) one step further. The key point in their approach is that they control the convergence rate of the shrinkage estimator by using a non-random guess sequence so that they can apply the James-Stein rule in the limit. However there is a strong assumption imposed on the non-random guess sequence, which makes this result difficult to apply in practice. We relax this assumption by shrinking a base estimator toward a data-dependent point (random guess). We show that the James-Stein type estimator shrunk toward a data-dependent point has a smaller asymptotic risk than the base estimator. Interestingly it turns out that using a random guess is basically equivalent to combining two estimators by using the Stein random combination weight. Most studies have been focused on either non-random combination weight or random combination weight. See Arthanar and Dodge (1981), Cohen (1976) and Green and Strawderman (1991), Laplace (1818) and Cahn (1994). Accordingly we propose the Optimal Weighting Scheme (OWS) estimator which includes both random and non-random combination as special cases.

1.2 Asymptotic Risk Improvement

Consider $y_t = x_t'\beta^0 + \varepsilon_t$ $t = 1, 2, \ldots, n$ where $\beta^0 \in R^k$ and ε_t is assumed to be identical and independent. We define $X^n = [x_1, x_2, \ldots, x_n]'$. Let b_n be an estimator for β^0. A function $L(b_n,\beta^0)$ is called the loss function if and only if (1) $L(b_n,\beta^0) \geq 0$ for all b_n and all β^0 and (2) $L(b_n,\beta^0) = 0$ if and only if $b_n = \beta^0$. The expectation of the loss function $E(L(b_n,\beta^0))$ is called the risk denoted by $R(b_n,\beta^0)$. An example of a loss function is the quadratic loss, $L(b_n,\beta^0) = (b_n-\beta^0)'Q_n(b-\beta^0)$ where Q_n is a symmetric and positive definite matrix. Let $\{b_n\}$ be a sequence of estimators of $\beta^0$and let $\{L(b_n,\beta^0)\}$ be a sequence of loss values. Suppose $L(b_n,\beta^0)$ converges to an integrable random variable Ψ in distribution. The asymptotic risk of $\{b_n\}$ for $\{L(b_n,\beta^0)\}$ is then defined by

$$AR(\{b_n\}, \beta^0) = E(\Psi).$$

Schmoyer and Arnold (1989) introduce a sequence of non-random guesses, $\{g_n\}$, where g_n is our guess about the true parameter, β^0. A general version of the James-Stein shrinkage estimator, $\delta(b_n,g_n)$, is then defined by

$$\delta(b_n,g_n) = n^{-1/2}K(U_n, n^{-1}Qn) + g_n$$

where $U_n = n^{1/2}(b_n-g_n)$ and Q_n is a known matrix. For example, if

$$K(U_n,n^{-1}Q_n) = \left(1-\frac{k-2}{U_n' n^{-1} Q_n U_n}\right)U_n \text{ and}$$

$$Q_n = \hat{\sigma}_n^2 X^{n\prime}X^n,$$

then $\delta(b_n,g_n) = \left(1-\frac{(k-2)}{(b_n-g_n)'Q_n(b_n-g_n)}\right)(b_n - g_n) + g_n$ which is the usual JS estimator for $g_n = 0$.[2]

So far we have considered only a quadratic loss function. Let us define a general loss function, $L(b_n,\beta^0) = h(n^{1/2}(b_n-\beta^0), n^{-1}Q_n)$ where Q_n is a known matrix. The quadratic loss function is a special case where $h(n^{1/2}(b_n-\beta^0), n^{-1}Q_n) = n^{1/2}(b_n-\beta^0)'n^{-1}Q_n n^{1/2}(b_n-\beta^0)$. By assuming that (1) $n^{1/2}(g_n-\beta^0) \to \theta$ where θ is a finite fixed vector (2) $n^{-1}Q_n \xrightarrow{p} Q$ where Q is a nonstochastic symmetric and positive definite matrix and (3) $h(a,b) = h(n^{1/2}a, n^{-1}b)$, Schmoyer and Arnold (1989) have proved that

(1) $n^{1/2}(\delta(b_n,g_n)-\beta^0) \xrightarrow{d} K(U,Q)+\theta$ where $n^{1/2}(b_n-g_n) \xrightarrow{d} U$

(2) $AR(\{\delta(b_n,g_n)\},\beta^0) = E\{L(K(U,Q), -\theta)\}$

(3) $AR(\{\delta(b_n,g_n)\},\beta^0) < AR(\{b_n\},\beta^0)$ if $\lambda \in (0, 2(k-2))$

provided that $AR(\{\delta(b_n,g_n)\},\beta^0) < \infty$ and $AR(\{b_n\},\beta^0) < \infty$. This result is interesting in the following senses. Firstly, The limiting random variable of the shrinkage estimator is not the usual normal random variable but a non-linear function of the normal random

---

[2] Note that $\delta(b_n,g_n) = (1 - w(k,L))(b_n-g_n) + g_n$ where $w(k,L) = (k-2)/L(b_n, g_n)$. The positive James-Stein shrinkage estimator is defined as $\delta(b_n,g_n) = (1 - w(k,L))^+(b_n-g_n) + g_n$ which is the proper convex combination of the estimator and the guess. The weight is increasing with k and decreasing with L. Since L is a weighted distance between the estimator and the guess, we have the following interpretation of this kind of shrinkage estimator. If the distance between the estimator and the guess is big, then this shrinkage scheme gives large weight to the estimator and small weight to the guess. If the distance is small, the other story is true.

variable. Secondly, we can achieve risk improvement even when the sample size approaches infinity.

It is well known that when $g_n = 0$ for all n, then $AR(\{\delta(b_n,g_n)\},\beta^0)$ is same as $AR(\{b_n^{LS}\},\beta^0)$; we have no improvement in large sample. The condition that $n^{1/2}(g_n-\beta^0) \rightarrow \theta$ plays the key role in obtaining the asymptotic risk improvement. However, the guess process is not random and must converge to the unknown truth β^0 with the rate $O(n^{-1/2})$. In other words, we have to have a very good idea about the DGP without looking at the data generated from the Data Generating Process (DGP). This assumption is not practical and it needs to be relaxed.

Accordingly we consider two alternatives to Schmoyer and Arnold. One is a changing parameter, non-random guess approach. The other is a fixed parameter, random guess approach. We consider the changing parameter, non-random guess approach first. Suppose we observe y_{nt} which is generated by $y_{nt} = X_{nt}\beta_n^0 + \varepsilon_{nt}$ where the true parameter process $\{\beta_n^0\}$, though changing with n, converges to a limit β^0 as n converges to infinity. Here n is the size of the sample and t is the index of an observation. In this case the truth depends on the size of the sample we draw from the population. This model is relevant if the population distribution depends on the sampling. We state some assumptions.

Assumption 2-1 $n^{1/2}(b_n-\beta_n^0) \xrightarrow{d} N(0,A)$ where A is a nonstochastic, symmetric and positive definite matrix.

Assumption 2-2 $n^{1/2}(g^*-\beta_n^0) \rightarrow \theta$ where θ is a fixed vector.

Assumption 2-3 $n^{-1}Q_n \xrightarrow{p} Q$ where Q is a nonstochastic symmetric and positive definite matrix.

Corollary 2-1 Asymptotic Risk with changing parameter and non-random guess

Suppose that Assumption 2-1, 2-2 and 2-3 hold. We define a loss function $L(b_n,\beta_n^0) = h(n^{1/2}(b_n-\beta_n^0), n^{-1}Q_n)$ where h is continuous in both arguments and it satisfies $h((b_n-\beta_n^0), Q_n) = h(n^{1/2}(b_n-\beta_n^0), n^{-1}Q_n)$. The extended James-Stein type estimator is defined by $\delta(b_n,g^*) = n^{-1/2}K(U_n, n^{-1}Q_n) + g^*$ where $U_n = n^{1/2}(b_n-g^*)$ and K is continuous in both arguments. Then

(1) $U_n \xrightarrow{d} U \sim N(-\theta,A)$.

(2) $n^{1/2}(\delta(b_n,g^*)-\beta_n^0) \xrightarrow{d} K(U,Q)+\theta$.

(3) $L(\delta(b_n,g^*),\beta_n^0) \xrightarrow{d} L(K(U,Q), -\theta)$.

(4) $AR(\{\delta(b_n,g^*)\},\beta_n^0) = E\{L(K(U,Q), -\theta)\}$ provided the expectation exists.

Proof: The proof is straightforward and omitted.

The other alternative is to randomize the guess process so that it depends on the data. A necessary condition for this random guess is

$$n^{1/2}(g_n-\beta^0) \xrightarrow{d} N(\theta,B)$$

where B is a nonstochastic matrix. However it turns out that this assumption is not sufficient because even though the limiting marginal distributions of $n^{1/2}(g_n-\beta^0)$ and $n^{1/2}(b_n-\beta_n^0)$ may be normal, the joint distribution could be another distribution. In order to avoid this problem we impose a joint normality condition.

Assumption 2-4 Joint Normality Condition

$$\begin{bmatrix} n^{1/2}(b_n-\beta^0) \\ n^{1/2}(g_n-\beta^0) \end{bmatrix} \xrightarrow{d} \begin{bmatrix} U_1 \\ U_2 \end{bmatrix} \sim N\left(\begin{bmatrix} 0_{k\times 1} \\ \theta_{k\times 1} \end{bmatrix}, \begin{bmatrix} A_{k\times k} & \Delta_{k\times k} \\ \Delta'_{k\times k} & B_{k\times k} \end{bmatrix} \right)$$

where θ, A, B, and Δ are bounded..

Theorem 2-2 Asymptotic Risk with fixed parameter and changing guess

Suppose that Assumption 2-3 and 2-4 hold. We define a loss function $L(b_n,\beta^0) = h(n^{1/2}(b_n-\beta^0), n^{-1}Q_n)$ where h is continuous in both arguments and it satisfies $h((b_n-\beta_n^0),$

$Q_n) = h(n^{1/2}(b_n-\beta_n^0), n^{-1}Q_n)$. The extended James-Stein type estimator is defined by $\delta(b_n,g_n) = n^{-1/2}K(U_n, n^{-1}Q_n) + g_n$ where $U_n = n^{1/2}(b_n-g^*)$ and K is continuous in both arguments. Then

(1) $U_n \xrightarrow{d} U \equiv U_1 - U_2$.

(2) $n^{1/2}(\delta(b_n,g_n)-\beta^0) \xrightarrow{d} K(U,Q)+U_2$.

(3) $L(\delta(b_n,g_n),\beta^0) \xrightarrow{d} L(K(U,Q),-U_2)$.

(4) $AR(\{\delta(b_n,g_n)\},\beta^0) = E\{L(K(U,Q), -U_2)\}$ provided the expectation exists.

Proof: See Appendix.

This result shows that we obtain the basically same formula using the random guess. This result is interesting in the following senses. First, it provides us with a way of choosing a finite sample random guess. It allows us to choose another estimator as our guess since many econometric estimators satisfy the asymptotic normality condition. For example, if we want a shrinkage transformation of the OLS estimator, one option is to shrink the OLS estimator towards the LAD estimator instead of zero if the LAD estimator satisfies the joint distribution assumption.[3] Secondly, this random guess approach provides us with a way of combining two estimators. In fact if we use another estimator as our guess, then the shrinkage transformation is equivalent to using a convex combination of those two estimators which we call the James-Stien Combination (JSC) Estimator. In the following we prove that we are still able to make an improvement on the base estimator even though we use the random guess.

Assumption 2-5 $\text{Prob}[b_n \neq g_n] = 1$ for all n.

Assumption 2-6 $\text{Prob}[U_1 \neq U_2] = 1$.

[3] One can show that the LAD estimator and the OLS estimator are converging to a joint normal distribution under some regularity condition using the Bahadur representation (She, P. (1992), Bates and White (1993)).

Definition 2-1 James-Stein Combination Estimator

The combination of two estimators using the James-Stein rule,

$$\delta_\lambda^{JS}(b_n,g_n) = \left(1-\frac{\lambda}{(b_n-g_n)'Q_n(b_n-g_n)}\right)(b_n - g_n) + g_n$$

where λ is a constant, is called the James-Stein Combination (JSC) Estimator[4].

Corollary 2-3 Suppose that Assumption 2-3, 2-4, 2-5 and 2-6 hold. Define

$K(U,Q) = \left(1-\frac{\lambda}{U'QU}\right)U$. Then

(1) $n^{1/2}(\delta_\lambda^{JS}(b_n,g_n)-\beta^0) \xrightarrow{d} \delta_\lambda^{JS}(U_1,U_2)$.

(2) $AR(\{\delta_\lambda^{JS}(b_n,g_n)\},\beta^0) = R(\delta_\lambda^{JS}(U_1,U_2),0)$ provided the expectation exists.

Proof: The proof is straightforward and omitted.

Theorem 2-4 Suppose that Assumption 2-3, 2-4, 2-5 and 2-6 hold. Then

(1) $AR(\{\delta_\lambda^{JS}(b_n,g_n)\},\beta^0)$ is strictly convex in λ.

(2) Let $\lambda^* \in \text{argmin}\, AR(\{\delta_\lambda^{JS}(b_n,g_n)\},\beta^0)$. Then $\lambda^* = \nu/\omega$ where

$$\nu = E\left[\frac{U_1'Q(U_1-U_2)}{(U_1-U_2)'Q(U_1-U_2)}\right] \text{ and } \omega = E\left[\frac{1}{(U_1-U_2)'Q(U_1-U_2)}\right].$$

(3) $AR(\{\delta_{\lambda^*}^{JS}(b_n,g_n)\},\beta^0) = -\nu^2/\omega + \kappa$ where $\kappa = AR(\{b_n\},\beta^0) = E[U_1'QU_1]$.

(4) $AR(\{\delta_{\lambda^*}^{JS}(b_n,g_n)\},\beta^0) \leq AR(\{b_n\},\beta^0)$ where the equality holds only when $\nu = 0$.

(5) $AR(\{\delta_\lambda^{JS}(b_n,g_n)\},\beta^0) \leq AR(\{b_n\},\beta^0)$ if $\lambda \in [\min\{0,2\nu/\omega\}, \max\{0,2\nu/\omega\}]$ where the equality holds only when $\nu = 0$.

Proof: See Appendix.

As long as $\nu \neq 0$, which we call the 'Relative Non-Efficiency Condition (RNEC)' for the base estimator[5], we can achieve an asymptotic risk improvement with respect to the base

[4] Note that the JSC estimator is well-defined by Assumption 2-5.

estimator by choosing the shrinkage factor λ properly. One example where $\nu = 0$ is that the base estimator is asymptotically efficient. While the sign of ω is positive by Assumption 2-6, the sign of ν is not determined. Therefore the sign of λ^* depends on whether $E\left[\frac{U_1'Q(U_1-U_2)}{(U_1-U_2)'Q(U_1-U_2)}\right]$ is negative or positive. It is interesting that the ratio ν/ω is equal to k-2 when U_1 and U_2 are independent; i.e. $\Delta = 0$. Therefore in this case, the optimal combination weight is exactly equal to the James-Stein optimal weight. The deviation of the ratio ν/ω from k-2 depends on the degree of the correlation between the base estimator and the guess estimator.

Example 2-1 Suppose that $A = \sigma^2 I$, $B = \tau^2 I$, $\Delta = 0$, $Q = I$ and $k \geq 3$. Then

(1) $\nu = (k-2)\sigma^2\omega$.

(2) $\lambda^* = (k-2)\sigma^2$.

(3) $AR(\{\delta_{\lambda^*}^{JS}(b_n,g_n)\},\beta^0) = -(k-2)^2\sigma^4/\omega + \sigma^2 k$.

Example 2-2 Suppose that $A = \sigma^2 I$, $B = \tau^2 I$, $\Delta = 0$, $Q = I$, $\theta \neq 0$ and $k = 4$. Then

(1) $\omega = 1/[(\sigma^2 + \tau^2)\theta'\theta]$.

(2) $AR(\{\delta_{\lambda^*}^{JS}(b_n,g_n)\},\beta^0) = -4\sigma^4/[(\sigma^2 + \tau^2)\theta'\theta] + 4\sigma^2$.

Example 2-3 Suppose that $A = \sigma^2 I$, $B = \tau^2 I$, $\Delta = 0$, $Q = I$, $\theta = 0$ and $k \geq 3$. Then

(1) $\omega = 1/[(k-2)(\sigma^2 + \tau^2)]$

---

[5] The relative non-efficiency condition does not allow us to choose any asymptotically efficient estimator as the base estimator (b_n) unless we select a "super-efficient estimator" as our guess (g_n) if we want to reduce asymptotic risk. If we know the error distribution is normal, the Maximum Likelihood Estimator (MLE) is asymptotically efficient and has the minimum asymptotic risk. There is no need to shrink the MLE to minimize the asymptotic risk. For an example of a "super-efficient estimator" and some discussion of asymptotic efficiency, see e.g. White (1994, pp 133). Suppose we choose an asymptotically efficient estimator as the base estimator. Then, any guess estimator, if not super-efficient, can be expressed as the sum of the asymptotically efficient estimator and a random noise, which converges to zero as n goes to infinity and is asymptotically uncorrelated with the asymptotically efficient estimator. Therefore, $Cov(U_1,U_1-U_2) = 0$ which is equivalent to U_1 being independent of U_1-U_2. This implies that $\nu = 0$. The condition, $\nu \neq 0$, also guarantees that $Prob[U_1 = U_2] < 1$.

(2) $AR(\{\delta_{\lambda^*}^{JS}(b_n,g_n)\},\beta^0) = k\sigma^2\tau^2/(\sigma^2+\tau^2) + 2\sigma^4/(\sigma^2+\tau^2)$.

Example 2-4 Suppose that $A = \sigma^2 I$, $B = \tau^2 I$, $\Delta = 0$, $Q = I$ and $k \geq 3$. Then

$AR(\{\delta_{\lambda^*}^{JS}(b_n,g_n)\},\beta^0) \leq k\sigma^2 - (k-2)^2\sigma^4/[\theta'\theta+(k-2)(\sigma^2+\tau^2)]$.

While the JSC estimator is a combination of two estimators using a random weight, conventional combined estimators use non-random weight. This non-random combination has been studied mainly for independent estimators. See Cohen (1976) and Green and Strawderman (1991). Laplace (1818), Cahn (1994) consider combining correlated estimators, but they analyze the one dimensional case and obtain the optimal weight by minimizing (asymptotic) variance. We consider combining multi-dimensional correlated estimators by minimizing asymptotic risk. We define the Non-Random Combination (NRC) Estimator formally.

Definition 2-2 Non-random Combination Estimator

The combination of two estimators using a non-random weight,

$$\delta_\lambda^{NR}(b_n,g_n) = (1-\lambda)(b_n - g_n) + g_n$$

where λ is a constant, is called the Non-Random Combination (NRC) Estimator.

Corollary 2-5 Suppose that Assumption 2-3 and 2-4 hold. Then

(1) $n^{1/2}(\delta_\lambda^{NR}(b_n,g_n)-\beta^0) \xrightarrow{d} \delta_\lambda^{NR}(U_1,U_2)$.

(2) $AR(\{\delta_\lambda^{NR}(b_n,g_n)\},\beta^0) = R(\delta_\lambda^{NR}(U_1,U_2),0)$ provided the expectation exists.

Proof: The proof is straightforward and omitted.

Corollary 2-6 Suppose that Assumption 2-3, 2-4, 2-5 and 2-6 hold. Then

(1) $AR(\{\delta_\lambda^{NR}(b_n,g_n)\},\beta^0)$ is strictly convex in λ.

(2) Let $\lambda^* \in \text{argmin } AR(\{\delta_\lambda^{NR}(b_n,g_n)\},\beta^0)$. Then $\lambda^* = \beta/\alpha$ where

$\alpha = E[(U_1 - U_2)'Q(U_1 - U_2)]$ and $\beta = E[U_1'Q(U_1 - U_2)]$.

(3) $AR(\{\delta_{\lambda*}^{NR}(b_n,g_n)\},\beta^0) = -\beta^2/\alpha + \kappa$ where $\kappa = AR(\{b_n\},\beta^0) = E[U_1'QU_1]$.

(4) $AR(\{\delta_{\lambda*}^{NR}(b_n,g_n)\},\beta^0) \leq AR(\{b_n\},\beta^0)$ where the equality holds only when $\beta = 0$.

(5) $AR(\{\delta_{\lambda}^{NR}(b_n,g_n)\},\beta^0) \leq AR(\{b_n\},\beta^0)$ if $\lambda \in [\min\{0,2\beta/\alpha\}, \max\{0,2\beta/\alpha\}]$ where the equality holds only when $\beta = 0$.

(6) $AR(\{\delta_{\lambda*}^{NR}(b_n,g_n)\},\beta^0) \leq AR(\{g_n\},\beta^0)$ where the equality holds only when $\gamma = 0$ with $\gamma = E[(U_1 - U_2)'QU_2]$.

Proof: See Appendix.

Therefore if both $\beta \neq 0$ and $\gamma \neq 0$, then the asymptotic risk of the NRC estimator is strictly smaller than both the base estimator and the guess estimator when the optimal weight is chosen.

Example 2-5 Suppose that $A = \sigma^2 I$, $B = \tau^2 I$, $\Delta = 0$ and $Q = I$. Then

(1) $\alpha = \theta'\theta + k(\sigma^2 + \tau^2)$.

(2) $\beta = k\sigma^2$

(3) $\lambda^* = k\sigma^2/[\theta'\theta + k(\sigma^2 + \tau^2)]$.

(3) $AR(\{\delta_{\lambda*}^{NR}(b_n,g_n)\},\beta^0) = k\sigma^2 - k^2\sigma^4/[\theta'\theta + k(\sigma^2 + \tau^2)]$.

Example 2-6 Suppose that $A = \sigma^2 I$, $B = \tau^2 I$, $\Delta = 0$, $Q = I$ and $\theta = 0$. Then

(1) $\lambda^* = \sigma^2/(\sigma^2 + \tau^2)$.

(3) $AR(\{\delta_{\lambda*}^{NR}(b_n,g_n)\},\beta^0) = k\sigma^2\tau^2/(\sigma^2 + \tau^2)$.

By comparing Example 2-3 and Example 2-6, we can conclude that $AR(\{\delta_{\lambda*}^{NR}(b_n,g_n)\},\beta^0) < AR(\{\delta_{\lambda*}^{JS}(b_n,g_n)\},\beta^0)$ when $A = \sigma^2 I$, $B = \tau^2 I$, $\Delta = 0$, $Q = I$ and $\theta = 0$. The following example provides a condition under which $AR(\{\delta_{\lambda*}^{NR}(b_n,g_n)\},\beta^0) > AR(\{\delta_{\lambda*}^{JS}(b_n,g_n)\},\beta^0)$.

Example 2-7 Suppose that $A = \sigma^2 I$, $B = \tau^2 I$, $\Delta = 0$, $Q = I$, $k \geq 3$ and $\theta \neq 0$. Then

$$AR(\{\delta_{\lambda^*}^{NR}(b_n,g_n)\},\beta^0) > AR(\{\delta_{\lambda^*}^{JS}(b_n,g_n)\},\beta^0) \Leftrightarrow \omega < \left(\frac{k}{k-2}\right)^2\left[\frac{1}{\theta'\theta + k(\sigma^2+\tau^2)}\right].$$

Example 2-8 Suppose that $A = \sigma^2 I$, $B = \tau^2 I$, $\Delta = 0$, $Q = I$, $k = 4$ and $\theta \neq 0$. Then

$$AR(\{\delta_{\lambda^*}^{NR}(b_n,g_n)\},\beta^0) > AR(\{\delta_{\lambda^*}^{JS}(b_n,g_n)\},\beta^0) \Leftrightarrow 4(\sigma^2+\tau^2) < \theta'\theta[4(\sigma^2+\tau^2)-1].$$

We now propose a general combination scheme, which includes both the JSC estimator and the NRC estimator as special cases.

Assumption 2-7 $(U_1-U_2)'Q(U_1-U_2)$ is nondegenerate.[6]

Definition 2-3 Optimal Weighting Scheme Estimator

The combination of two estimators defined by

$$\delta_\lambda^{OW}(b_n,g_n) = \left(1-\lambda_1-\frac{\lambda_2}{(b_n-g_n)'Q_n(b_n-g_n)}\right)(b_n - g_n) + g_n$$

where $\lambda = [\lambda_1\ \lambda_2]'$, is called the Optimal Weighting Scheme (OWS) Estimator.

Corollary 2-7 Suppose that Assumption 2-3, 2-4, 2-5 and 2-6 hold. Then

(1) $n^{1/2}(\delta_\lambda^{OW}(b_n,g_n)-\beta^0) \xrightarrow{d} \delta_\lambda^{OW}(U_1,U_2)$.

(2) $AR(\{\delta_\lambda^{OW}(b_n,g_n)\},\beta^0) = R(\delta_\lambda^{OW}(U_1,U_2),0)$ provided the expectation exists.

Proof: The proof is straightforward and omitted.

The following theorem provides the optimal combination weight for the OWS estimator and some conditions under which the asymptotic risk of the OWS estimator is smaller than the asymptotic risk of the base estimator.

---

[6] A random variable X is called degenerate if there exists a constant c such that Prob[X=c] = 1.

Theorem 2-8 Suppose that Assumption 2-3, 2-4, 2-5, 2-6 and 2-7 hold. Then

(1) $AR(\{\delta_\lambda^{OW}(b_n,g_n)\},\beta^0)$ is strictly convex in λ.

(2) Let $\lambda^* \in \text{argmin } AR(\{\delta_\lambda^{NR}(b_n,g_n)\},\beta^0)$. Then

$\lambda_1^* = (\alpha\omega-1)^{-1}(\beta\omega-\nu)$ and $\lambda_2^* = (\alpha\omega-1)^{-1}(\alpha\nu-\beta)$.

(3) $AR(\{\delta_{\lambda^*}^{OW}(b_n,g_n)\},\beta^0) = (\alpha\omega-1)^{-2}[-\alpha\beta^2\omega^2-(2\alpha\beta\nu-\alpha^2\nu^2+\beta^2)\omega+(\alpha\nu^2-2\beta\nu)] + \kappa$.

(4) $AR(\{\delta_{\lambda^*}^{OW}(b_n,g_n)\},\beta^0) \leq AR(\{b_n\},\beta^0)$ where the equality holds only when $\beta = 0$ and $\nu = 0$.

Proof: See Appendix.

We call $(\lambda_1^*, \lambda_2^*)$ optimal weighting shrinkage factors and $(\alpha,\beta,\nu,\omega)$ combination control parameters.

In the following we identify some properties of optimal weighting shrinkage factors and combination control parameters.

Corollary 2-9 Suppose that Assumption 2-3, 2-4, 2-5, 2-6 and 2-7 hold. Then

(1) $\alpha\omega - 1 > 0$.

(2) $$\lambda_1^* \geq 0 \Leftrightarrow \text{Cov}\left[U_1'Q(U_1-U_2), \frac{1}{(U_1-U_2)'Q(U_1-U_2)}\right] \leq 0.$$

$$\lambda_1^* < 0 \Leftrightarrow \text{Cov}\left[U_1'Q(U_1-U_2), \frac{1}{(U_1-U_2)'Q(U_1-U_2)}\right] > 0.$$

(3) $$\lambda_2^* \geq 0 \Leftrightarrow \text{Cov}\left[(U_1-U_2)'Q(U_1-U_2), \frac{U_1'Q(U_1-U_2)}{(U_1-U_2)'Q(U_1-U_2)}\right] \leq 0.$$

$$\lambda_2^* < 0 \Leftrightarrow \text{Cov}\left[(U_1-U_2)'Q(U_1-U_2), \frac{U_1'Q(U_1-U_2)}{(U_1-U_2)'Q(U_1-U_2)}\right] > 0.$$

(4) Suppose that b_n is asymptotically efficient and g_n is consistent and not super-efficient. Then

$\lambda_1^* = 0$ and $\lambda_2^* = 0$.

Proof: See Appendix.

We now prove that the OWS estimator has no larger asymptotic risk than both the JSC estimator and the NRC estimator.

Corollary 2-10 Suppose that Assumption 2-3, 2-4, 2-5, 2-6 and 2-7 hold. Then

(1) $AR(\{\delta_{\lambda^*}^{OW}(b_n,g_n)\},\beta^0) \leq AR(\{\delta_{\lambda^*}^{JS}(b_n,g_n)\},\beta^0)$ where the strict inequality holds if λ_1^* is not equal to zero.

(2) $AR(\{\delta_{\lambda^*}^{OW}(b_n,g_n)\},\beta^0) \leq AR(\{\delta_{\lambda^*}^{NR}(b_n,g_n)\},\beta^0)$ where the strict inequality holds if λ_2^* is not equal to zero.

Proof: See Appendix.

Example 2-9 Suppose that $A = \sigma^2 I$, $B = \tau^2 I$, $\Delta = 0$, $Q = I$, $k \geq 3$ and $\theta = 0$. Then

(1) $\lambda_1^* = \sigma^2/(\sigma^2 + \tau^2)$.

(2) $\lambda_2^* = 0$.

Example 2-9 is interesting in that it tells us that the NRC estimator is optimal and the JS random combination part does not make any contribution when $A = \sigma^2 I$, $B = \tau^2 I$, $\Delta = 0$, $Q = I$, $k \geq 3$ and $\theta = 0$; especially when there is no asymptotic bias[7]. The following example provides a condition under which the JS random part makes contribution to an asymptotic risk reduction.

Example 2-10 Suppose that $A = \sigma^2 I$, $B = \tau^2 I$, $\Delta = 0$, $Q = I$, $k = 4$ and $\theta \neq 0$. Then

[7] Since we are in particular interested in combining the OLS estimator and the LAD estimator in which case θ is most likely equal to zero, it is interesting to see whether λ_2^* is still equal to zero when A and B are general covariance matrices and Δ is non-zero. The necessary and sufficient condition is given in Corollary 2-9 as

$$Cov(X,Y) = 0$$

where $X = (U_1-U_2)'Q(U_1-U_2)$ and $Y = U_1'Q(U_1-U_2)/(U_1-U_2)'Q(U_1-U_2)$. It is not easy to verify this condition analytically, but it is possible to estimate the covariance and test the null hypothesis that the covariance is equal to zero. According to the simulation study carried out in Appendix, there is some evidence that the covariance is close to zero, but not equal to zero. See Appendix for detailed discussion.

$\lambda_2^* = 0 \Leftrightarrow \theta'\theta = 4(\sigma^2+\tau^2)/[2(\sigma^2+\tau^2)-1]$.

Even though the OWS estimator has some nice properties, the OWS estimator cannot be estimated directly because it contains 4 unknown parameters; α, β, ν, ω. We show how to estimate those parameters consistently in the following. We consider only the case where there is no asymptotic bias; $\theta = 0$. Before we proceed, we define some random variables used in the result. Define

$$U = \begin{bmatrix} U_1 \\ U_1 - U_2 \end{bmatrix}.$$

Then $U \sim N(0_{2k\times1}, \Sigma)$ where $\Sigma = \begin{bmatrix} A & A-\Delta \\ A-\Delta' & A-\Delta-\Delta'+B \end{bmatrix}$. There exists a matrix, P such that $\Sigma = PP'$. Let $Z = P^{-1}U$. Then $Z \sim N(0_{2k\times1}, I_{2k\times2k})$. Define

$M_1 = P'N_1P$ where $N_1 = 1/2\begin{bmatrix} 0_{k\times k} & Q \\ Q & 0_{k\times k} \end{bmatrix}$ and

$M = P'NP$ where $N = \begin{bmatrix} 0_{k\times k} & 0_{k\times k} \\ 0_{k\times k} & Q \end{bmatrix}$.

It can be shown by some simple algebra that

(1) $U_1'Q(U_1-U_2) = Z'M_1Z$.

(2) $(U_1-U_2)'Q(U_1-U_2) = Z'MZ$.

This transformation permits us to use Ullah (1990)'s results on moments of the ratio of quadratic forms. Suppose that $\hat{A}, \hat{B}, \hat{\Delta}$ are consistent estimators for A, B, Δ respectively. We consider the following estimators for the combination control parameters.

(1) $\hat{\alpha} = tr\left[(\hat{A}-\hat{\Delta}-\hat{\Delta}'+\hat{B})Q\right]$.

(2) $\hat{\beta} = tr\left[(\hat{A}-\hat{\Delta}')Q\right]$.

(3) $\hat{\omega} = (\Gamma(1))^{-1}\int_0^\infty |\hat{N}_{1t}|^{-1/2}\,dt$ where $\hat{N}_{1t} = I + 2t\hat{\Sigma}_{22}Q$ and $\hat{\Sigma}_{22} = \hat{A}-\hat{\Delta}-\hat{\Delta}'+\hat{B}$.

$$(4)\ \hat{\nu} = (\Gamma(1))^{-1}\int_0^\infty |\hat{N}_{0t}|^{-1/2}\, tr\left[\hat{M}_1 \hat{N}_{0t}^{-1}\right]dt$$

where (1) $\Gamma(.)$ is the gamma function;

(2) $\hat{N}_{0t} = I + 2t\hat{M}$

Before we show the consistency of these estimators, we need to establish that the combination control parameters are not infinite.

Assumption 2-8 $\Sigma_{22} \equiv A - \Delta - \Delta' + B$ is positive definite.

Corollary 2-11 Suppose that Assumption 2-4 and 2-8 hold.

Then

(1) $|\alpha| < \infty$.

(2) $|\beta| < \infty$.

(3) $|\omega| < \infty$ if $k > 2$.

(4) $|\nu| < \infty$ if $k \neq 2$ and $k \neq 4$.

Proof: See Appendix.

Now we prove that the estimators defined above converges to the combination control parameters in probability.

Corollary 2-12 Suppose that Assumption 2-3, 2-4, 2-5, 2-6, 2-7 and 2-8 hold. Suppose that $k > 2$ and $k \neq 4$. Then

(1) $\hat{\alpha}_n \xrightarrow{p} \alpha$.

(2) $\hat{\beta}_n \xrightarrow{p} \beta$.

(3) $\hat{\omega}_n \xrightarrow{p} \omega$.

(4) $\hat{\nu}_n \xrightarrow{p} \nu$.

Proof: See Appendix.

In reality, we need to use the estimator ($\hat{\lambda}$) for the optimal weighting shrinkage factor (λ^*) and the estimated OWS estimator is given by

$$\delta_{\hat{\lambda}}^{OW}(b_n,g_n) = \left(1-\hat{\lambda}_1-\frac{\hat{\lambda}_2}{(b_n-g_n)'Q_n(b_n-g_n)}\right)(b_n - g_n) + g_n$$

where $\hat{\lambda}_1 = (\hat{\alpha}\hat{\omega}-1)^{-1}(\hat{\beta}\hat{\omega}-\hat{v})$ and $\hat{\lambda}_2 = (\hat{\alpha}\hat{\omega}-1)^{-1}(\hat{\alpha}\hat{v}-\hat{\beta})$.

An important question is whether we can still achieve the optimality (minimum asymptotic risk) when we use the estimated OWS estimator. The following corollary answers this question.

Corollary 2-13 Suppose that Assumption 2-3, 2-4, 2-5, 2-6, and 2-8 hold. Suppose that $k > 2$ and $k \neq 4$. Then

(1) $\hat{\lambda}_{1n} \xrightarrow{p} \lambda_1^*$ and $\hat{\lambda}_{2n} \xrightarrow{p} \lambda_2^*$.

(2) $n^{1/2}(\delta_{\lambda^*}^{OW}(b_n,g_n), \beta^0) \xrightarrow{d} \delta_{\lambda^*}^{OW}(U_1,U_2)$.

(3) $n^{1/2}(\delta_{\hat{\lambda}_n}^{OW}(b_n,g_n), \beta^0) \xrightarrow{d} \delta_{\lambda^*}^{OW}(U_1,U_2)$.

(4) $AR(\{\delta_{\hat{\lambda}_n}^{OW}(b_n,g_n)\}, \beta^0) = AR(\{\delta_{\lambda^*}^{OW}(b_n,g_n)\}, \beta^0)$

Proof: See Appendix.

The estimated OWS estimator has the same limiting distribution of the OWS estimator with true combination control parameters. Therefore, the asymptotic risk of the estimated OWS estimator achieve the minimum bound even though only estimators are used in place of the true combination control parameters.

1.3 Asymptotic Risk of the LAD estimator

In this section we will apply the results in the previous section to the LAD estimator[8]. One necessary condition is the asymptotic normality of the LAD estimator. Many versions of asymptotic theories have been developed. We use Bloomfield and Steiger's (1983) result because it allows the processes to be dependent. We first state their theorem. As usual we have $y = X\beta^0 + \varepsilon$ where $\beta^0 \in R^k$, $y \in R^n$.

Asymptotic Normality Theorem for the LAD Estimator

Suppose the following conditions hold.

A1. (y_t, X_t) is stationary and ergodic and for finite $\beta^0 \in R^k$, $k \in N$, $\varepsilon_t \equiv y_t - X_t'\beta^0$ is a Martingale Differences Sequence.

A2. X is independent of ε.

A3. X'X is positive definite almost surely for all n sufficiently large.

A3. ε_t has a continuous density, f(.), at zero such that f(0) > 0.

---

[8] An estimator, $b \in R^k$ is called the LAD estimator if and only if

$$b \in \operatorname{argmin} \sum_{t=1}^{n} | y_t - X_t'\beta |$$

The main reason we are interested in the LAD estimator is that it is robust to the outliers in the dependent variable. However the LAD estimator is not robust to the outliers in independent variables. This can be explained intuitively by the fact that the finite sample breakdown point of a univariate sample median is 50% and the LAD estimator is the conditional median estimator of the dependent variable as a function of the independent variables. Intuitively speaking, the LAD estimator is not stable with respect to explanatory variables and therefore there could be room for stabilizing the estimator by using some shrinkage techniques. Outliers can be interpreted in many ways. One way is to view them as coming from the fat tailed distributions. Recently many data, especially financial data tend to exhibit very fat tailed distributions. In this case the OLS estimator is too vulnerable to be relied on. We need other estimators such as the LAD estimator.

Despite this advantage, little attention has been paid to the LAD estimators in the development of econometrics. There are two main reasons. One is the difficulty in obtaining the LAD estimator because the usual differentiation cannot be used. Charnes et al. (1954) showed that the LAD estimator can be obtained by the simplex method linear programming. But this method was not efficient in that the parameter space grows along with the number of observations and as a result, it requires a long search time. Barrodale and Roberts (1974) proposed a modified version of simplex algorithm called Barrodale and Roberts L_1 Algorithm. This algorithm is much more efficient than the simplex method and greatly reduces the computation time. In the paper we use the Barrodale and Roberts L_1 Efficient Algorithm. The other reason is again computational burden in obtaining the standard error. In order to get the standard error, one should either estimate the density function of the dependent variable evaluated at zero or do the bootstrapping. With the rapidly growing computer power, this is not a great problem either. In the appendix, we explain heuristically why minimizing mean absolute error gives the median estimator and discuss how to convert the minimization problem into a linear programming.

A4. ε_t has a unique median

Let b^{LAD} be the LAD estimator. Then

$$n^{1/2}(b_n^{LAD}-\beta^0) \xrightarrow{d} N(0, (2f(0))^{-2}Q^{-1})$$

where Q is the second moment matrix of X.

Based on the asymptotic normality results we have the following facts. We cover only the non-random guess case for simplicity, and all results can easily be extended to the random guess case.

Corollary 3-1 Asymptotic Risk of the LAD Estimator

Suppose we have the quadratic loss function[9], $L(b_n,\beta^0) = (b_n-\beta^0)' X^{n\prime}X^n(b_n-\beta^0)$.

Under the relevant assumptions, $AR(\{b_n^{LAD}\},\beta^0) = (2f(0))^{-2}k$.

Proof: See Appendix.

Theorem 3-2 Asymptotic Risk of the JSLAD Estimator

Suppose we have the quadratic loss function, $L(b_n,\beta^0) = (b_n-\beta^0)' X^{n\prime}X^n(b_n-\beta^0)$.

Define the JSLAD estimator, b^{JSLAD}, as

$$b^{JSLAD} = (n^{1/2}2\hat{f}(0))^{-1} K(U_n, Q_n) + g_n$$

$$K(U_n,Q_n) = \left(1-\frac{\lambda}{U_n'Q_nU_n}\right)U_n$$

$$U_n = n^{1/2}2\hat{f}(0)(b_n- g_n)$$

where $\hat{f}(0)$ is a consistent estimator for f(0). Then

$$AR(\{b_n^{JSALD}\},\beta^0) = (2f(0))^{-2}\{k-\lambda(2(k-2)-\lambda)E\left(\frac{1}{k+2P-2}\right)\}$$

[9] Whenever we mention a loss function, we implicitly have the quadratic loss function in mind. But any function which satisfies the definition of the loss function can be used and the choice of a loss function depends on the nature of the study or the convenience of calculation. In many cases the quadratic loss function is easy to deal with. Although the LAD estimator minimizes the sum of the absolute value of the errors, we consider only the quadratic loss function; this is mainly for its computational convenience.

where $P \sim$ Poisson $(\theta'Q\theta/2)$ and $2(2f(0))^2(\beta^0-g_n)'X^{n\prime}X^n(\beta^0-g_n) \xrightarrow{p} \theta'Q\theta/2$.

Proof: See Appendix.

Finally we have the following fact showing the asymptotic inadmissibility of the LAD estimator.

Corollary 3-3 Asymptotic Inadmissibility of the LAD Estimator

Under the same assumption in Corollary 3-1 and Theorem 3-2,

$$AR(\{b_n^{JSLAD}\},\beta^0) < AR(\{b_n^{LAD}\},\beta^0) \text{ if } k > 2.$$

1.4 The Existence of An Improved LAD Estimator

Schmoyer and Arnold also prove that for a robust estimator, b_R, there exists a better estimator with strictly smaller risk in finite sample. In this section we apply their result to the LAD estimator. The improved robust estimator, $\delta(b_R)$, is defined as

$$\delta(b_R) = (I - \left(\frac{b}{a+\|b_R\|}\right)(X'X)^{-1})b_R$$

where a, b are some positive constants.

In order to apply their result to the LAD estimator, we need to show at least that the LAD estimator is translation equivariant and unbiased. An estimator, b, is called regression equivariant if and only if $b(\{(y_t+X_t'v, X_t)\ t = 1,2, \ldots n\}) = b(\{(y_t, X_t)\ t = 1,2, \ldots n\}) + v$ where v is any column vector[10].

Corollary 4-1 The LAD estimator is regression equivariant.

Proof: Let $b_1 \in \operatorname{argmin} \sum_{t=1}^{n} |y_t - X_t'\beta|$.

[10] See pp 116 in Rousseeuw and Leroy (1987) for definition.

Let $b_2 \in \operatorname{argmin} \sum_{t=1}^{n} |(y_t + X_t'v) - X_t'\beta| = \operatorname{argmin} \sum_{t=1}^{n} |y_t - X_t'(\beta - v)|$

Hence $b_2 - v = b_1$ implying that $b_2 = b_1 + v$. Q.E.D.

Next we need to check if the LAD estimator is unbiased or not. As mentioned earlier, Taylor (1974) prove that under some conditions, the LAD estimator is unbiased. One of condition is that the distribution of the error is symmetrical about zero. In addition to this, there are three more conditions. Andrews (1986) provide us with a general proof of unbiasedness of estimators including GLS, quasi-maximum likelihood, robust, adaptive, and spectral estimators. He shows that if an estimator is a solution of an optimization programming with some mild regularity conditions and the error is symmetric, then the estimator is unbiased. We use his result to show the unbiasedness of the LAD estimator.

Corollary 4-2 Unbiasedness of the LAD Estimator[11]

If the distribution of ε is symmetrical, then $E(b^{LAD}) = \beta$.

Now we are in the position to discuss the improved LAD estimator. Combing the properties shown above, we are able to obtain the Improved LAD estimator (ILAD).

Corollary 4-3 Improved LAD Estimator Existence

A1. The distribution of ε is symmetrical.

A2. $R(b^{LAD}, \beta) < \infty$.

A3. $Var(y) < \infty$.

A4. There exists $\varepsilon > 0$ such that $tr(\Sigma) - 2\xi > \varepsilon$ where $\Sigma = Cov(b^{LAD})$ and ξ = the largest eigenvalue of Σ.

A5. There exists $N > 0$ such that $N \geq E(\|b^{LAD} - \beta\|)^4 (Var(y))^{-2}$.

---

[11] In his proof, Taylor (1974) needed $E(\varepsilon) = 0$. It can be shown that if $E(\varepsilon)$ is not equal to zero, then only intercept estimator is biased and the other slop estimators still remain unbiased. Hence if this is the case, we can remove the intercept estimator using Frisch-Waugh-Lovell Theorem.

Then there exist positive constants a, b such that the estimator

$$\delta(b^{LAD}) = (I - \left(\frac{b}{a+\|b^{LAD}\|}\right)Q^{-1})b^{LAD}$$

has strictly smaller risk than the LAD estimator for all $\beta \in R^k$. We call this estimator Improved LAD Estimator (ILAD).[12]

Proof: Proof is straightforward and omitted.

---

[12] This ILADE has a finite sample property. We can also show the existence of an asymptotically improved extimator which can be applied to the LAD estimator. Suppose we have the following assumptions.

(1) $n^{1/2}(b_n-\beta^0) \xrightarrow{d} N(0,\sigma^2 A)$.

(2) $n^{1/2}(g_n-\beta^0) \longrightarrow \theta$. ($\{g_n\}$ is non-random guess sequence.)

(3) $n^{-1}Q_n \xrightarrow{p} Q$.

(4) $K(U_n, n^{-1}Q_n) = \left(1 - \frac{b}{a + U_n n^{-1} Q_n U_n}\right) U_n$ where a, b are constants and

$U_n = n^{1/2}(b_n - \beta^0)$ and $Q_n = \hat{\sigma}_n^{-2}(X^{n\prime}X^n)$.

(5) $\delta(b_n,g_n) = n^{-1/2}K(U_n, n^{-1}Q_n) + g_n$.

Then, for sufficiently small b and large a, $AR(\{\delta(b_n,g_n)\},\beta^0) < AR(\{b_n\},\beta^0)$.

Proof

Using the ART,

$$AR(\{\delta(b_n,g_n)\},\beta^0) = E\{L(K(U,Q), -\theta)\}$$

$$= E\{[\left(1 - \frac{b}{a + U'QU}\right) U+\theta]'Q[\left(1 - \frac{b}{a + U'QU}\right) U+\theta]\}$$

where $U \sim N(-\theta,Q^{-1})$.

Since $E(U) = -\theta$, we can utilize the technique used in James and Stein (1960) to get the following.

$$E\{L(K(U,Q), -\theta)\} < E\{L(U,-\theta)\}.$$

Therefore $AR(\{\delta(b_n,g_n)\},\beta^0) < AR(\{b_n\},\beta^0)$.

We call this estimator (applied to the LADE) Asymptotically Improved LADE (AILADE).

The finite sample formula for AILADE is

$$\delta(b_n,g_n) = n^{-1/2}K(U_n, n^{-1}Q_n) + g_n$$

$$= n^{-1/2}\left(1 - \frac{b}{a + n^{1/2}(b_n - g_n)n^{-1}Q_n n^{1/2}(b_n - g_n)}\right) n^{1/2}(b_n-g_n) + g_n$$

$$= \left(1 - \frac{b}{a + (b_n - g_n)Q_n(b_n - g_n)}\right)(b_n-g_n) + g_n.$$

Advantages over the finite-sample Improved Estimator are as follows

(1) b_n is not necessarily translation equivariant.
(2) b_n could be biased. This may allow us to do an iterative estimation. In other words, there is a possibility of using b_n^s as the guess of n+1 step iteration.
(3) The other technical assumptions such as A4 and A5 are no longer needed.

Even though the ILAD estimator looks like the usual JSLAD Estimator (JSLAD), there are several important differences. Since the ILAD estimator has a matrix shrinkage transformation, it shrinks each elements in the coefficient vector at a different rate. On the other hand, the JSLAD estimator has a single shrinkage rate for all coefficients. The improvement of the JSLAD estimator is guaranteed only asymptotically, but the improvement of the ILAD estimator can be obtained in the small sample whenever we have the right values of the two dimensional ridge parameters. The difficulty in choosing the parameters is essentially same as what we have when we are dealing with the Ridge estimator. That is why we call these parameters ridge parameters. An analysis of how to choose the right values of the ridge parameters is beyond the scope of this chapter. We conjecture that cross validation method may help in finding the right ridge values.

1.5 Simulation

In previous section, we have shown that the asymptotic risk of the JSLAD estimator is strictly less than the asymptotic risk of the LAD estimator. In this section, we discuss a Monte Carlo simulation designed to investigate the small sample properties of the shrinkage LAD estimators. We compare the approximated risks of all combination estimators developed in previous sections as well as the corresponding positive combination estimators where the weight is constrained to be positive. In the simulation, we specifically choose the LAD estimator and the OLS estimator to combine.

The basic model for the simulation is $y_t = x_t'\beta^0 + \varepsilon_t$ where $t = 1, 2, \ldots, n$, $\beta^0 \in R^k$, $n = 300$ and $k = 4$. We set $\beta^0 = 1$. The number of iteration is 1,000. We choose four symmetric distributions and two non-symmetric distribution for ε_t. The Uniform distribution within [-4,4], the standard normal distribution, the student t-distribution with 3 degrees of freedom and the Cauchy distribution with interquartile range 1 are selected for symmetric distributions. These represent moderate, heavy and very heavy tailed distribution. For the non-symmetric distribution, we choose the shifted Chi-square distribution centered at zero with 12 degrees of freedom and the shifted Rayleigh distribution centered at zero with parameter being 4. x_t is generated by the joint normal distribution, $N(0,\Sigma)$ where covariances are all 0.5 and variances are one[13]. The first entry of x_t is one.

In order to compute the combination of the two estimator, we need to estimate the error density evaluated at zero and the covariance matrix between two estimators. We estimate the density using a Kernel method with Gaussian Kernel. See the Appendix for detail discussion. Since both estimators are in the RCASOI (Regular Consistent Asymptotically Second Order Indexed) class, we can exploit the score and Hessian representations of both estimators to compute the covariance matrix. See Bates and White (1993) for a detail discussion. For each iteration, we compute the quadratic loss value for each estimator. We approximate the risk by averaging the loss value over simulation. Once we obtain the approximated risk for each estimator, then we compute the risk improvement (in percentage) relative to both the LAD estimator and the OLS estimator. All results are collected in Table 1.5.1 to 1.5.8.

It is known that the performance of the median compared with the sample mean is the worst among symmetric distributions in terms of asymptotic variance when the error is distributed uniformly. As expected, the risk of the LAD estimator (62.822) is greater than the risk of the OLS estimator (21.988) for the Uniform distribution with [-4,4]. All combination methods give negative weight to the LAD estimator. Both the OWS estimator and the NRC estimator dominate the JSC estimator. The risks of both the OWS estimator and the NRC estimator are smaller than the risk of the OLS estimator. When the regression error is normal, this is the case where the guess estimator (OLS) is asymptotically efficient. The risk of the OLS estimator (4.1213) is mildly smaller compared with the risk of the LAD estimator (6.2219). All combination methods again give negative weight to the LAD estimator. Both the OWS estimator and the NRC

---

[13] We use the multivariate normal random vector generator, G05EAF AND G05EZF in the MATLAB NAG Foundation Toolbox. We initialize the generator using G05CBF with the input, 22824 for each iteration. This means that we have the exactly same explanatory variables for each error distribution and for each guess so that we can compare the effects of different error distributions and different guesses.

estimator dominate the JSC estimator. All combination estimators have smaller risk than the LAD estimator, but greater than the OLS estimator. However, the deterioration of the OWS estimator and the NRC estimator relative to the OLS estimator is not large (-0.86 % and -0.83 % respectively). The Student-t distribution with 3 degrees of freedom has a relatively fat tail. As expected, the risk of the LAD estimator (7.348) is smaller than the risk of the OLS estimator (11.253). The weight to the LAD estimator is about 0.65 - 0.67 for all combinations. All combination estimators have smaller risk than both the LAD estimator and the OLS estimator. The improvement over the LAD estimator and the OLS estimator is about 1 - 3 % and 35 - 37 % respectively. Again, the OWS estimator and the NRC estimator dominate the JSC estimator. The Cauchy distribution represents a heavy tailed distribution. The performance of the OLS estimator is worst compared with the LAD estimator in terms of risk (236944.289 and 10.406 respectively). Nevertheless, combining the LAD estimator with the worst OLS estimator makes an improvement on the LAD estimator. The improvement over the LAD estimator and the OLS estimator is about 1 - 1.3 % and 100 % respectively.

The LAD estimator is worst compared with the OLS estimator in terms of risk (241.174 and 82.235) when the regression error is Chi-square distribution with 10 degrees of freedom. However, all combination methods fail to give negative weight to the LAD estimator. For the OWS estimator and the NRC estimator, the weight is very small (about 0.089 - 0.096). On the other hand, the JSC estimator give a large positive weight to the LAD estimator (0.807), which clearly shows the inferiority of the JSC estimator when the regression error is not symmetric. The failure can be explained by the bias in the constant coefficient, which makes the distance between two estimators very large (172.039). This in turns makes the JS weight over-estimated. The bias might also explain why the OWS estimator and NRC estimator fail to combine the two estimators optimally. All combination estimators are better than the LAD estimator, but worse than the OLS estimator. When the error has the Rayleigh distribution with parameter being 4, the result is basically same as in the Chi-square distribution. However, the skewness (-0.373) is smaller than in the Chi-square distribution (-0.660). As a result, the bias in the constant term is much smaller. The OWS estimator and the NRC estimator now give small negative weight to the LAD estimator.

We summarize the simulation results. First, the OWS estimator and the NRC estimator always dominate the JSC estimator. Second, the performance of the OWS estimator is virtually same as the performance of the NRC estimator. The contribution of the random weight is negligible. We recommend to use the NRC estimator because the weight is much easier to estimate. Third, when the regression error has a symmetric density (except normal density), all combination methods make improvement on both the LAD estimator and the OLS estimator. Lastly, when the density of the regression error is not symmetric, all combination methods seem to fail to find the optimal weight. The degree of failure seems to depend on the bias in the constant coefficient in the LAD estimator which is affected by the skewness of the density of the regression error. The bias in finite sample seems to have a big impact.

1.6 Out-of-Sample Predictive Ability: Empirical Study

In this section we investigate the out-of-sample predictive ability of the combination estimators developed in previous sections using two different data sets. In this application we combine the OLS estimator and the LAD estimator. We perform a comparison study where various estimators such as the OLS estimator, the LAD estimator, and stable estimators[14] (e.g. the Ridge estimator, the Garrotte estimator, the Non-negative Garrotte estimator and the LASSO estimator) are compared. We also estimate pre-determined combination (PDC) of the LAD estimator and the OLS estimator

[14] There is a class of estimators which has been developed in particular to improve out-of-sample predictability. Brieman (1994) argues that the good in-sample fit and the poor out-of-sample performance of the OLS estimator comes from the fact that it is unstable in the sense that a small change in data can lead to a big change in the predictor. His idea is to stabilize the OLS estimator by shrinking it towards zero. Based on this idea, Brieman (1994, 1995) proposes new estimators such as the Garrotte and the Non-Negative Garrotte. By a variety of simulation experiments he shows that the out-of-sample performance of the Garrotte, the Non-Negative Garrotte, and the Ridge estimators can be better than the OLS estimator. Motivated by Brieman's work, Tibshirani (1994) developed the LASSO (Least Absolute Shrinkage and Selection Operator). Even though there is no formal definition of stability, we call this class of estimators stable.

using a pre-determined weight called "Active Weight". We vary the weight with the increment being 0.05 over the closed interval [0,1]. In fact, the PDC contains the OLS estimator and the LAD estimator as special cases where the weight is 0 and 1 respectively. Therefore a total of 28 estimators are considered in the study. Their definitions are given in Table 1.6.1.

In order to evaluate the forecasting power, we use the following forecasting error measurements. Let y be the out-of-sample actual values and let e be the prediction errors.

$$PMSE(e) = e'e/T \qquad PMAD(e) = \frac{\sum_{k=1}^{T} |e_k|}{T}$$

$$R^2 = 1 - \frac{PMSE(e)}{S^2(y)} \qquad R^{2A} = 1 - \frac{PMAD(e)}{MAD(y)}.$$

$S^2(y) = MSE(y) - \bar{y}^2$, $MSE(y) = y'y/T$ and $MAD(y) = \frac{\sum_{k=1}^{T} |y_k - \bar{y}|}{T}$ where $\bar{y}$ is the sample mean and T is the number of out-of-sample observations[15].

We consider two data sets: daily Korean interest rates and daily US stock returns. The Korean interest rates considered are the 3 year Corporate Bond Rate (CBR) and the 3 month Certificate of Deposit Rate (CDR). See Figure 1.6.1 and Figure 1.6.2 for time-series plots for the variables. The US data set contains daily returns on ADC TeleCom Co. and HomeStake Co. stocks which are randomly chosen from the DATASTREAM database. We use daily excess returns which are obtained by subtracting the 3 month US T-bill rate from the returns. Table 1.6.2 and 1.6.3 provide summary statistics for both data sets.

---

[15] The prediction R^2 is not necessarily positive because out-of-sample predictions are not guaranteed to be orthogonal to out-of-sample residuals. The prediction R^2 compares the performance of a predictor to the imaginary situation where we are given the sample mean of the target variable over the entire out-of-sample period in advance and use it as our predictor. Therefore, positive prediction R^2 indicates that the predictor is better than the sample mean assumed to known in advance in terms of PMSE. If we define e* to be the out-of-sample error of the sample mean, then

$S^2(y) = PMSE(e^*)$.

Therefore, $R^2 > 0$ is equivalent to $PMSE(e^*) < PMSE(e)$.

Since we fail to reject the unit root hypothesis for the Korean interest rates, we use differenced variables. The forecasting model is

$$\Delta Dep_t = \alpha + \sum_{i=1}^{k1} \beta_i \Delta CBR_{t-i} + \sum_{i=1}^{k2} \gamma_i \Delta CDR_{t-i} + \varepsilon_t.$$

We use ΔCBR_t and ΔCDR_t as our dependent variables. We set $k_1 = k_2 = 1$. This model is for the one-step forecast horizon forecasting model.

We use 100 observations to estimate the model. Once we estimate the model, we then form 1 one-step out-of sample forecast, which completes one cycle of the estimation-forecasting process. After one complete cycle, we delete the first in-sample observation and add the first out-of-sample observation which we try to forecast during the first cycle. With this new 100 in-sample observations, we do the second cycle in the exactly same way. We complete 100 cycles in total generating 100 one step ahead point forecasts[16]. After we get the 100 point forecasts, we calculate the PMSE, PMAD, R^2 and R^{2A}. In other words, we use a rolling window prediction method where the estimation window size is 100, the prediction window size is 1 and the total number of windows is 100. We repeat the 100 estimations and forecasts for each of the 28 estimators and for each of the target variables[17]. The outcomes are summarized by Table 1.6.4 and Figure 1.6.3 through Figure 1.6.8.

---

[16] The out-of-sample period is 2/2/94 - 6/7/94.

[17] In order to estimate the optimal combination weight for the JSC and OWS estimators, we need to obtain a Choleskey decomposition of an estimated joint covariance matrix ($\hat{\Sigma}$) where

$$\hat{\Sigma} = \begin{bmatrix} \hat{A} & \hat{A} - \hat{\Delta} \\ \hat{A} - \hat{\Delta}' & \hat{A} - \hat{\Delta} - \hat{\Delta}' - \hat{B} \end{bmatrix}.$$

While the population joint covariance matrix is theoretically (semi) positive definite, the estimated sample covariance matrix sometimes fails to be so. In such case, the typical problem is that one of eigenvalues is negative. Since $\hat{\Sigma}$ is real and symmetric, there exit $\hat{Q}, \hat{\Lambda}$ such that $\hat{\Sigma} = \hat{Q}\hat{\Lambda}\hat{Q}'$ where columns of $\hat{Q}$ are eigenvectors and $\hat{\Lambda}$ contains eigenvalues on diagonal and zeros off diagonal. As long as all eigenvalues are non-negative we can obtain $\hat{P} = \hat{Q}\hat{\Lambda}^{1/2}$ as square root matrix. If there exists at least one negative eigenvalue, we cannot obtain such matrix. This is a potential problem in an empirical study. In this application, we encounter this problem when we analyses the Korean interest rate data sets, but not the US stock return data set. For the Korean data set we adjust the prediction period such a way we can avoid the problem. We guess that this problem has to do with the estimated covariance, $\hat{\Delta}$ because in two extreme cases where (1) $\hat{\Delta} = 0$ (2) $\hat{\Delta} = \Delta$, $\hat{\Sigma}$ is guaranteed to be (semi) positive definite. By the continuity property, there exits a $c \in (0,1)$ such that

The performance of the PDC estimators can be visually described by two plots. One is the plot of the PMSE against the active weights (PMSE plot, See Figure 1.6.3 and Figure 1.6.6) and the other is the plot of the PMAD against the active weights (PMAD plot, See Figure 1.6.4 and Figure 1.6.7). It is interesting that the LAD estimator has smaller PMSE than the OLS estimator and the PMSE is decreasing with active weight as visualized in Figure 1.6.3 and Figure 1.6.6. We can represent an estimator as a point in PMAD-PMSE space by drawing the scatter diagram of the PMAD and the PMSE (Figure 1.6.5 and Figure 1.6.8). In this diagram, we prefer the estimator which is located closer to the origin because the PMAD and the PMSE can be treated as bad commodities. We call the scatter diagram generated by the PDC estimators "active line". On each point of the active line, there is a corresponding active weight. We represent other estimators by its first initial in PMAD-PMSE space except the NRC estimator denoted by "c" and the Random Walk predictor denoted by "w". For example, "r" stands for the Ridge estimator and "g" for the Garrotte estimator and so on. For the CBR target variable, the Garrotte, Non-negative Garrotte and JSC estimators outperform all other estimators (See Figure 1.6.5). The NRC and OWS estimators have virtually same performance and are located near the active line. The out-of-sample forecasting for the CDR gives very similar pictures. The LAD estimator attains the best performance in terms of both PMSE and PMAD. The performance of the combination estimators are located between the LAD estimator and the OLS estimator. According to prediction R^2 the CBR is more difficult to predict than the CDR. All estimators give negative prediction R^2 when they are used to predict the CBR. However, for the CDR the LAD estimator and all combination estimators attain positive prediction R^2.

We have the following forecasting model for the US excess returns.

$$r_t = \alpha + \sum_{i=1}^{k1} \beta_i r_{t-i} + \sum_{i=1}^{k2} \gamma_i r_{m_{t-i}} + \varepsilon_t$$

$$\hat{\Sigma}(c) = \begin{bmatrix} \hat{A} & \hat{A} - c\hat{\Delta} \\ \hat{A} - c\hat{\Delta}' & \hat{A} - c\hat{\Delta} - c\hat{\Delta}' - \hat{B} \end{bmatrix}$$

is (semi) positive definite. Using $\hat{\Sigma}(c)$ seems to be better than using $\hat{\Sigma}(0)$, but in either case it is not consistent for Δ.

where r_{mt} is the daily excess returns on the S&P500 index and $k_1 = k_2 = 1$. The efficient market hypothesis predicts that $\alpha = \beta = \gamma = 0$. In other words the best predictor is zero. We call this Random Walk predictor and we include this in our comparison study. We set the size of estimation window and the number of prediction window to be 520 which is about two year sample period[18]. Again we use a fixed rolling window method.

The outcomes are summarized in Table 1.6.5 and Figure 1.6.9 through Figure 1.6.14. For ADC TeleCom Co. stock return, the LAD estimator again has smaller PMSE than the OLS estimator, but the PMSE is quadratic so that the PDC estimator with weight 0.75 achieve smaller PMSE than both estimators. Due to the quadratic effect, we have a curved active line in Figure 1.6.11. In this case, all combination estimators outperforms both the LAD and the OLS estimators in terms of PMSE. They also achieve better performance than stable regression estimators. Interestingly all estimators beat the Random Walk predictor in terms of PMSE, but the Random Walk predictor beats all estimators in terms of PMAD. Therefore, we have the situation where "The random walk wins partially." We have even stronger quadratic effect for HomeStake Co. return where the OLS estimator has smaller PMSE than the LAD estimator. As a result, the active line in Figure 1.6.14 is strongly convex toward the origin. The Non-negative Garrotte, Garrotte, Ridge and JSC estimators outperform all other estimators in terms of both PMSE and PMAD. The Random Walk predictor shows almost identical patterns. For the two US stock return data set, the combinations estimators generally achieve better performance than the LAD and OLS estimators but the magnitude of improvement is very small.

1.7 Conclusion

We have proposed an extension of the JS estimator in a direction that its risk improvement can be preserved when the sample size goes to infinity. This extension can

---

[18] The out-of-sample period is 4/5/94 - 3/29/96.

provide a reasonable answer to the problem of the JS estimator that it bas been confined primarily to small sample situation. This extension turns out to be equivalent to combining two estimators using a random combination weight. We have proposed the Optimal Weighting Scheme Estimator which includes both random combination and non-random combination as special cases. Using the asymptotic risk improvement results, we have applied shrinkage techniques to robust regression, especially the LAD Estimator, proposing three combination estimators. Our simulation shows that the shrinkage LAD estimator works, and makes some improvement over the LAD estimator as well as the OLS estimator except several cases. Even though I currently consider only the OLS guess, it seems plausible that I can refine the guess by choosing a robust estimator with an asymptotic normality distribution. It should be interesting to see how out-of-sample predictability, robustness, and risk improvement are affected by using a robust guess. The application of combination estimators to interest rate and stock return forecasting shows that it has some potential to improve out-of-sample forecasts.

Appendix: Proofs and Discussions

Proof of Theorem 2-2

(1) $U_n = n^{1/2}(b_n - g_n) = n^{1/2}(b_n - \beta^0 + \beta^0 - g_n) = n^{1/2}(b_n - \beta^0) - n^{1/2}(g_n - \beta^0) \xrightarrow{d} U \equiv U_1 - U_2$ by Assumption 2-4 and the continuity property.

(2) $n^{1/2}(\delta(b_n,g_n) - \beta^0) = n^{1/2}[n^{-1/2}K(U_n,n^{-1}Q_n) + (g_n - \beta^0)] = K(U_n, n^{-1}Q_n) + n^{1/2}(g_n - \beta^0) \xrightarrow{d} K(U,Q) + U_2$ by (1), Assumption 2-3.

(3) $L(\delta(b_n,g_n),\beta^0) = h(n^{1/2}(\delta(b_n,g_n)-\beta^0), n^{-1}Q_n) \xrightarrow{d} h(K(U,Q) + U_2,Q) = h(n^{1/2}K(U,Q) + U_2, n^{-1}Q) = L(K(U,Q),-U_2)$ by (2).

(4) $AR(\{\delta(b_n,g_n)\},\beta^0) = E\{L(K(U,Q), -U_2)\}$ by (3). Q.E.D.

Proof of Theorem 2-4

Let $f(\lambda) = AR(\{\delta_\lambda^{JS}(b_n,g_n)\},\beta^0)$. By Corollary 2-3 $f(\lambda) = \omega\lambda^2 - 2\nu\lambda + \kappa$ where ω, ν and κ are defined as in Theorem 2-4..

(1) $f'(\lambda) = 2\omega\lambda - 2\nu$ and $f''(\lambda) = 2\omega$. Since $\omega > 0$ by Assumption 2-6, $f''(\lambda) > 0$ for any λ.

(2) Setting $f'(\lambda) = 0$ and solving for λ, we have $\lambda^* = \nu/\omega$.

(3) By plugging $\lambda^* = \nu/\omega$ in $f(\lambda)$, we have $f(\lambda^*) = AR(\{\delta_{\lambda^*}^{JS}(b_n,g_n)\},\beta^0) = -\nu^2/\omega + \kappa$

(4) and (5) are obvious and omitted. Q.E.D.

Proof of Corollary 2-6

Let $f(\lambda) = AR(\{\delta_\lambda^{NR}(b_n,g_n)\},\beta^0)$. By Corollary 2-5 $f(\lambda) = R(\delta_\lambda^{NR}(U_1,U_2),0)$. It can be shown using some algebra that $f(\lambda) = \alpha\lambda^2 - 2\beta\lambda + \kappa$.

(1) $f'(\lambda) = 2\alpha\lambda - 2\beta$ and $f''(\lambda) = 2\alpha$. Since $\alpha > 0$ by Assumption 2-6, $f''(\lambda) > 0$ for any λ.

(2) Setting $f'(\lambda) = 0$ and solving for λ, we have $\lambda^* = \beta/\alpha$.

(3) By plugging $\lambda^* = \beta/\alpha$ in $f(\lambda)$, we have $f(\lambda^*) = AR(\{\delta_{\lambda^*}^{NR}(b_n,g_n)\},\beta^0) = -\beta^2/\alpha + \kappa$

(4) and (5) are obvious.

(6) We can show that $f(\lambda^*) = -\gamma^2/\alpha + AR(\{g_n\},\beta^0)$ which completes the proof. Q.E.D.

Proof of Theorem 2-8

Let $f(\lambda) = AR(\{\delta_\lambda^{OW}(b_n,g_n)\},\beta^0)$. By Corollary 2-7 $f(\lambda) = R(\delta_\lambda^{NR}(U_1,U_2),0)$. It can be shown using some algebra that $f(\lambda) = \alpha\lambda_1^2 - 2\beta\lambda_1 + 2\lambda_1\lambda_2 + \omega\lambda_2^2 - 2v\lambda_2 + \kappa$.

(1) The first order conditions are given by

$f_1(\lambda) = 2\alpha\lambda_1 - 2\beta + 2\lambda_2 = 0.$

$f_2(\lambda) = 2\lambda_1 - 2v + 2\omega\lambda_2 = 0.$

Since $f_{11}(\lambda) = 2\alpha$, $f_{22}(\lambda) = 2\omega$, and $f_{12}(\lambda) = 0$, the Hessian matrix is given by

$$H = \begin{bmatrix} 2\alpha & 0 \\ 0 & 2\omega \end{bmatrix} \text{which is positive definite.}$$

(2) Setting $f'(\lambda) = 0$ and solving for λ,

$\lambda_1^* = (\alpha\omega-1)^{-1}(\beta\omega-v)$.

$\lambda_2^* = (\alpha\omega-1)^{-1}(\alpha v-\beta)$.

(3) By plugging $\lambda_1^* = (\alpha\omega-1)^{-1}(\beta\omega-v)$ and $\lambda_2^* = (\alpha\omega-1)^{-1}(\alpha v-\beta)$.in $f(\lambda)$ and doing some algebra,

$$f(\lambda^*) = AR(\{\delta_{\lambda^*}^{NR}(b_n,g_n)\},\beta^0) = (\alpha\omega-1)^{-2}[-\alpha\beta^2\omega^2-(2\alpha\beta v-\alpha^2v^2+\beta^2)\omega+(\alpha v^2-2\beta v)] + \kappa.$$

(4) Let $h(\omega)$ $-[-\alpha\beta^2\omega^2-(2\alpha\beta v-\alpha^2v^2+\beta^2)\omega+(\alpha v^2-2\beta v)]$. We want to show $h(\omega) \geq 0$ for all ω.

Case 1: $\beta = 0$ & $v = 0$. Then $h(\omega) = 0$.

Case 2: $\beta = 0$ & $v \neq 0$. Then $h(\omega) = \alpha v^2(\alpha\omega-1) > 0$ because $\alpha > 0$, $v \neq 0$ and $\alpha\omega > 1$.

Case 3: $\beta \neq 0$ & $v = 0$. Then $h(\omega) = \beta^2\omega(\alpha\omega-1) > 0$ because $\omega > 0$, $\beta \neq 0$ and $\alpha\omega > 1$.

Case 4: $\beta \neq 0$ & $v \neq 0$. Let $\omega^* \in \{\omega \mid h(\omega) = 0\}$.

Sub-case 1: $\alpha v - \beta = 0$. Then $\omega^* = -(\alpha^2v^2-\beta^2-2\alpha\beta v)/2\alpha\beta^2$.

It can be shown that $\omega > \omega^*$ because $\alpha\omega > 1$ and $\beta \neq 0$.

This implies that $h(\omega) > 0$ for all ω.

Sub-case 2: $\alpha v - \beta \neq 0$. Then

$\omega_-^* = v(2\beta-\alpha)/\beta^2$ and $\omega_+^* = 1/\alpha$ with $\omega_-^* < \omega_+^*$.

It can be shown that $\omega > \omega_+^*$ because $\alpha\omega > 1$.

This implies that $h(\omega) > 0$ for all ω. Q.E.D.

Proof of Corollary 2-9

(1) Let X be a random variable which satisfies (1) Prob[X = μ] < 1 and (2) Prob[X > 0] = 1.
Then we can show using the strong form of Jensen's Inequality that 1/E(X) < E(1/X).
Let X be $(U_1-U_2)'Q(U_1-U_2)$. Then the first condition is satisfied by Assumption 2-7 and the second condition is satisfied by Assumption 2-3 and Assumption 2-6.

Therefore, we have $1/\mathrm{E}\left[(U_1-U_2)'Q(U_1-U_2)\right] < \mathrm{E}\left[\frac{1}{(U_1-U_2)'Q(U_1-U_2)}\right]$ which implies $1/\alpha < \omega$.

(2) $\lambda_1^* \geq 0 \Leftrightarrow \beta\omega - \nu \geq 0 \Leftrightarrow \nu - \beta\omega \leq 0 \Leftrightarrow$

$$\mathrm{E}\left[\frac{U_1'Q(U_1-U_2)}{(U_1-U_2)'Q(U_1-U_2)}\right] - \mathrm{E}\left[U_1'Q(U_1-U_2)\right]\mathrm{E}\left[\frac{1}{(U_1-U_2)'Q(U_1-U_2)}\right] \leq 0 \Leftrightarrow$$

$$\mathrm{Cov}\left[U_1'Q(U_1-U_2), \frac{1}{(U_1-U_2)'Q(U_1-U_2)}\right] \leq 0.$$

(3) $\lambda_2^* \geq 0 \Leftrightarrow \alpha\nu - \beta \geq 0 \Leftrightarrow \beta - \alpha\nu \leq 0 \Leftrightarrow$

$$\mathrm{E}\left[U_1'Q(U_1-U_2)\right] - \mathrm{E}\left[(U_1-U_2)'Q(U_1-U_2)\right]\mathrm{E}\left[\frac{U_1'Q(U_1-U_2)}{(U_1-U_2)'Q(U_1-U_2)}\right] \leq 0 \Leftrightarrow$$

$$\mathrm{Cov}\left[(U_1-U_2)'Q(U_1-U_2), \frac{U_1'Q(U_1-U_2)}{(U_1-U_2)'Q(U_1-U_2)}\right] \leq 0.$$ Q.E.D.

Proof of Corollary 2-10

Let $f(\lambda) = AR(\{\delta_\lambda^{OW}(b_n,g_n)\},\beta^0)$ which has a global minimum at λ^*. Therefore for any λ we have
$AR(\{\delta_{\lambda^*}^{OW}(b_n,g_n)\},\beta^0) \leq AR(\{\delta_\lambda^{OW}(b_n,g_n)\},\beta^0)$.
(1) Let $\lambda = [0\ \lambda_2^*]'$. Then $AR(\{\delta_{\lambda^*}^{OW}(b_n,g_n)\},\beta^0) \leq AR(\{\delta_\lambda^{OW}(b_n,g_n)\},\beta^0)$.
It can be shown by some algebra that $AR(\{\delta_\lambda^{OW}(b_n,g_n)\},\beta^0) = -\nu^2/\omega + \kappa = AR(\{\delta_{\lambda^*}^{JS}(b_n,g_n)\},\beta^0)$. We know that $[\lambda_1^*, \lambda_2^*]$ is the unique and global solution. Hence, if λ_1^* is not
equal to zero, then $[0\ \lambda_2^*]$ is not optimal and the strict inequality holds.
(2) Let $\lambda = [\lambda_1^*\ 0]'$. Then $AR(\{\delta_{\lambda^*}^{OW}(b_n,g_n)\},\beta^0) \leq AR(\{\delta_\lambda^{OW}(b_n,g_n)\},\beta^0)$.
It can be shown by some algebra that $AR(\{\delta_\lambda^{OW}(b_n,g_n)\},\beta^0) = -\beta^2/\alpha + \kappa = AR(\{\delta_{\lambda^*}^{NR}(b_n,g_n)\},\beta^0)$. We know that $[\lambda_1^*, \lambda_2^*]$ is the unique and global solution. Hence, if λ_2^* is not equal to zero, then $[\lambda_1^*\ 0]$ is not optimal and the strict inequality holds. Q.E.D.

Proof of Corollary 2-11

(1) $\alpha = E(V'QV)$ where $V = U_1 - U_2 \sim N(0, \Sigma_{22})$. Since α is the sum of variances and covariances of normal random variables, its absolute value is finite.
(2) $\beta = E(U_1QV)$. Since β is also the sum of variances and covariances of normal random variables, its absolute value is finite.

(3) Note that $\frac{1}{V'QV} = \frac{1}{Y'RY} = \frac{1}{Y'Y}\frac{Y'Y}{Y'RY}$ where Y = $P^{-1}V$ with PP' = Σ_{22} and R = P'QP.

It can be shown that $\frac{1}{\lambda_M} \le \frac{Y'Y}{Y'RY} \le \frac{1}{\lambda_m}$ where λ_M and λ_m is the largest and smallest eigenvalues of R (hence, of Q) respectively. This implies that $\frac{1}{\lambda_M} E\left[\frac{1}{Y'Y}\right] \le \omega \le \frac{1}{\lambda_m} E\left[\frac{1}{Y'Y}\right]$. If k > 2, then $\frac{1}{\lambda_M(k-2)} \le \omega \le \frac{1}{\lambda_m(k-2)}$.

(4) Note that $v^2 \le E\left[(U_1'QV)^2\right] E\left[\frac{1}{(V'QV)^2}\right]$ by the Cauchy-Schwarz inequality. Since U_1 and V are normal random variable, $E\left[(U_1'QV)^2\right] < \infty$. Form (3) we have

$\frac{1}{\lambda_M^2} E\left[\frac{1}{(Y'Y)^2}\right] \le E\left[\frac{1}{(V'QV)^2}\right] \le \frac{1}{\lambda_m^2} E\left[\frac{1}{(Y'Y)^2}\right]$. If k ≠ 2 and k ≠ 4, then

$\frac{1}{\lambda_M^2(k-2)(k-4)} \le E\left[\frac{1}{(V'QV)^2}\right] \le \frac{1}{\lambda_m^2(k-2)(k-4)}$. Therefore, $|v| < \infty$. Q.E.D.

Proof of Corollary 2-12

(1) $\alpha = E\left[(U_1 - U_2)'Q(U_1 - U_2)\right] = \text{tr}\{ E\left[(U_1 - U_2)'Q(U_1 - U_2)\right]\}$

$= E\{\text{tr}\left[(U_1 - U_2)'Q(U_1 - U_2)\right]\} = E\{\text{tr}\left[(U_1 - U_2)(U_1 - U_2)'Q\right]\}$

$= \text{tr}\{E\left[(U_1 - U_2)(U_1 - U_2)'Q\right]\} = \text{tr}[(A-\Delta-\Delta'+B)Q]$.

Since tr( ) is a continuous function, $\hat{\alpha} = tr\left[(\hat{A} - \hat{\Delta} - \hat{\Delta}' + \hat{B})Q\right]$ is consistent.

(2) By the same reasoning, $\hat{\beta} = tr\left[(\hat{A} - \hat{\Delta}')Q\right]$ is consistent.

(3) We define $g(\Sigma_{22}Q, t) \equiv \det(I + 2t\Sigma_{22}Q)$. Then we want to show that there exists a domination function, d(t), such that

1) $|g(\Sigma_{22}Q,t)| \le d(t)$ for all Σ_{22} and Q in a compact parameter space.

2) $\int_0^\infty d(t)dt < \infty$.

Using the relationship between determinant and eigenvalues of a matrix, we can express g(.) in terms of eigenvalues as follows.

$$g(\Sigma_{22}Q,t) = \left\{\prod_{i=1}^{k} \lambda_i\right\}^{1/2}$$

where λ_i is an eigenvalue of the inverse matrix of $I+2t\Sigma_{22}Q$. Using some linear algebra, we can obtain an upper bound given by

$$g(\Sigma_{22}Q,t)^2 \le \left\{\frac{1}{2|\bar{\kappa}|t+1}\right\}^k$$

where $\bar{\kappa}$ is the minimum (in absolute value) eigenvalue of $\Sigma_{22}Q$. Now we define d(t) as

$$d(t)=\left\{\frac{1}{2|\bar{\kappa}|t+1}\right\}^{k/2}.$$

As long as k > 2 which is assumed in the corollary, the domination function, D(t), can satisfy the other condition. Hence we have the desired result.

(4) We define $G(M,M_1,t)\equiv tr(M_1(I+2tM)^{-1})[\det(I+2tM)]^{-1/2}$. Then we want to show that there exists a domination function, D(t), such that

1) $|G(M,M_1,t)|\le D(t)$ for all M and M_1 in a compact parameter space.

2) $\int_0^{\infty} D(t)dt<\infty$.

Define $G_1(M,M_1,t)\equiv tr(M_1(I+2tM)^{-1})$ and $G_2(M,M_1,t)\equiv[\det(I+2tM)]^{-1/2}$. By using the similar argument, we can show that

$$G_1(M,M_1,t)^2\le\left\{\frac{1}{2|\bar{\kappa}|t+1}\right\}^{k}$$

where $\bar{\kappa}$ is the minimum (in absolute value) eigenvalue of M.

Using the trace version of Caucy-Schwarts inequality, we can obtain an upper bound for $G_2(M,M_1,t)$.

$$G_2(M,M_1,t)^2\le 2k|\bar{\xi}|\left\{\frac{1}{2|\bar{\kappa}|t+1}\right\}.$$

Where $\bar{\xi}$ is the maximum eigenvalue of M_1. By combining the two results, we have

$$G(M,M_1,t)\le 2k|\bar{\xi}|\left\{\frac{1}{2|\bar{\kappa}|t+1}\right\}^{k+1}.$$

Now we define D(t) as

$$D(t)=2k|\bar{\xi}|\left\{\frac{1}{2|\bar{\kappa}|t+1}\right\}^{k+1}.$$

The domination function, D(t), can satisfy the other condition. Hence we have the desired result. Q.E.D.

Proof of Corollary 2-13

(1) Both $\hat{\lambda}_1$ and $\hat{\lambda}_2$ are continuous function of the consistent estimators. Since the limit of continuous function of consistent estimators is the value of function evaluated at the limit of the consistent estimators, we have the desired results; $\hat{\lambda}_{1n}\xrightarrow{a.s.}\lambda_1^*$ and $\lambda_{2n}\xrightarrow{a.s.}\lambda_2^*$.

(2) The result immediately follows from Corollary 2-7; $n^{1/2}(\delta_{\lambda^*}^{OW}(b_n,g_n)-\beta^0)\xrightarrow{d}\delta_{\lambda^*}^{OW}(U_1,U_2)$.

(3) The result follows from the result in (1) and (2);

$$n^{1/2}(\delta_{\hat{\lambda}_n}{}^{OW}(b_n,g_n)-\beta^0) \xrightarrow{d} \delta_{\lambda^*}{}^{OW}(U_1,U_2).$$

(4) The result follows from (3) and Corollary 2-7. Q.E.D.

Proof of Corollary 3-1

$$L(b_n,\beta^0) = (b_n-\beta^0)'X^{n\prime}X^n(b_n-\beta^0)$$

$$= n^{1/2}(b_n-\beta^0)'\frac{X^{n\prime}X^n}{n}n^{1/2}(b_n-\beta^0)$$

$$= [n^{1/2}2f(0)(2f(0))^{-1}PP^{-1}(b_n-\beta^0)]'\frac{X^{n\prime}X^n}{n}[n^{1/2}2f(0)(2f(0))^{-1}PP^{-1}(b_n-\beta^0)]$$

where $PP' = Q^{-1}$.

$$\xrightarrow{d} (2f(0))^{-2}Z'Z \text{ where } Z \sim N(0,I).$$

Hence $AR(\{b_n\},\beta^0) = E((2f(0))^{-2}Z'Z) = (2f(0))^{-2}k$. Q.E.D.

Proof of Theorem 3-2

$$AR(\{b_n^{JSLAD}\},\beta^0) = (2f(0))^{-2}E\{L(K(U,Q),\theta)\}$$

$$= (2f(0))^{-2}E([K(U,Q)-\theta]'Q[K(U,Q)-\theta])$$

where $K(U,Q) = \left(1-\frac{\lambda}{U'QU}\right)U$ and $E(U) = \theta$.

$$= (2f(0))^{-2}\{k-\lambda(2(k-2)-\lambda)E\left(\frac{1}{k+2P-2}\right)\}$$

where $P \sim$ Poisson $(\theta'Q\theta/2)$ and $2(f(0))^2(\beta^0-g_n)'X^{n\prime}X^n(\beta^0-g_n) \xrightarrow{p} \theta'Q\theta/2$. (See Stein (1955) for detail.) Q.E.D.

Density Estimation using Gaussian Kernel

Suppose we have a random sample $\{e_i, i = 1, \ldots, n\}$ from a population with a density, f(x). The Kernel estimator of this density is given by

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x-e_i}{h}\right)$$

where K(x) is the standard normal density function, and h is the bandwidth. We use Silverman (1986)'s method to select the optimal bandwidth, h. We want to choose h such that it minimizes the mean integrated square error (MISE) of the density estimator. That is

$$h \in \text{argmin MISE}(\hat{f}(x)) = E\left(\int\{\hat{f}(x)-f(x)\}^2dx\right).$$

The solution is given by

$$h^* = k_2^{-2/5}\left(\int K(t)^2dt\right)^{1/2}\left(\int f''(x)^2dx\right)^{-1/2}n^{-1/2}$$

where $k_2 = \int t^2K(t)dt$. The problem with this solution is that it depends on the second derivative of the density which has to be estimated. If f(x) is the standard normal density, then $\int f''(x)^2 dx \approx 0.212\sigma^{-5}$. In this case, $h^* \approx 1.06\sigma n^{-1/5} \approx 0.79Rn^{-1/5}$ where R is the interquartile. It turns out that this is a good approximation for any density function. This approximation is sensitive to the skewness of the underlying density function, but not to kurtosis. All density functions except the Chi-square distribution we consider are symmetric. Hence this choice of the bandwidth can be justified.

Why does minimizing sum of absolute errors give the median estimator?

Let $\hat{\mu} \in \text{argmin } n^{-1}\sum_{t}^{n}|y_i - \mu|$. Then $\hat{\mu}$ is the sample median.

Heuristic Proof: Let m be the sample median. For simplicity, assume that n is odd. Then

$$\sum_{t}^{n} 1[y_i < m] = (n-1)/2$$

$$\sum_{t}^{n} 1[y_i > m] = (n-1)/2$$

$$\sum_{t}^{n} 1[y_i = m] = 1$$

We want to show $\hat{\mu} = m$. Suppose that $\hat{\mu} \neq m$. Without loss of generality, suppose that $\hat{\mu} = m + \varepsilon$. Then by the definition of $\hat{\mu}$ we have the following.

$$n^{-1}\sum_{t}^{n}|y_i - \hat{\mu}| < n^{-1}\sum_{t}^{n}|y_i - m|.$$

Let $L_i \equiv |y_i - \hat{\mu}|$. Then $L_i = |y_i - \hat{\mu}|1[y_i < m] + |y_i - \hat{\mu}|1[y_i > m] + |y_i - \hat{\mu}|1[y_i = m]$ where

$|y_i - \hat{\mu}|1[y_i < m] = |y_i - m| + |m - \hat{\mu}|$

$|y_i - \hat{\mu}|1[y_i > m] = |y_i - m| - |m - \hat{\mu}|$

$|y_i - \hat{\mu}|1[y_i = m] = |y_i - m| + |m - \hat{\mu}|$.

Therefore,

$$n^{-1}\sum_{t}^{n}|y_i - \hat{\mu}| = n^{-1}\sum_{t}^{n}|y_i - m| + n^{-1}(n-1)/2|m - \hat{\mu}| - n^{-1}(n-1)/2|m - \hat{\mu}| + n^{-1}|m - \hat{\mu}|$$

$$= n^{-1}\sum_{t}^{n}|y_i - m| + n^{-1}|m - \hat{\mu}|$$ which is a contradiction. Q.E.D.

How to convert L_1 problem into a linear programming?

The θth quantile problem is given by

(I) $\min \sum_{t}^{n}|y_t - x_t\beta|\{\theta 1[y_t \geq x_t\beta] + (1-\theta)1[y_t < x_t\beta]$.

If θ = ½, then this reduces to L_1 problem. Define

$u_t \equiv 1[y_t \geq x_t\beta]|y_t - x_t\beta|$

$v_t \equiv 1[y_t < x_t\beta]|y_t - x_t\beta|$

$a - c \equiv \beta$.

Then (I) is equivalent to the following linear programming.

$$(\text{II}) \quad \min \sum_{t}^{n}(\theta u_t + (1-\theta)v_t)$$

subject to (1) $y_t = x_t(a-c) + u_t + v_t \quad t = 1, 2, \ldots n$

(2) $a, c \geq 0$

(3) $u, v \geq 0.$

Testing whether $\lambda_2^* = 0$ by Simulation

The basic model for the simulation is $y_t = x_t'\beta^0 + \varepsilon_t$ where $t = 1, 2, \ldots, n$, $\beta^0 \in R^k$, $n = 200$ and $k = 8$. We set $\beta^0 = 1$. The number of iteration is 2,000. We choose standard normal distribution for ε_t. x_t is generated by the joint normal distribution, $N(0,\Sigma)$ where covariances are all 0.5 and variances are one[19]. The first entry of x_t is one.

We compute the OLS and LAD estimators and covariance matrices $\hat{A}, \hat{B}, \hat{\Delta}$ and generate M (2000) random numbers $\begin{bmatrix} u_{1i} \\ u_{2i} \end{bmatrix}$ from the joint normal distribution $N\left(\begin{bmatrix} 0_{k\times 1} \\ 0_{k\times 1} \end{bmatrix}, \begin{bmatrix} \hat{A}_{k\times k} & \hat{\Delta}_{k\times k} \\ \hat{\Delta}'_{k\times k} & \hat{B}_{k\times k} \end{bmatrix}\right)$. Then we compute the following random variates.

(1) $\alpha_i = (u_{1i}-u_{2i})'Q(u_{1i}-u_{2i})$

(2) $v_i = u_{1i}'Q(u_{1i}-u_{2i})/(u_{1i}-u_{2i})'Q(u_{1i}-u_{2i})$

Once we generate $\{\alpha_i, \beta_i, \omega_i, v_i, i = 1, 2, \ldots, M\}$, we compute the following statistics.

(1) $\bar{\alpha} = (1/M)\sum_{i}^{M}\alpha_i$

(2) $\bar{\beta} = (1/M)\sum_{i}^{M}\beta_i$

(3) $\bar{\omega} = (1/M)\sum_{i}^{M}\omega_i$

(4) $\bar{v} = (1/M)\sum_{i}^{M}v_i$

(5) $\bar{\lambda}_2 = (\bar{\alpha}\bar{\omega} - 1)^{-1}(\bar{\alpha}\bar{v} - \bar{\beta})$

(6) $\hat{C}ov(\alpha_i, v_i) = (1/M)\sum_{i}^{M}(\alpha_i - \bar{\alpha})(v_i - \bar{v})$: Sample Covariance

(7) P-value[20] for the null hypothesis H_0: $Cov(\alpha_i, v_i) = 0$

---

[19] We use the multivariate normal random vector generator, G05EAF AND G05EZF in the MATLAB NAG Foundation Toolbox. We initialize the generator using G05CBF with the input, 22824 for each iteration. This means that we have the exactly same explanatory variables for each error distribution and for each guess so that we can compare the effects of different error distributions and different guesses.

[20] The variance of sample covariance is given in Kendall M.G. and A. Stuart (1969) by

This completes one single simulation. We repeat this simulation 2000 times which produces 2000 realizations of each of $\overline{\lambda}_2$, $\hat{C}ov(\alpha_i, v_i)$, and P-value. Figure A shows the histograms with number of bins being 500. $\overline{\lambda}_2$ is distributed around 0.1048 and $\hat{C}ov(\alpha_i, v_i)$ is distributed around -0.0428. Even though many covariances are close to zero, the mean (-0.0428) seems significantly different from zero, which is supported by the fact that the percentage of P-value less than 0.1 is 55.3%.

Reference

Arthanari, T.S. and Dodge, Y. (1982) *Mathematical Programming in Statistics*, Wiley, New York.

Askin, Ronald G. and Montgomery, Douglas C. (1980), Augmented Robust Estimators, *Technometrics*, Vol. 22, No. 3, 333-341.

Barrodale, I. and Roberts, F.D.K. (1974), Algorithm 478: Solution of an Over-determined System of Equations in the L_1 Norm, *Communications of the Association for Computing Machinery*, 17, 319-320.

Bates, Charles E. and Halbert White (1993), Determination of Estimators with Minimum Asymptotic Covariance Matrices, *Econometric Theory*, 9, 633-648

Bloomfield, Peter and William L. Steiger (1983), *Least Absolute Deviations: Theory, Applications, and Algorithms.*

Breiman, L. (1994), Heuristics of Instability and Stabilization in Model Selection, *Technical Report No. 416*, Univ. of Cal., Berkeley

Breiman, L. (1995), Better Subset Regression Using the Nonnegative Garrote, Technometrics, Vol. 37, No. 4, 373-384

Charnes, A. and C. E. Lemke (1954), Computational Theory of Linear Programming: The Bounded Variables Problem, Graduate School of Ind. Administration, Carnegie Inst. Of Technol., Pittsburgh, Pennsylvania.

Cohen, Arthur (1976), Combining Estimates of Location, *Journal of the American Statistical Association*, Vol. 71, No. 353, 172-175.

Green, Edwin J. and Strawderman William E. (1991), James-Stein Type Estimator

---

$(1/n)(m_{22} - m_{11}^2 + m_{20}m_{01}^2 + m_{02}m_{10}^2 + 2m_{11}m_{01}m_{10} - 2m_{21}m_{01} - 2m_{12}m_{10}$

where $m_{ij} = (1/n)\sum_{t=1}^{n}(x_t - \overline{x})^i (y_t - \overline{y})^j$.

for Combining Unbiased and Possibly biased Estimators, *Journal of the American Statistical Association,* Vol 86, No 416, 1001-1006.

Hoerl, Arthur E. and Kennard, Robert W. (1970), Ridge Regression: Biased Estimation for Nonorthogonal Problems, *Technometrics*, Vol. 12, No. 1, 55-82.

James, W. and Stein, C. (1960), Estimation with Quadratic Loss. Proc. *Fourth Berkeley Symp. On Math. Statis. And Prob.*, Vol 1, 361-379.

Rousseeuw, Peter J. and Leroy, Annick M. (1987), *Robust Regression and Outlier Detecton.*

Saleh, A.K.M.E. and Sen, P.K. (1985), On Shinkage M-estimators of Location Parameters, *Commun. Statist. A - Theory Methods*, 14, 2313-2329.

Schmoyer, R. and Arnold, S. (1989), Shrinking Techniques for Robust Regression, *Contributions to Probability and Statistics: Essays in Honor of Ingram Olkin*, 368-384.

Sen, P.K. and Saleh, A.K.M.E. (1987), On preliminary Test and Shrinkage M-estimation in Linear Models, *Ann. Statist.*, 15, 1580-1592.

She, P. (1992), Another Solution, *Bulletin of the Institute of Mathematical Statistics*, 21, 234.

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, Chapman and Hall.

Stein, C. (1955), Inadmissibility of the Usual Estimator for the Mean of a Multivariate Distribution, *Proc. Third Berkeley Symp. Math. Statist. Prob. 1*, 197-206, Univ. of California Press.

Taylor, Lester D. (1974), Estimation by Minimizing the Sum of Absolute Errors, *Frontiers in Econometrics*, Academic Press.

Tibshirani, R. (1994), Regression Shrinkage and Selection via the LASSO, *Technical Report*, Univ. of Toronto.

Ullah, A. (1990), Contribution to Econometric Theory and Application, *Essays in Honour of A. L. Naga*, 242-292

Vinod, Hrishikesh D, and Ullah, Aman (1981), *Recent Advances in Regression Methods.*

Table 1.5.1 Averaged Parameter Estimates

	Uniform	Normal	Student-t	Cauchy	Chi-square	Rayleigh
LAD β_1	0.993	0.998	1.004	0.994	0.353	0.632
β_2	0.990	1.001	1.004	1.006	1.004	1.000
β_3	0.997	0.998	1.001	1.002	1.007	1.010
β_4	1.003	1.002	0.997	0.989	0.993	1.004
OLS β_1	0.996	0.998	1.001	1.783	0.999	0.998
β_2	0.994	1.006	1.007	1.259	1.001	0.991
β_3	1.003	0.998	1.005	1.856	1.021	0.999
β_4	1.001	1.001	0.995	0.095	0.998	1.004
f(0)	0.131	0.389	0.348	0.272	0.091	0.117
$[2f(0)]^{-2}$	14.548	1.651	2.070	3.389	29.866	18.319
σ^2	5.310	1.001	2.877	91380.121	19.979	10.745
α	1.369	1.540	3.537	93185.433	1.676	1.532
β	1.955	1.560	0.997	-11.369	1.502	1.594
ν	1.440	1.034	0.330	-0.018	0.922	1.064
ω	1.386	1.265	0.645	0.010	1.178	1.272
λ_1	1.424	1.009	0.338	-0.015	0.900	1.038
λ_2	0.011	0.019	-0.014	-6.772	0.018	0.020
L	20.503	2.312	6.676	236897.640	172.039	67.491
w(NRC)	-0.432	-0.022	0.666	1.015	0.089	-0.051
w(JSC)	-0.514	-0.129	0.652	1.013	0.807	0.636
w(OWS)	-0.440	-0.033	0.669	1.015	0.096	-0.046
$w(NRC)^+$	0.000	0.039	0.666	1.015	0.112	0.031
$w(JSC)^+$	0.233	0.315	0.710	1.013	0.810	0.669
$w(OWS^+$	0.000	0.037	0.669	1.015	0.117	0.033

Note: The combination weights are defined as

$w(NRC) = 1 - (\beta/\alpha)$ $\quad w(NRC)^+ = \max\{0, w(NRC)\}$

$w(JS)C = 1 - (\nu/\omega)L^{-1}$ $\quad w(JSC)^+ = \max\{0, w(JSC)\}$

$w(OWS) = 1 - \lambda_1 - \lambda_2 L^{-1}$ $\quad w(OWS)^+ = \max\{0, w(OWS)\}$

where $L = (b^{LAD}-b^{OLS})'X'X(b^{LAD}-b^{OLS})$.

Table 1.5.2 Summary Statistics for Error Distribution.

	Uniform	Normal	Student-t	Cauchy	Chi-square	Rayleigh
Mean	-0.003	-0.002	0.001	0.806	-0.001	-0.002
Median	-0.008	-0.001	0.001	-0.007	-0.660	-0.373
Std. Dev.	2.304	1.001	1.676	71.288	4.465	3.274
Skewness	0.000	-0.001	0.027	0.264	0.867	0.622
Kurtosis	1.813	2.980	17.325	147.022	4.073	3.216

Table 1.5.3 Risk Comparison: Uniform Distribution

	Risk	Improvement relative to LAD	Improvement relative to OLS
LAD	62.822	0.00 %	-185.71 %
OLS	21.988	65.17 %	0.00 %
NRC	17.340	72.40 %	21.14 %
JSC	41.046	34.66 %	-86.86 %
OWS	17.339	72.40 %	21.14 %
Positive NRC	21.988	65.17 %	0.00 %
Positive JSC	34.874	44.49 %	-58.60 %
Positive OWS	21.988	65.17 %	0.00 %

Table 1.5.4 Risk Comparison: Standard Normal Distribution

	Risk	Improvement relative to LAD	Improvement relative to OLS
LAD	6.2219	0.00 %	-50.97 %
OLS	4.1213	33.76 %	0.00 %
NRC	4.1569	33.19 %	-0.86 %
JSC	5.1314	17.53 %	-24.51 %
OWS	4.1556	33.21 %	-0.83 %
Positive NRC	4.1306	33.61 %	-0.23 %
Positive JSC	4.7193	24.15 %	-14.51 %
Positive OWS	4.1309	33.61 %	-0.23 %

Table 1.5.5 Risk Comparison: Student-t Distribution

	Risk	Improvement relative to LAD	Improvement relative to OLS
LAD	7.348	0.00 %	34.70 %
OLS	11.253	-53.15 %	0.00 %
NRC	7.124	3.05 %	36.70 %
JSC	7.269	1.08 %	35.41 %
OWS	7.124	3.05 %	36.70 %
Positive NRC	7.124	3.05 %	36.70 %
Positive JSC	7.182	2.25 %	36.17 %
Positive OWS	7.124	3.05 %	36.70 %

Table 1.5.6 Risk Comparison: Cauchy Distribution

	Risk	Improvement relative to LAD	Improvement relative to OLS
LAD	10.406	0.00 %	100.00 %
OLS	236944.289	-22768.91 %	0.00 %
NRC	10.269	1.32 %	100.00 %
JSC	10.303	0.99 %	100.00 %
OWS	10.273	1.28 %	100.00 %
Positive NRC	10.269	1.32 %	100.00 %
Positive JSC	10.303	0.99 %	100.00 %
Positive OWS	10.273	1.28 %	100.00 %

Table 1.5.7 Risk Comparison: Shifted Chi-square Distribution

	Risk	Improvement relative to LAD	Improvement relative to OLS
LAD	241.174	0.00 %	-193.27 %
OLS	81.235	65.90 %	0.00 %
NRC	86.196	64.26 %	-4.82 %
JSC	199.925	17.10 %	-143.11 %
OWS	86.249	64.24 %	-4.88 %
Positive NRC	85.370	64.60 %	-3.81 %
Positive JSC	199.958	17.09 %	-143.15 %
Positive OWS	85.553	64.53 %	-4.03 %

Table 1.5.8 Risk Comparison: Shifted Rayleigh Distribution

	Risk	Improvement relative to LAD	Improvement relative to OLS
LAD	113.001	0.00 %	-165.90 %
OLS	42.499	62.39 %	0.00 %
NRC	43.997	61.06 %	-3.52 %
JSC	87.585	22.50 %	-106.08 %
OWS	43.976	61.08 %	-3.48 %
Positive NRC	43.204	61.77 %	-1.66 %
Positive JSC	87.138	22.89 %	-105.03 %
Positive OWS	43.255	61.72 %	-1.78 %

Table 1.6.1 Definition of Estimators

Estimator	Definition
PDC ($\delta_A(w)$) with weight (w) = 0	$\delta^{PD}(w) = b^{LS}$
PDC ($\delta_A(w)$) with weight (w) $\in$ (0,1)	$\delta^{PD}(w) = wb^{LAD} + (1-w)\, b^{LS}$
PDC ($\delta_A(w)$) with weight (w) = 1	$\delta^{PD}(w) = b^{LAD}$
Ridge (b^R)	$b^R \in \text{argumin} \lVert y - Xb \rVert^2 \quad \text{st. } b'b < s$
Garrotte (b^G)	$b^G \in \text{argumin} \lVert y - Z\gamma \rVert^2$ st. $Z_{ij} = X_{ij}b_j^{LS}$, $\gamma'\gamma < s$
Non-Negative Garrotte (b^N)	$b^G \in \text{argumin} \lVert y - Z\gamma \rVert^2$ st. $Z_{ij} = X_{ij}b_j^{LS}$, $\gamma' 1 < s$, $\gamma \geq 0$
LASSO (b^L)	$b^L \in \text{argumin} \lVert y - Xb \rVert^2 \quad \text{st. } b' 1 < s$

1. i is the unit vector.
2. Values for s are determined by multi-fold cross-validation.

Table 1.6.2 Summary Statistics (Korean Interest Rates)

	cbr	Δcbr	cdr	Δcdr
mean	13.18	-0.0026	13.53	-0.0023
median	13.10	0.0000	13.40	0.0000
maximum	15.50	0.4300	17.00	1.4000
minimum	10.97	-0.4500	11.25	-1.7300
standard error	1.08	0.0842	1.38	0.1908
skewness	0.21	-0.0629	0.31	-0.7832
kurtosis	2.16	7.4924	2.22	23.5690
ADF statistics	-1.70	-7.34	-1.85	-8.69
Critical Value (10%)	-2.56	-1.61	-2.56	-1.61

1. ADF statistics for level was calculated with constant in the regression with 10 lags.
2. ADF statistics for difference was calculated with no deterministic components in regression. Again 10 lags were used.

Table 1.6.3 Summary Statistics (US Stock Returns)

	ADC TelCom.	HomeStake Co.
mean	0.1472	0.0106
median	-0.0142	-0.0210
maximum	11.9167	11.2545
minimum	-22.1296	-12.4531
standard error	2.9448	2.5201
skewness	-0.1703	0.0780
kurtosis	3.9286	1.9583

Table 1.6.4 Prediction Result (Korean Interest Rates)

	CBR Forecasting				CDR Forecasting			
	PMSE	PMAD	R^2	R^{2A}	PMSE	PMAD	R^2	R^{2A}
OLS	0.001566	0.022570	-0.20003	-0.038250	0.006382	0.048341	-0.017570	0.014544
0.05	0.001558	0.022424	-0.19339	-0.031550	0.006358	0.048213	-0.013750	0.017145
0.10	0.001549	0.022278	-0.18707	-0.024840	0.006335	0.048085	-0.009980	0.019746
0.15	0.001541	0.022132	-0.18105	-0.018130	0.006311	0.047958	-0.006260	0.022347
0.20	0.001534	0.021987	-0.17534	-0.011430	0.006288	0.047830	-0.002590	0.024948
0.25	0.001527	0.021841	-0.16994	-0.004720	0.006265	0.047703	0.001028	0.027549
0.30	0.001520	0.021695	-0.16485	0.001983	0.006243	0.047575	0.004596	0.030150
0.35	0.001514	0.021549	-0.16007	0.008689	0.006221	0.047447	0.008115	0.032751
0.40	0.001508	0.021404	-0.15560	0.015394	0.006199	0.047320	0.011583	0.035352
0.45	0.001503	0.021258	-0.15143	0.022100	0.006178	0.047192	0.015002	0.037953
0.50	0.001498	0.021112	-0.14758	0.028805	0.006157	0.047066	0.018370	0.040538
0.55	0.001493	0.020966	-0.14403	0.035511	0.006136	0.046953	0.021689	0.042828
0.60	0.001489	0.020820	-0.14079	0.042216	0.006115	0.046850	0.024957	0.044933
0.65	0.001485	0.020675	-0.13786	0.048922	0.006095	0.046747	0.028176	0.047037
0.70	0.001482	0.020529	-0.13524	0.055627	0.006075	0.046647	0.031344	0.049060
0.75	0.001479	0.020383	-0.13293	0.062333	0.006056	0.046551	0.034462	0.051020
0.80	0.001476	0.020237	-0.13093	0.069038	0.006037	0.046455	0.037531	0.052979
0.85	0.001474	0.020092	-0.12923	0.075744	0.006018	0.046359	0.040549	0.054938
0.90	0.001472	0.019946	-0.12785	0.082450	0.005999	0.046263	0.043517	0.056898
0.95	0.001471	0.019800	-0.12677	0.089155	0.005981	0.046168	0.046435	0.058825
LAD	0.001470	0.019654	-0.12600	0.095861	0.005963	0.046089	0.049303	0.060443
NRC	0.001551	0.022204	-0.18829	-0.021440	0.006141	0.047022	0.020892	0.041416
JSC	0.001459	0.019984	-0.11822	0.080683	0.006014	0.046295	0.041128	0.056237
OWS	0.001549	0.022183	-0.18677	-0.020470	0.006140	0.047013	0.021096	0.041609
RIDGE	0.001524	0.020747	-0.16791	0.045605	0.006404	0.048086	-0.021010	0.019738
GAR	0.001447	0.020805	-0.10881	0.042921	0.006317	0.047330	-0.007210	0.035147
NNGAR	0.001450	0.020789	-0.11097	0.043669	0.006335	0.047189	-0.010070	0.038014
LASSO	0.001577	0.021926	-0.20838	-0.008650	0.006384	0.048369	-0.017940	0.013963

Table 1.6.5 Prediction Result (US Stock Returns)

	ADC TeleCom Co				HomeStake Co.			
	PMSE	PMAD	R^2	R^{2A}	PMSE	PMAD	R^2	R^{2A}
OLS	11.02739	2.401711	-0.001410	0.000879	4.275478	1.578444	0.005622	-0.00955
0.05	11.02243	2.401033	-0.000970	0.001161	4.273883	1.577932	0.005993	-0.00923
0.10	11.01783	2.400355	-0.000550	0.001443	4.272477	1.577507	0.006320	-0.00895
0.15	11.01356	2.399721	-0.000160	0.001706	4.271260	1.577082	0.006603	-0.00868
0.20	11.00965	2.399154	0.000196	0.001942	4.270232	1.576675	0.006842	-0.00842
0.25	11.00608	2.398710	0.000520	0.002127	4.269392	1.576272	0.007037	-0.00816
0.30	11.00286	2.398321	0.000813	0.002289	4.268741	1.575870	0.007189	-0.00791
0.35	10.99998	2.397954	0.001075	0.002442	4.268279	1.575474	0.007296	-0.00765
0.40	10.99745	2.397595	0.001304	0.002591	4.268006	1.575094	0.007360	-0.00741
0.45	10.99526	2.397325	0.001503	0.002703	4.267922	1.574863	0.007379	-0.00726
0.50	10.99342	2.397056	0.001670	0.002815	4.268026	1.574732	0.007355	-0.00718
0.55	10.99193	2.396820	0.001806	0.002914	4.268319	1.574646	0.007287	-0.00712
0.60	10.99078	2.396632	0.001910	0.002992	4.268801	1.574561	0.007175	-0.00707
0.65	10.98998	2.396474	0.001983	0.003057	4.269471	1.574478	0.007019	-0.00702
0.70	10.98952	2.396353	0.002024	0.003107	4.270331	1.574396	0.006819	-0.00696
0.75	10.98941	2.396233	0.002034	0.003158	4.271379	1.574314	0.006575	-0.00691
0.80	10.98964	2.396112	0.002013	0.003208	4.272616	1.574232	0.006288	-0.00686
0.85	10.99023	2.395991	0.001960	0.003258	4.274042	1.574152	0.005956	-0.00681
0.90	10.99115	2.395876	0.001876	0.003306	4.275656	1.574075	0.005580	-0.00676
0.95	10.99243	2.395761	0.001760	0.003354	4.277460	1.574037	0.005161	-0.00673
LAD	10.99405	2.395646	0.001613	0.003402	4.279452	1.574043	0.004698	-0.00674
NRC	10.99121	2.396691	0.001871	0.002967	4.271405	1.577233	0.006569	-0.00878
JSC	10.99218	2.395732	0.001782	0.003366	4.269610	1.573878	0.006987	-0.00663
OWS	10.99136	2.396698	0.001858	0.002964	4.271247	1.577195	0.006606	-0.00875
RIDGE	11.10845	2.395347	-0.000540	0.003894	4.272639	1.572277	0.006282	-0.00561
GAR	11.02588	2.398849	-0.001280	0.002069	4.269367	1.573136	0.007043	-0.00616
NNGAR	11.01529	2.398690	-0.000320	0.002136	4.254709	1.569847	0.010452	-0.00405
LASSO	11.00619	2.393585	0.000511	0.004259	4.275490	1.575523	0.005619	-0.00768
Random Walk	11.03380	2.392080	-0.001999	0.004883	4.300280	1.565840	-0.000147	-0.00149

Figure 1.5.1 Uniform Distribution

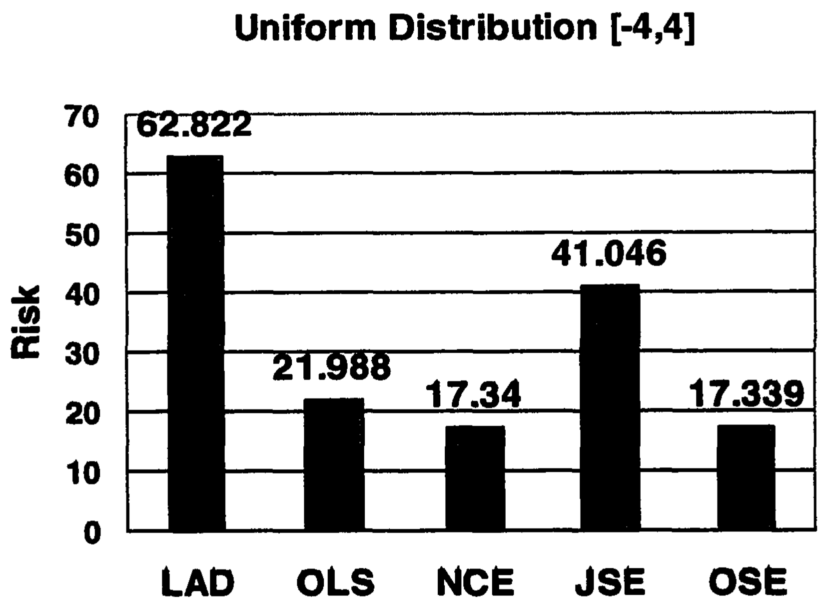


Figure 1.5.2 Standard Normal Distribution

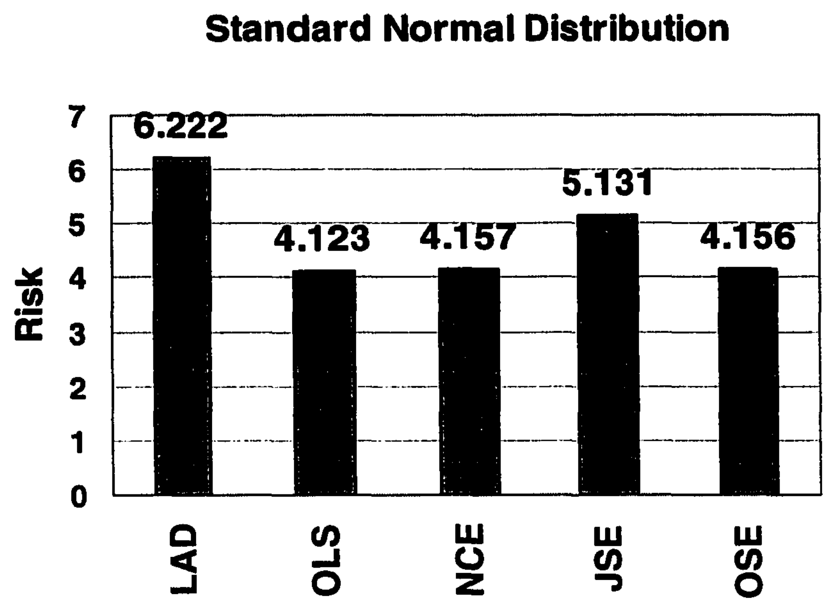


Figure 1.5.3 Student-t Distribution (Degree of Freedom = 3)

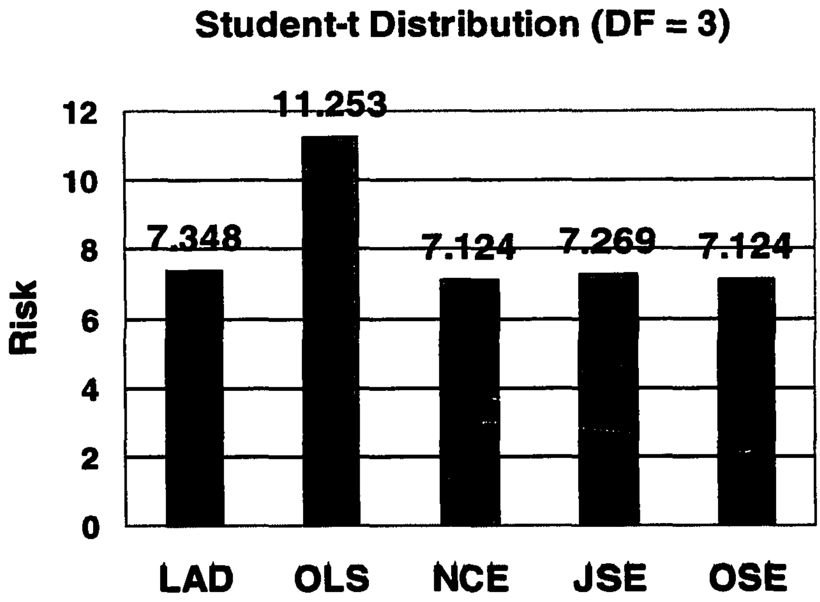


Figure 1.5.4 Cauchy Distribution (InterQuartile = 1)

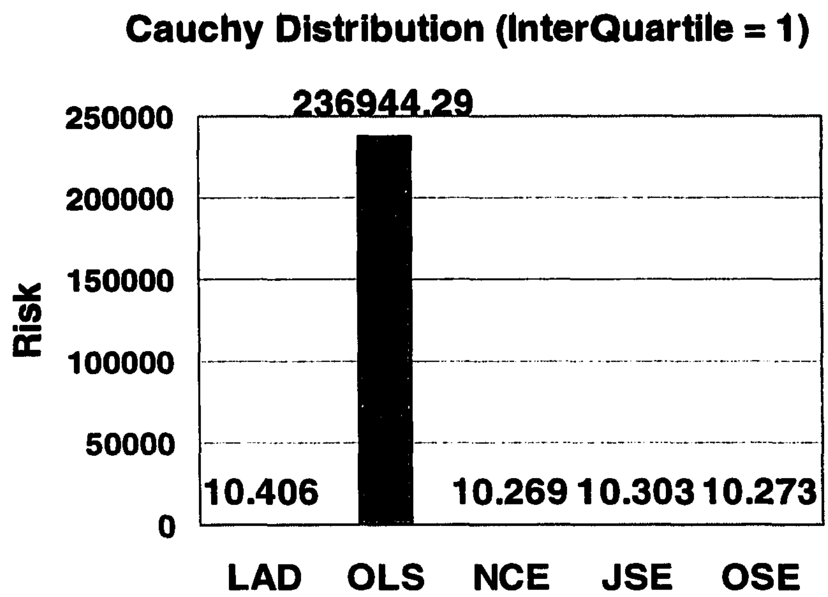


Figure 1.5.5 Chi-square Distribution (centered at zero)

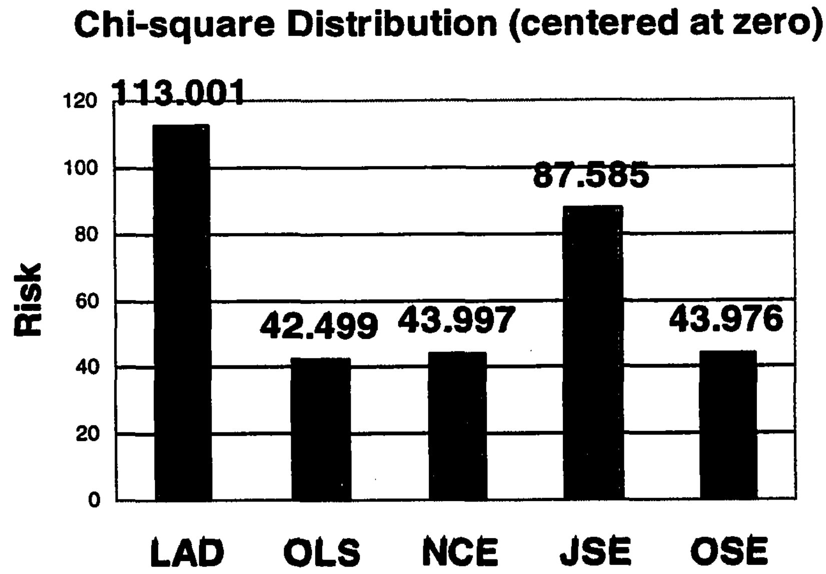


Figure 1.5.6 Rayleigh Distribution (centered at zero)

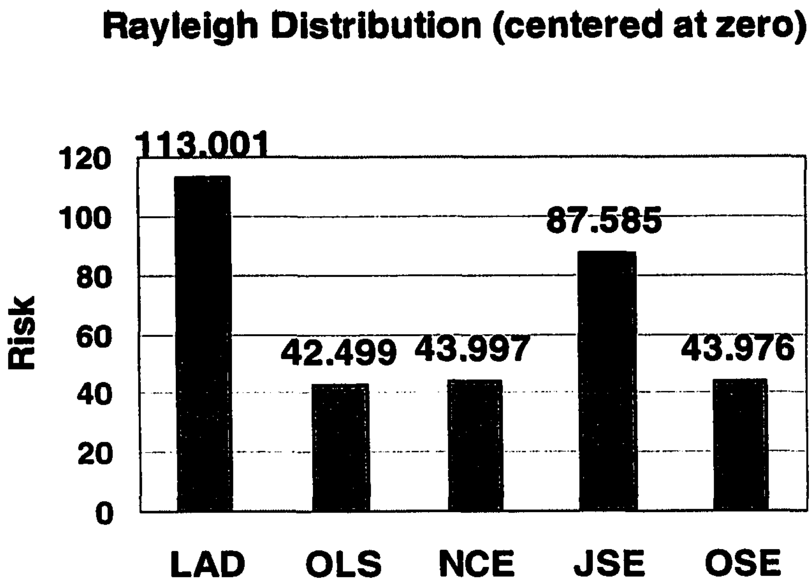


Figure 1.6.1 Korean 3 Year Corporate Bond Rate

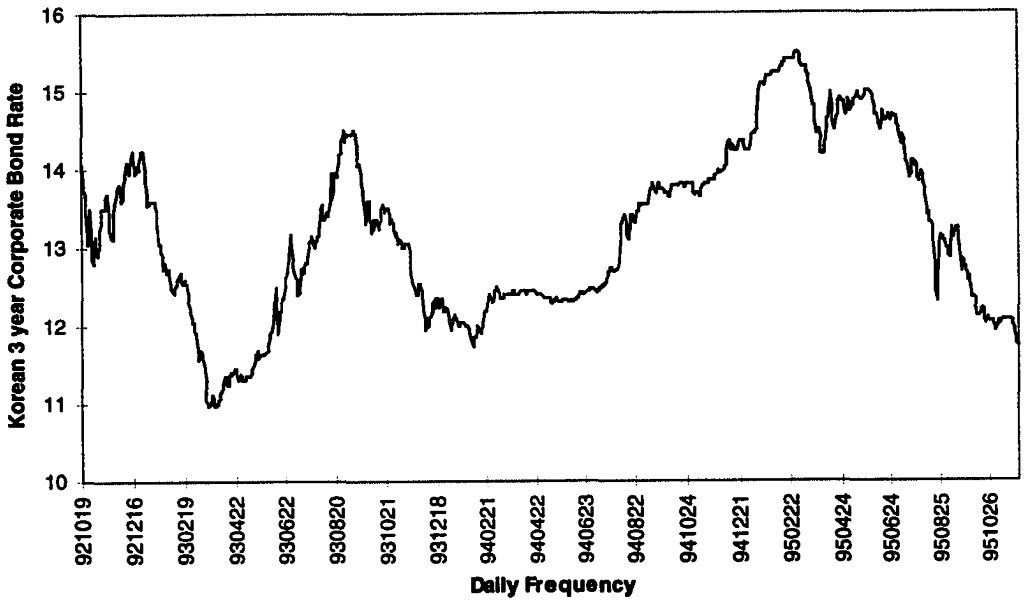


Figure 1.6.2 Korean 3 Month Certificate Deposit Rate

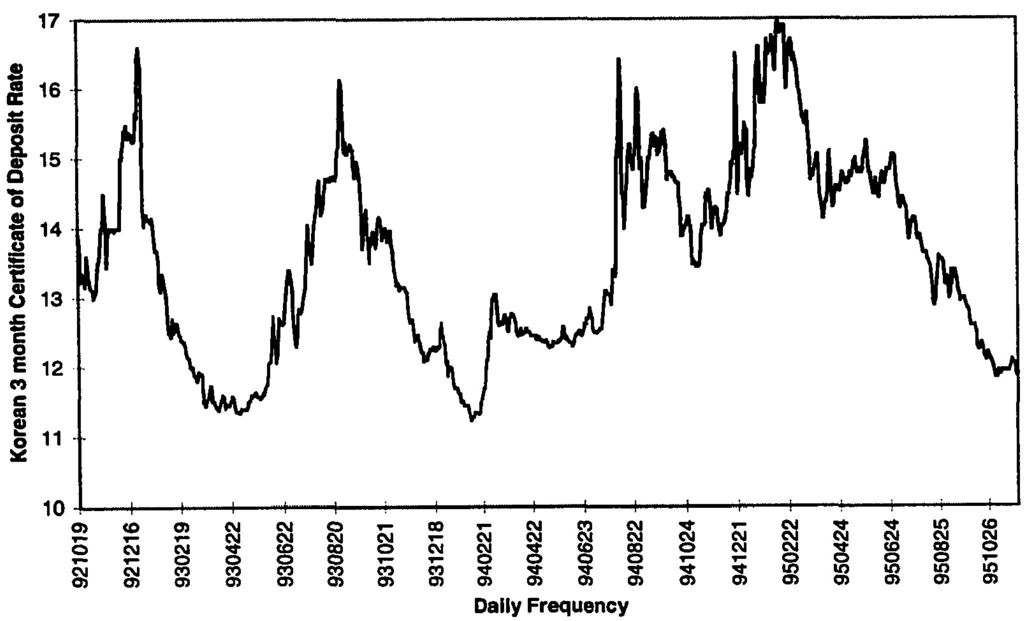


Figure 1.6.3 Out-of-Sample MSE (CBR, Training Sample Size = 100)

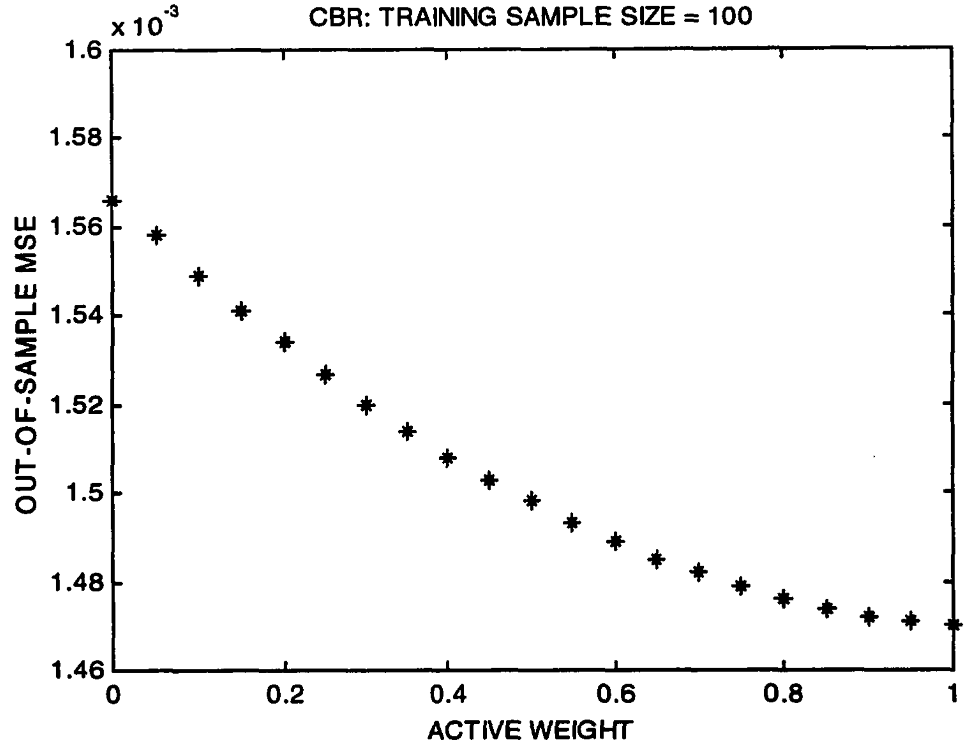


Figure 1.6.4 Out-of-Sample MAD (CBR, Training Sample Size = 100)

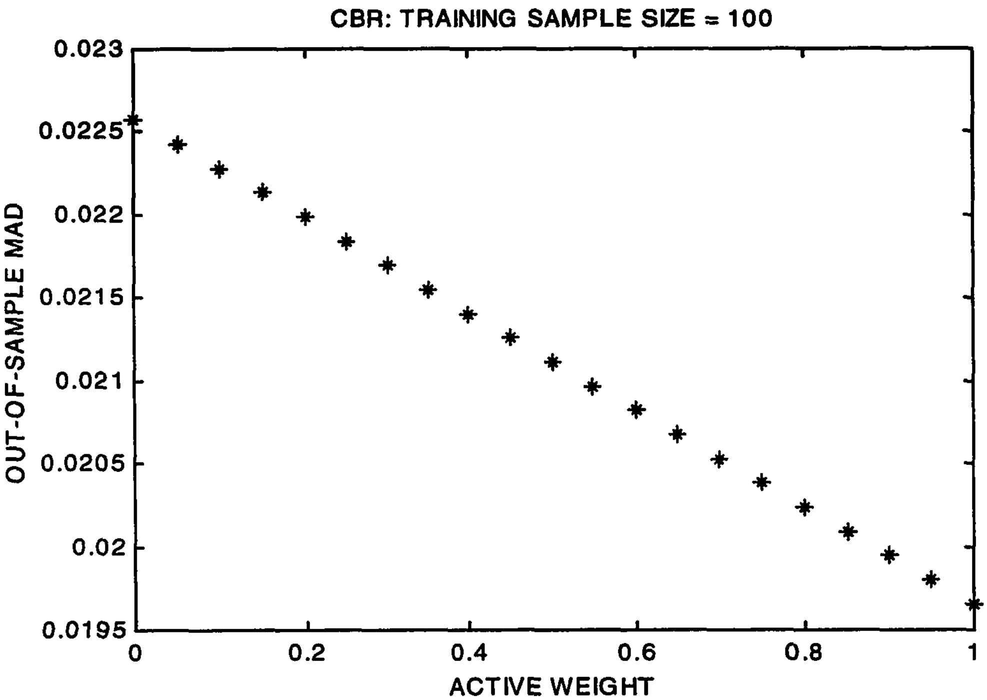


Figure 1.6.5 Out-of-Sample Prediction (CBR, Training Sample Size = 100)

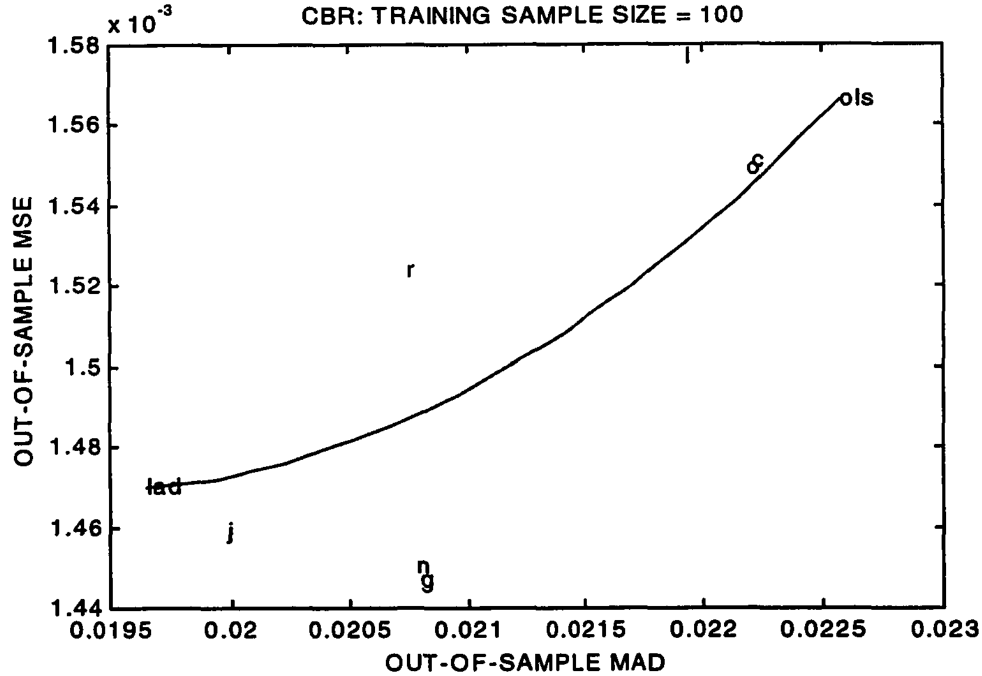


Figure 1.6.6 Out-of-Sample MSE (CDR, Training Sample Size = 100)

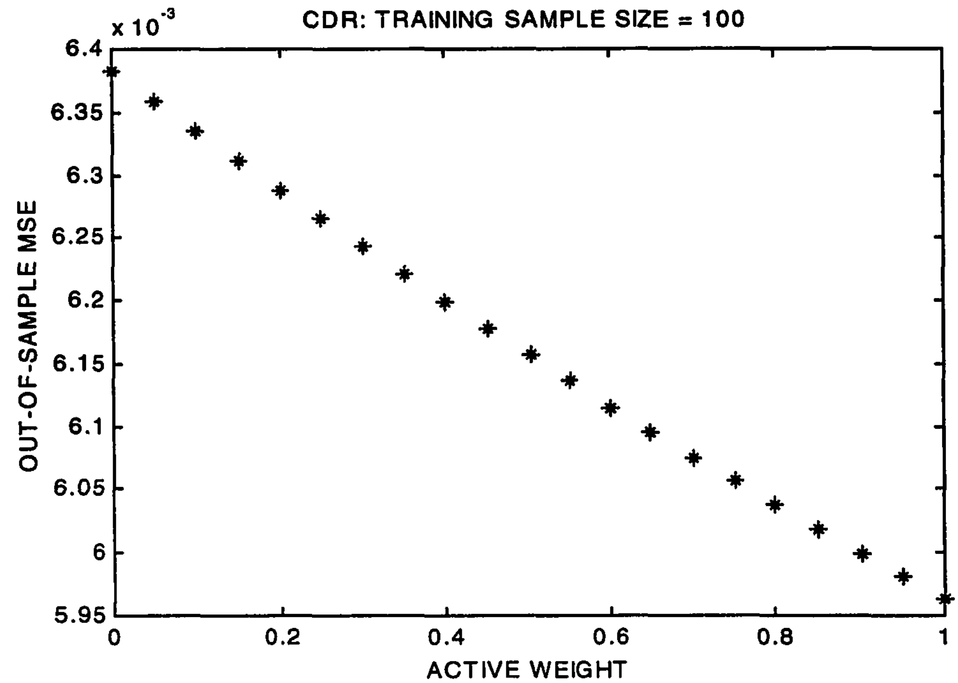


Figure 1.6.7 Out-of-Sample MAD (CDR, Training Sample Size = 100)

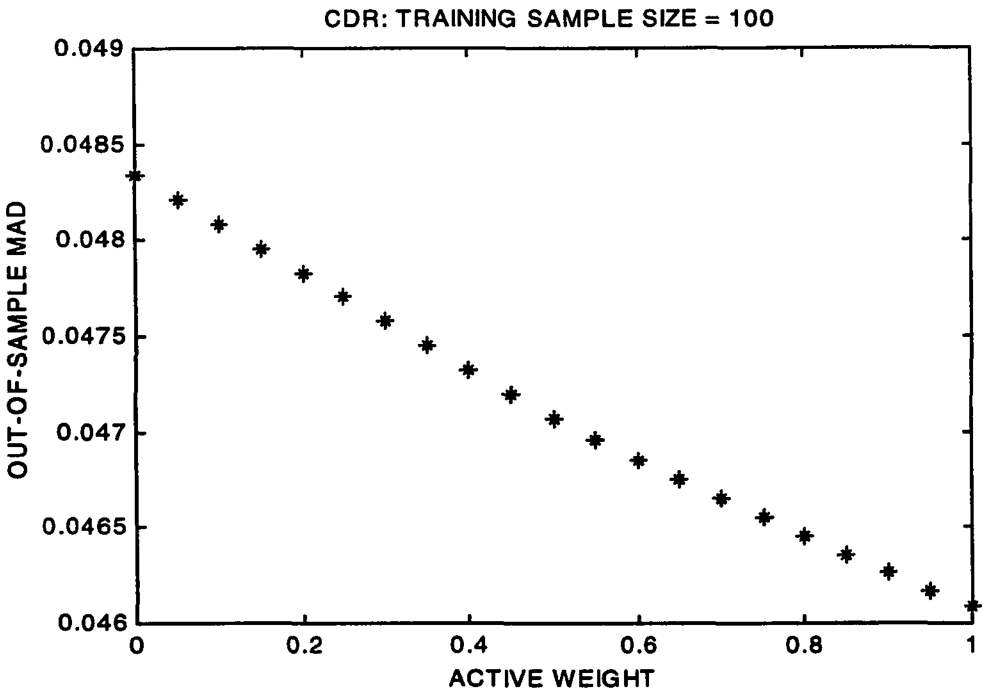


Figure 1.6.8 Out-of-Sample Prediction (CDR, Training Sample Size = 100)

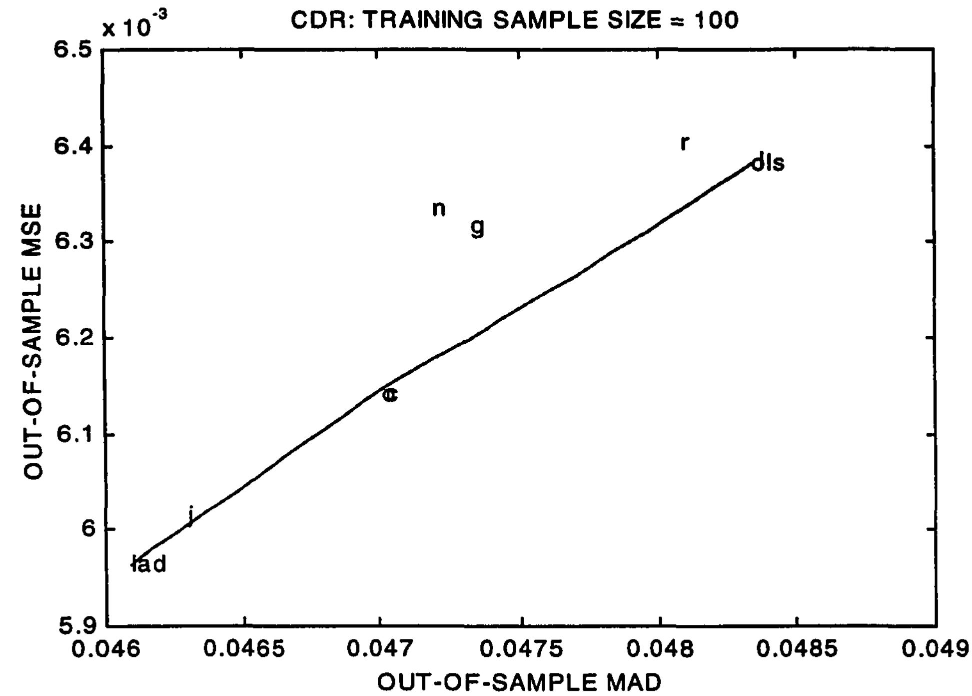


Figure 1.6.9 Out-of-Sample MSE (ADC, Training Sample Size = 520)

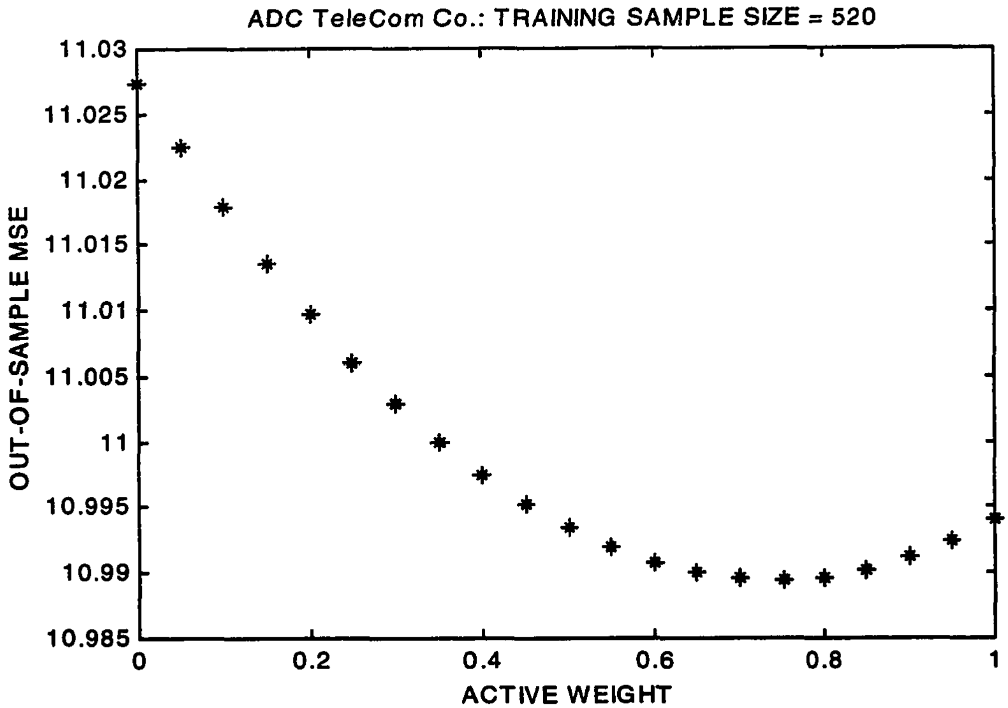


Figure 1.6.10 Out-of-Sample MAD (ADC, Training Sample Size = 520)

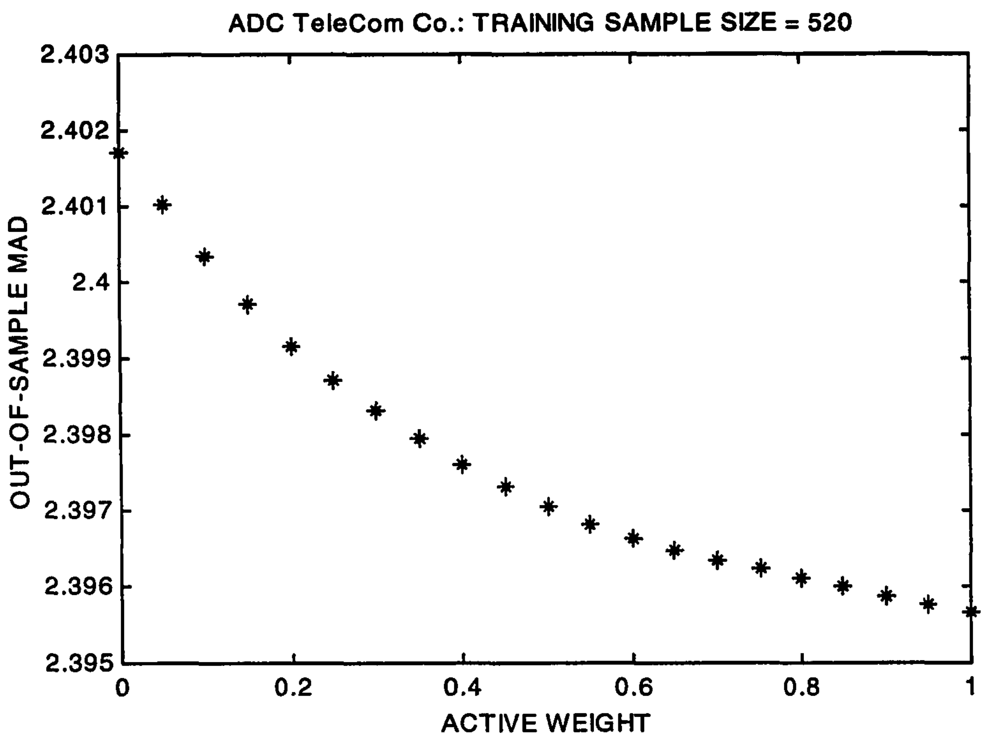


Figure 1.6.11 Out-of-Sample Prediction (ADC, Training Sample Size = 520)

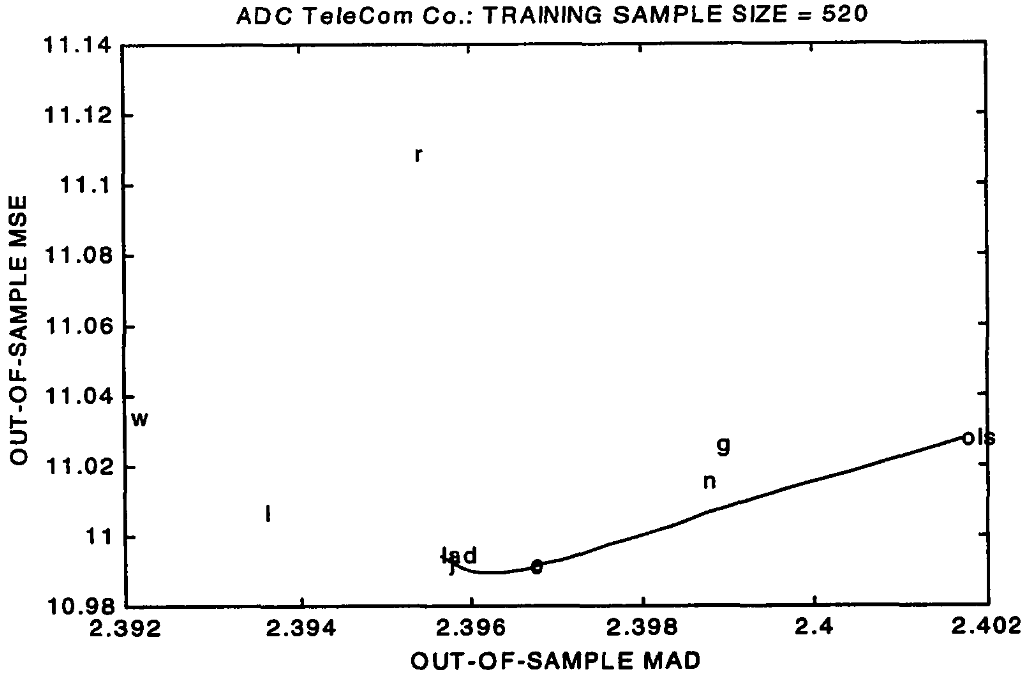


Figure 1.6.12 Out-of-Sample MSE (HomeStake, Training Sample Size = 520)

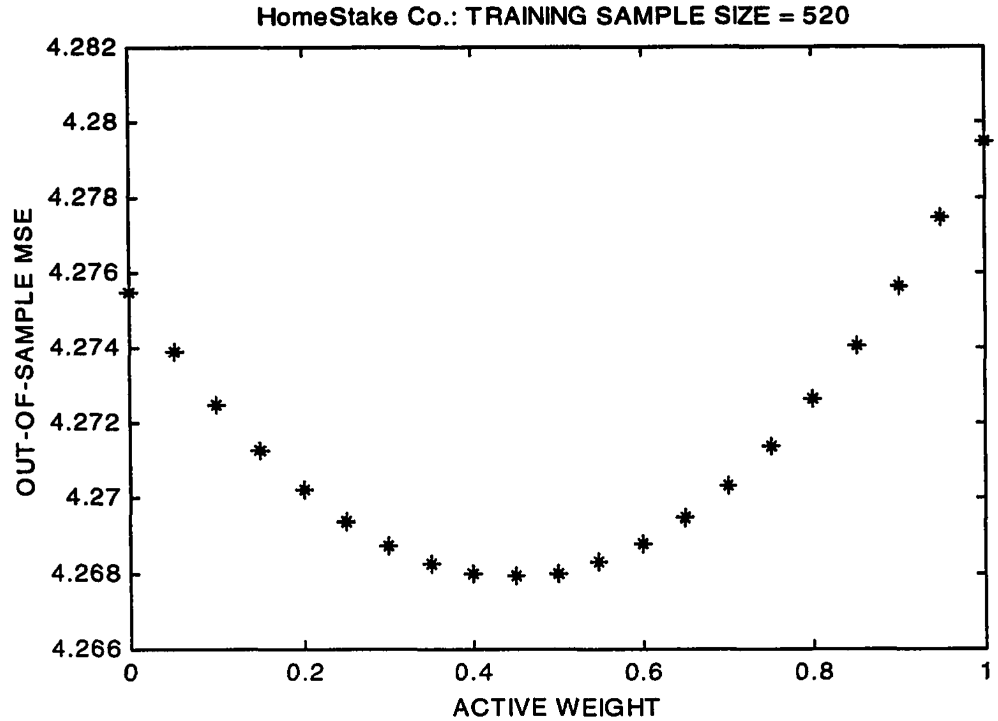


Figure 1.6.13 Out-of-Sample MAD (HomeStake, Training Sample Size = 520)

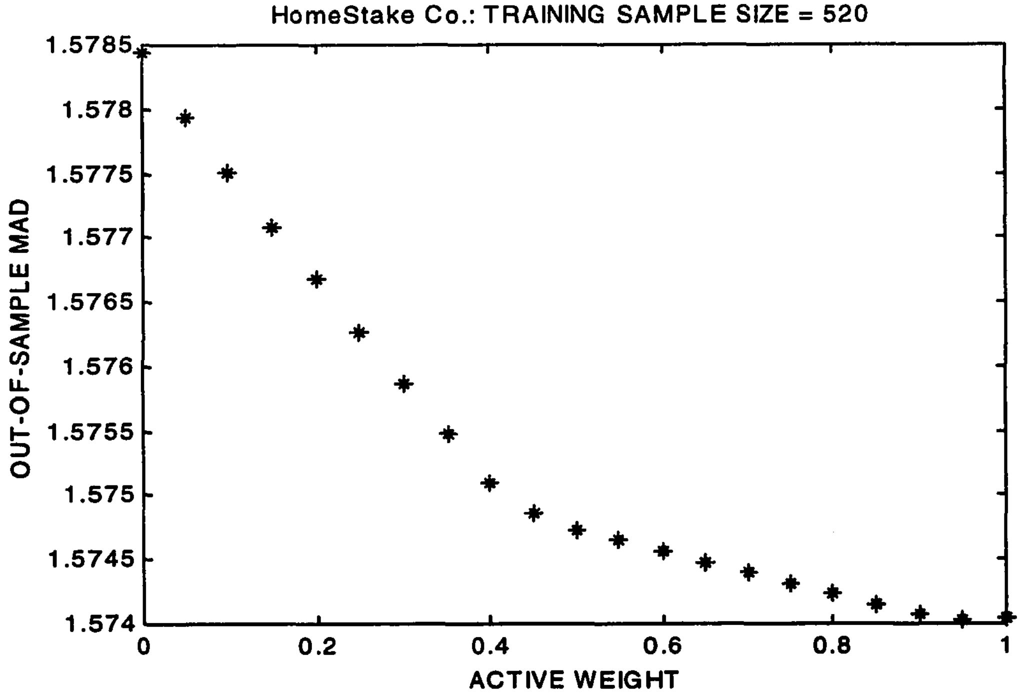


Figure 1.6.14 Out-of-Sample Prediction (HomeStake, Training Sample Size = 520)

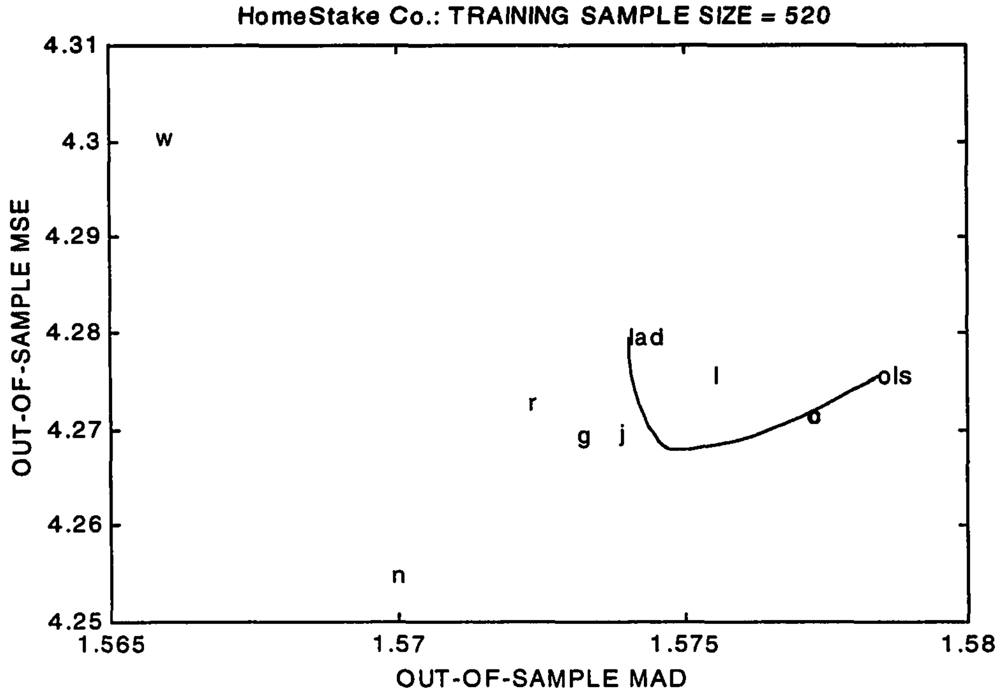
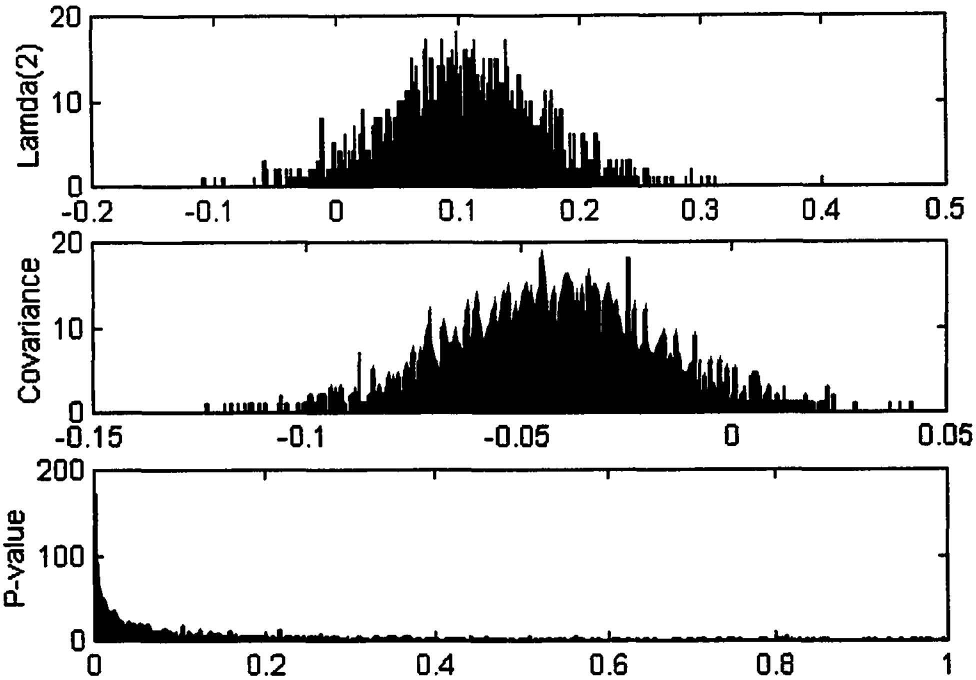


Figure A Distribution of λ_2, Covariance and P-value



Chapter 2

Scaling Estimation of the Shrinkage Least Absolute Deviation Estimator using a Bootstrap Approach

2.1 Introduction

One drawback of the shrinkage estimator generally and in particular those of Chapter 1 is the lack of a measure of precision. Bootstrapping the standard errors or the confidence intervals is one solution for this problem.

Delaney and Chatterjee (1986) combine the bootstrap and cross-validation to estimate the ridge parameter. Vinod and Raj (1988) apply the bootstrapping method to a shrinkage estimator (a ridge estimator) to investigate the economic issues in Bell System divestiture. They identify a bootstrap problem: a lack of a pivot[1] for the ridge estimator. For the ridge estimator denoted by b_λ where λ is the ridge parameter, even the distribution of b_λ-β depends on β. Brownstone (1990) bootstraps two improved estimators: Mundlak's restricted principle-components estimator and a Stein-rule estimator which shrinks the OLS estimator toward Mundlak's estimator. He uses non-pivotal statistics, i.e., the percentile method to get the sampling distributions. He shows that the non-parametric bootstrap provides a good estimate of the estimator's risk and standard errors. Vinod (1995) provides a solution to the non-pivotal problem for ridge regression by applying Beran's double bootstrap. This method involves a bootstrap within a bootstrap which is computationally intensive.

We provide a way to obtain the sampling distributions and confidence intervals for shrinkage estimators with a fixed and random guess based on a bootstrapping method. The asymptotic normality approximation does not apply because the limiting distribution of the shrinkage estimator is not a standard normal but is a nonlinear function of a normal random variable. It is well known that in order to get a better bootstrap confidence interval, one should use pivotal or asymptotically pivotal statistics. The importance of using pivotal statistics have been intensively discussed in bootstrap literature. See Hartigan (1986), Beran (1987), Hall (1987), Hall (1988). For general shrinkage

[1] "A function T of both the data and an unknown parameter is said to be pivotal if it has the same distribution for all values of the unknowns. It is asymptotically pivotal if, for sequences of known constants $\{a_n\}$ and $\{b_n\}$, a_nT+b_n has a proper non-degenerate limiting distribution not depending on unknowns." This definition is taken from Hall (1992) pp 14.

estimators other than ridge estimators, it is unlikely that the necessary scaling estimators are available.

The basic idea is to use the Lemma 1 and Lemma 2 in Ullah (1990) to derive the first moment and the second moment. Ullah (1990) proves two lemmas which allow us to obtain the moments of the function of normal random variables by taking derivatives and to easily calculate the moments of the ratio of quadratic normal random variables. By taking derivatives of the limiting random variable of the shrinkage estimator, we calculate the asymptotic variance of the shrinkage estimator. We use a consistent estimator for this asymptotic variance to obtain the bootstrapping pivotal statistics.

2.2 Asymptotic Moments: James-Stein Estimator with Non-Random Guess

First we derive the asymptotic moments of the shrinkage estimator with a non-random guess. Consider $y = X\beta^0 + \varepsilon$ where $\beta^0 \in R^k$, $k > 2$ and ε is iid random vector. Let b_n be an estimator satisfying that $n^{1/2}(b_n-\beta^0) \xrightarrow{d} N(0,A)$. We define the JS estimator, $\delta^{JS}(b_n,g_n)$,

$$\delta^{JS}(b_n,g_n) = \left(1-\frac{\lambda}{(b_n-g_n)'Q_n(b_n-g_n)}\right)(b - g_n) + g_n$$

where $\lambda \in (0, 2(k-2))$ and g_n is a non-random guess such that $n^{1/2}(g_n-\beta^0)$ converges to a fixed vector, θ which is called "finite sample guess bias". In most cases, the optimal λ is chosen to be k-2. We have the following result.

$$n^{1/2}(\delta^{JS}(b_n,g_n)-\beta^0) \xrightarrow{d} \left(1-\frac{\lambda}{U'QU}\right)U+\theta$$

where $n^{1/2}(b_n-g_n) \xrightarrow{d} U \sim N(-\theta,A)$ and $n^{-1}Q_n \xrightarrow{p} Q$.

Define $h(U) = \left(1-\frac{\lambda}{U'QU}\right)U + \theta$. Since A is assumed to be symmetric and positive definite, there exists a matrix, P such that $A = PP'$. Let $\mu = -P^{-1}\theta$ and $z = P^{-1}U$. Then z is normally distributed with mean vector, μ and identity covariance matrix. Define H(z) as follows.

$$H(z) = \left[\left(1 - \frac{\lambda}{z'z}\right)z - \mu\right]$$

Some simple algebra shows that h(U) = P'H(z). Note that E(h(U) = PE(H(z)) and Var(h(U)) = PVar(H(z))P'. In the following we will give explicit expression of E(H(z)) and Var(H(z)).

Theorem 2-1 Asymptotic First Moment of Shrinkage Estimator with Non-random Guess

Let $H_i(z)$ be the ith component of H(z), ie $H_i(z) = (1-\lambda/z'z)e_iz-e_i\mu$ where e_i is a k-dimensional vector whose elements are all zeros except the ith element being one. Then under some regularity conditions which allow us to exchange limit and integral, we have the following:

$$E(H_i(z)) = -\lambda\mu_iW(\mu) - \lambda\Gamma(1)^{-1}\int_0^\infty (1+2t)^{-k/2}\exp(-(t/1+2t)\mu'\mu)(-(2t/1+2t)\mu_i)dt$$

where $W(\mu) = \int_0^\infty (1+2t)^{-k/2}\exp(-(t/1+2t)\mu'\mu)dt$.

Proof: See Appendix.

It is well known that most shrinkage estimators are not unbiased. As Theorem 2-1 shows, the expectation of the limiting random variable of the centered shrinkage estimator with a non-random guess is not equal to zero. This means that shrinkage estimators with a non-random guess are not even asymptotically unbiased.

Theorem 2-2 Asymptotic Second Moment of Shrinkage Estimator with Non-random Guess

Let v_{ij} = (i,j) element of E(H(z)H(z)'). Then under the same conditions as in the Theorem 2-1 we have the following.

$$v_{ij} = a_{ij} - (b_{ij} - c_{ij}) - (b_{ji} - c_{ji}) + d_{ij}$$

where (1) $a_{ij} = 1$ if $i = j$ and $a_{ij} = 0$ otherwise.

$$(2)\ b_{ij} = \lambda\Gamma(1)^{-1}\int_0^\infty (1+2t)^{-k/2}(tr(I_{ij}N_{0t}^{-1}) + \mu' N_{2t}\mu \exp(-1/2\mu' N_{1t}\mu)dt$$

where $N_{0t} = (1+2t)I$

$N_{1t} = 2t/(1+2t)I$

$N_{2t} = (1/1+2t)^2 I_{ij}$

I_{ij} = zero matrix except (i,j) element being equal to one.

$$(3)\ c_{ij} = \lambda\mu_i\{\mu_j W(\mu) + \Gamma(1)^1\int_0^\infty (1+2t)^{-k/2}\exp(-(t/1+2t)\mu'\mu)(-(2t/1+2t)\mu_i)dt\}.$$

$$(4)\ d_{ij} = \lambda^2\Gamma(2)^{-1}\int_0^\infty t\det(N_{0t})^{-k/2}(tr(I_{ij}N_{0t}^{-1}) + \mu' N_{2t}\mu \exp(-1/2\mu' N_{1t}\mu)dt.$$

2.3 Asymptotic Moments: Optimal Weighting Scheme Estimator

There is one potential problem to be considered, which is the non location-scale distribution problem. Typically the distribution of the limiting random variable for the shrinkage estimator is not a member of a location-scale family. If the distribution is from the location-scale family, then the usual studentization is enough to make a pivotal statistics. When this is not the case, Babu and Singh's (1983) result that the bootstrap estimates the true sampling distribution up to a second-order term nevertheless justifies studentization for non location-scale families.

Using the same assumptions and notations in the Chapter 1, we have the following.

$$n^{1/2}(\delta^{OW}(b_n,g_n)-\beta^0) \xrightarrow{d} \left(1-\lambda_1-\frac{\lambda_2}{(U_1-U_2)'Q(U_1-U_2)}\right)(U_1-U_2) + U_2 \equiv h(U)$$

where $U \equiv \begin{bmatrix} U_1 \\ U_2 \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} A & \Delta \\ \Delta' & B \end{bmatrix}\right)$. Also let $\mu = \begin{bmatrix} 0_{k\times l} \\ 0_{k\times 1} \end{bmatrix}$ and $\Sigma = \begin{bmatrix} A & \Delta \\ \Delta' & B \end{bmatrix}$. Since Σ is assumed to be symmetric and positive definite, there exists a matrix, P such that $\Sigma = PP'$.

Let $z = P^{-1}U$. Then z is normally distributed with mean vector μ and identity covariance matrix. The limiting random variable, h(U), can be rewritten as

$$h(U) = \left(1-\lambda_1-\frac{\lambda_2}{U'J_1'QJ_1U}\right)J_1U + J_2U$$

where $J_1 = [I_k \ {-I_k}]$ and $J_2 = [0_k \ I_k]$. Define H(z) as follows.

$$H(z) \equiv \left(1-\lambda_1-\frac{\lambda_2}{z'P'J_1'QJ_1Pz}\right)J_1Pz + J_2Pz.$$

Some simple algebra shows that h(U) = H(z). Let $M_1 = P'J_1'QJ_1P$, $M_2 = J_1P$, and $M_3 = J_2P$. Then we can further simplify H(z).

$$H(z) = Mz - \left[\frac{\lambda_2 M_2 z}{z' M_1 z}\right]$$

where $M = (1-\lambda_1)M_2 + M_3$.

Theorem 3-1 Asymptotic First Moment of the OWS estimator

Let $H_i(z)$ be the ith component of H(z). Then under some regularity conditions which allow us to exchange limit and integral, we have the following:

$$E(H_i(z)) = 0.$$

Proof: See Appendix.

This result shows that even though shrinkage estimators with a random guess could be biased in small sample, we can obtain the unbiasedness in the limit as n goes to infinity. In other words, shrinkage estimators with a random guess are asymptotically unbiased.

Theorem 3-2 Asymptotic Second Moment of OWS estimator

Let v_{ij} = (i,j) element of E(H(z)H(z)'). Then under the same conditions as in the Theorem 3-1 we have the following:

$$v_{ij} = a_{ij} - b_{ij} - c_{ij} + d_{ij}$$

where (1) a_{ij} = (i,j) element of MM'.

(2) $b_{ij} = \lambda_2 \sum_{m=1}^{2k}\sum_{n=1}^{2k} M_{im} E_{mn} M_{2jn}$

where $E_{mn} = \Gamma(1)^{-1} \int_0^{\infty} |N_{0t}|^{-1/2}\, tr(I_{ij} N_{0t}^{-1})dt$ and $N_{0t} = I + 2tM_1$.

(3) $c_{ij} = b_{ji}$.

(4) $d_{ij} = \lambda_2^2 \sum_{m=1}^{2k}\sum_{n=1}^{2k} M_{2im} F_{mn} M_{2jn}$

where $F_{mn} = \Gamma(2)^{-1} \int_0^{\infty} t|N_{0t}|^{-1/2}\, tr(I_{ij} N_{0t}^{-1})dt$ and $N_{0t} = I + 2tM_1$.

Proof: See Appendix.

2.4 Covariance Estimation

In order to bootstrap the JS estimator with a random guess, we need to have an estimator for the covariance matrix between the base estimator and the guess. We provide an example of the covariance estimator in a special case where the base estimator is the LAD estimator and the guess is the OLS estimator.

As shown in Bates and White (1993), both the LAD estimator and the OLS estimator are members in RCASOI (Regular Consistent Asymptotically Second Order Indexed) class under some regularity conditions. For any member, b_n, in RCASOI class, there is a "score" representation (s^0_n) and "Hessian" representation (H^0_n) such that

$$b_n - \beta^0 = H^{0}_{n}{}^{-1} s^0_n + o_p(n^{-1/2}).$$

Accordingly, we have the following representation for the two estimators.

$$s_n^{LS} = 2\sum_{t=1}^{n} X_t \varepsilon_t . \qquad H_n^{LS} = 2\sum_{t=1}^{n} E(X_t X_t') .$$

$$s_n^{LAD} = -2\sum_{t=1}^{n} X_t (1_{[\varepsilon_t \le 0]} - 1/2) . \quad H_n^{LAD} = 2f(0)\sum_{t=1}^{n} E(X_t X_t') .$$

where $u_t \equiv y_t - X_t'\beta^0$ and f(0) is the value of the density of u_t at zero. Note that the asymptotic covariance can be approximated by

$$\mathrm{Cov}[n^{1/2}(b_n-\beta^0),\, n^{1/2}(g_n-\beta^0)]$$
$$= E[n^{1/2}(b_n-\beta^0)(g_n-\beta^0)'n^{1/2}]$$
$$= E[(n^{-1} H_n^{LAD})^{-1} n^{-1/2} s_n^{LAD} n^{-1/2} s_n^{LS\prime}(n^{-1} H_n^{LS})^{-1}]$$
$$= (n^{-1} H_n^{LAD})^{-1} E(s_n^{LAD} s_n^{LS\prime}/n)(n^{-1} H_n^{LS})^{-1}.$$

We can simplify each term as follows.

(1) $n^{-1} H_n^{LAD} = 2f(0)E(X_tX_t')$

(2) $E(s_n^{LAD}s_n^{LS}/n) = E(S_{1t}S_{2t}')$

where $S_{1t} = -2X_t(1[\varepsilon_t \leq 0] - ½)$

$S_{2t} = 2X_t\varepsilon_t$

(3) $n^{-1} H_n^{LS} = 2E(X_tX_t')$.

Theorem 4-1 Consistent Estimator for the Asymptotic Covariacne

Suppose

(i) $\{y_t, X_t\}$ is iid.

(ii) $\mathrm{tr}E(X_tX_t') < \infty$

(iii) $\hat{f}(0) \xrightarrow{p} f(0)$

(iv) $E(S_{it}S_{2t}') < \infty$

Then $[2f(0)n^{-1}\sum_{t=1}^{n} X_tX_t']^{-1}[n^{-1}\sum_{t=1}^{n} S_{1t}S_{2t}'][2n^{-1}\sum_{t=1}^{n} X_tX_t']^{-1} \xrightarrow{p} \Delta$.

Proof: straightforward using the law of large number.

2.5 Application

In this section, we show how the method developed in previous sections can be implemented. We choose the LAD estimator as the base estimator and the OLS estimator as the guess. Bootstrapping the LAD estimator itself is not new in the literature. Hahn (1995) bootstraps the quantile estimator and shows that the bootstrap distribution converges weakly to the limiting distribution of the quantile estimator. However bootstrapping a shrinkage LAD estimator has not previously been studied, to our knowledge. We use a simulated data set instead of real data so that we can evaluate performance in terms of the true parameters.

An artificial data set is generated as $y = X\beta^0 + \varepsilon$ where $\varepsilon \in R^n$, $\beta^0 \in R^k$, n = 80 and k = 3. We set $\beta^0 = 0$. We choose the standard normal distribution for the error distribution. Each row of X is generated by the joint normal distribution[2], $N(1,\Sigma)$, where covariances are all 0.8 and variances are one. Once one set of data is generated we consider it as our original data set and pretend not to know the true value of β^0. The OWS estimator, δ_{ni}, and its asymptotic standard deviation, s_i, are computed using the original data set.

We consider the equal tail percentile-t method (studentized), equal tail percentile method (unstudentized), naïve percentile method and "normal approximation" method for constructing confidence intervals for β^0. For the percentile-t method[3], the population equation is given by

$$\text{Prob}[t_L^P < n^{1/2}(\delta_{ni} - \beta^0{}_i)/s_i < t_U^P] = 1\text{-}\alpha.$$

The ideal $(1\text{-}\alpha)\%$ confidence interval is (t_L^P, t_U^P), but we cannot obtain this because the exact distribution of δ_{ni} is not known. This population equation can be approximated by the following sample equation.

$$\text{Prob}[t_L^s < n^{1/2}(\delta_{ni}^* - \delta_{ni})/s_i^* < t_U^s] = 1\text{-}\alpha.$$

where δ_{ni}^*, s_i^* are bootstrap estimates. Even though the sample equation can be solved in principle, in most cases it is intractable to solve it analytically because the empirical distribution from which the bootstrap re-sampling is taken is not continuous. Hence we approximate the solution to the sample equation using bootstrap re-sampling. By

---

[2] We use the multivariate normal random vector generator, DNRVG, which is a FORTRAN subroutine in NSWC (Naval Surface Warfare Center) Library. We set the seed to be 3833981.

[3] For the percentile method the population equation is $\text{Prob}[t_L^P < \delta_{ni} - \beta^0{}_i < t_U^P] = 1\text{-}\alpha$ and the sample equation is $\text{Prob}[t_L^s < (\delta_{ni}^* - \delta_{ni}) < t_U^s] = 1\text{-}\alpha$. We compute the naïve percentile confidence intervals by taking $\alpha/2$ and $(1\text{-}\alpha/2)$ percentile of $\{\delta_{ni}^*, i = 1, .. B\}$. The formula for the normal approximation interval is $[\delta_{ni} - 1.96s_i/n^{1/2}, \delta_{ni} + 1.96s_i/n^{1/2}]$.

bootstrapping (y_t, X_t) pairs[4], we generate $\{\gamma_i : i = 1, \dots, B\}$ where $\gamma_i = n^{1/2}(\delta_{ni}^* - \delta_{ni})/s_i^*$. We take the $\alpha/2$ percentile ($\hat{t}_L^s$) and (1-α/2) percentile ($\hat{t}_U^s$) as an approximation for t_L^s and t_U^s respectively. The (1-α)% bootstrap confidence interval is given by

$$[\delta_{ni} - \hat{t}_U^s s_i/n^{1/2}, \delta_{ni} - \hat{t}_L^s s_i/n^{1/2}].$$

Table 2.5.1 summarizes the regression results using the original data set. The LAD estimates have uniformly higher standard errors than the OLS estimates as we should expect, since the OLS estimator is the Maximum Likelihood estimator in this experiment. Interestingly the OWS estimates have smaller standard errors than the LAD estimates as well as the OLS estimates' standard errors even though theoretically we expect both the OLS and OWS estimators to have the same standard errors.

Table 2.5.2 shows the bootstrap 95% confidence intervals and some descriptive statistics about them[5]. The percentile-t bootstrap confidence intervals are [-.243, .434], [-.336, .296], and [-.349, .406]. They cover the true β^0 correctly. Note that they are not symmetric intervals according to the 'shape' statistics. The non-symmetry can be visualized by the histograms of the standardized bootstrap estimates. In many cases, the enforced symmetry based on a limiting distribution argument causes size distortion. The asymmetric property can be considered as an advantage of using bootstrapping method. The percentile confidence intervals are fairly comparable to the percentile-t confidence intervals. They are also not symmetric and a little bit shorter. The naïve percentile intervals have the exactly same length as the percentile intervals as expected because percentile intervals are a shifted version of the percentile intervals. Even though the limiting distribution is not normal, we include a "normal approximation" for comparison.

---

[4] Brownstone (1990) and Vinod (1995) bootstrap residuals. We prefer bootstrapping pairs because it is more robust to assumptions on the error term. See Efron (1993). Also we use an iid bootstrap method. For a dependent stationary data, the stationary bootstrap proposed by Politis and Romano can be used.

[5] A bootstrap iteration takes 2.46 seconds using FORTRAN program on a PC with 75 MHz Pentium processor. One reason for the long computation time is that we have to compute integrals over infinite intervals 72 times per iteration. See Theorem 3-2 in section 2.3. Since we have 3 regressors, the number of $\{E_{ij}\}$, $\{F_{ij}\}$ where i,j = 1, . . . , 6 is 72. We use the DQAGI subroutine in NSWC Library which allows us to compute the infinite integrals. We provide a graph showing what a typical integrand in E_{ij} and F_{ij} looks like.

The intervals are computed as if the distribution is close to normal. In order to evaluate each method in terms of coverage error, we need to show it analytically or run some Monte Carlo experiment, which has not been done in this paper. We leave this for further consideration.

2.6 Conclusion

We have used Ullah's Lemmas to derive the asymptotic moments of the shrinkage estimator with non-random guess and random guess. It has been shown that using a consistent estimator for the asymptotic moments, a non-parametric bootstrap pivotal statistic can be constructed. In order to illustrate the practical implementation, we have applied the result to the OWS estimator to compute the Bootstrap confidence intervals.

Appendix

Proof of Theorem 2-1 Asymptotic First Moment of James-Stein Estimator with Non-random Guess

Let $f(z) \equiv z - \mu$ and $g(z) \equiv (\lambda/z'z)z$. Let $f_i(z)$ and $g_i(z)$ be the ith element of f(z) and g(z) respectively where $f_i(z) = e_i z - e_i\mu$ and $g_i(z) = (\lambda/z'z)e_i z$. Then $H_i(z) = f_i(z) - g_i(z)$.

(1) $E(f_i(z)) = e_i E(z) - e_i\mu = e_i\mu - e_i\mu = 0$.

(2) Define $g_{1i}(z) = e_i'z$ and $g_2(z) = \lambda/z'z$. Then $g_i(z) = g_{1i}(z)g_2(z)$.

By using Ullah's differential operator, $d \equiv [\mu + \frac{\partial}{\partial\mu}]$,

$$g_{1i}(d) = e_i'd = e_i'[\mu + \frac{\partial}{\partial\mu}] \equiv \mu_i + \frac{\partial}{\partial\mu_i}.$$

On the other hand, we can show using Ullah's Lemma2 that

$$E(g_2(z)) = \lambda\Gamma(1)^{-1}\int_0^\infty |N_{0t}|^{-1/2}\exp(-1/2\mu' N_{1t}\mu)dt$$

where (1) $N_{0t} = (1+2t)I$ and $|N_{0t}| = (1+2t)^k$

(2) $N_{1t} = 2t/(1+2t)I$

(3) $\Gamma(\ )$ is the gamma function.

$$= \lambda\Gamma(1)^{-1}\int_0^\infty (1+2t)^{-k/2}\exp(-(t/1+2t)\mu'\mu)dt.$$

Define $W(\mu) = \Gamma(1)^{-1}\int_0^\infty (1+2t)^{-k/2}\exp(-(t/1+2t)\mu'\mu)dt$.

Therefore we have the following.

$$\begin{aligned} E(g_i(z)) &= E(g_i^1(z)g_i^2(z)) \\ &= g_i^1(d)E(g_i^2(z)) \\ &= [\mu_i + \frac{\partial}{\partial\mu_i}]\lambda\Gamma(1)^{-1}\int_0^\infty (1+2t)^{-k/2}\exp(-(t/1+2t)\mu'\mu)dt \\ &= \lambda\mu_i W(\mu) + \lambda\Gamma(1)^{-1}\int_0^\infty (1+2t)^{-k/2}\exp(-(t/1+2t)\mu'\mu)(-(2t/1+2t)\mu_i)dt. \end{aligned}$$

Hence $E(H_i(z)) = -E(g_i(z))$

$$= -\lambda\mu_i W(\mu) - \lambda\Gamma(1)^{-1}\int_0^\infty (1+2t)^{-k/2}\exp(-(t/1+2t)\mu'\mu)(-(2t/1+2t)\mu_i)dt. \quad \text{Q.E.D}$$

Proof of Theorem 2-2 Asymptotic Second Moment of James-Stein Estimator with Non-random Guess

Note that $H(z)H(z)' = f(z)f(z)' - f(z)g(z)' - g(z)f(z)' + g(z)g(z)'$.

(1) $E(f(z)f(z)') = E((z-\mu)(z-\mu)') = Var(z) = I$.

Hence $a_{ij} = 1$ if $i = j$ and $a_{ij} = 0$ otherwise.

(2) Note that $f(z)g(z)' = \lambda(zz'/z'z - \mu z'/z'z)$.

Let b_{ij} be the (i,j)th element of $\lambda E(zz'/zz)$. Then using Lemma 2

$$\begin{aligned} b_{ij} &= \lambda E(z'I_{ij}z/z'z) \\ &= \lambda\Gamma(1)^{-1}\int_0^\infty (1+2t)^{-k/2}(tr(I_{ij}N_{0t}^{-1}) + \mu' N_{2t}\mu\exp(-1/2\mu' N_{1t}\mu)dt \end{aligned}$$

where $N_{0t} = (1+2t)I$

$N_{1t} = 2t/(1+2t)I$

$N_{2t} = (1/1+2t)^2 I_{ij}$

I_{ij} = zero matrix except (i,j) element being equal to one.

Let c_{ij} be the (i,j)th element of $aE(\mu z'/zz)$. Then

$$\begin{aligned} c_{ij} &= \lambda E(\mu_i e_j'z/z'z) \\ &= \mu_i E(\lambda e_j'z/z'z) \\ &= \mu_i E(g_j(z)) \\ &= \mu_i\{\lambda\mu_j W(\mu) + \lambda\Gamma(1)^{-1}\int_0^\infty (1+2t)^{-k/2}\exp(-(t/1+2t)\mu'\mu)(-(2t/1+2t)\mu_j)dt\} \end{aligned}$$

using the result in the proof of Theorem 2-1.

Hence the (i,j) element of $E(f(z)g(z)')$ is given by $b_{ij} - c_{ij}$.

(3) By symmetry, the (i,j) element of $E(g(z)f(z)')$ is given by $b_{ji} - c_{ji}$.

(4) Let d_{ij} be the (i,j)th element of $E(g(z)g(z)')$. Then

$d_{ij} = \lambda^2 E(z'I_{ij}z/(z'z)^2)$

$$= \lambda^2\Gamma(2)^{-1}\int_0^\infty t\det(N_{0t})^{-k/2}(tr(I_{ij}N_{0t}^{-1}) + \mu' N_{2t}\mu)\exp(-1/2\mu' N_{1t}\mu)dt$$ using Ullah's Lemma 2. Q.E.D

Proof of Theorem 3-1 Asymptotic First Moment of OWS Estimator

Let $f(z) \equiv Mz$ and $g(z) \equiv \left[\frac{\lambda_2 M_2 z}{z' M_1 z}\right]$. Let $f_i(z)$ and $g_i(z)$ be the ith element of f(z) and g(z) respectively. Then $H_i(z) = f_i(z) - g_i(z)$.

(1) $E(f_i(z)) = E(M_i z) = M_i E(z) = 0$ where M_i is the ith row of M.

(2) Define $g_{1i}(z) = M_{2i}z$ and $g_2(z) \equiv \left[\frac{\lambda_2}{z' M_1 z}\right]$. Then $g_i(z) = g_{1i}(z)g_2(z)$.

By definition, $g_{1i}(d) \equiv M_{2i}d \equiv M_{2i}[\mu + \frac{\partial}{\partial\mu}]$ where M_{2i} is the ith row of M_2.

$$E(g_2(z)) = \lambda_2\Gamma(1)^{-1}\int_0^\infty |N_{0t}|^{-1/2}\exp(-1/2\mu' N_{1t}\mu)dt$$

where (1) $N_{0t} = I + 2tM_1$

(2) $N_{1t} = 2tM_1N_{0t}^{-1}$

(3) $\Gamma(\ )$ is the gamma function.

Define $W(\mu) = \Gamma(1)^{-1}\int_0^\infty |N_{0t}|^{-1/2}\exp(-1/2\mu' N_{1t}\mu)dt$.

Therefore $E(g_i(z)) = E(g_i^1(z)g_i^2(z))$

$= g_i^1(d)E(g_i^2(z))$

$= M_{2i}[\mu + \frac{\partial}{\partial\mu}]a\, W(\mu)|_{\mu=0} = 0.$

Note that $\frac{\partial}{\partial\mu}W(\mu)|_{\mu=0} = \Gamma(1)^{-1}\int_0^\infty |N_{0t}|^{-1/2}\exp(-1/2\mu' N_{1t}\mu)(-N_{1t}\mu)dt|_{\mu=0} = 0.$

Hence $E(H_i(z)) = 0$. Q.E.D

Proof of Theorem 3-2 Asymptotic Second Moment of OWS Estimator

Note that $H(z)H(z)' = f(z)f(z)' - f(z)g(z)' - g(z)f(z)' + g(z)g(z)'$.

(1) $E(f(z)f(z)') = E(Mzz'M') = ME(zz')M' = MM'$.

Hence a_{ij} = (i,j) element of MM'.

(2) $E(f(z)g(z)') = E(Mz\left[\frac{\lambda_2 M_2 z}{z' M_1 z}\right]') = \lambda_2 E\left[\frac{Mzz'M_2'}{z'M_1z}\right] = \lambda_2 ME\left[\frac{zz'}{z'M_1z}\right]M_2'$.

Let E_{mn} be the (m,n) element of $E\left[\frac{zz'}{z'M_1z}\right]$. Then E_{mn} can be computed using Ullah's lemmas as follows.

$$E_{mn} = E\left[\frac{z'I_{mn}z}{z'M_1z}\right] \text{ where } I_{mn} \text{ is defined as before.}$$

$$= \Gamma(1)^{-1}\int_0^\infty |N_{0t}|^{-1/2}(tr(I_{mn}N_{0t}^{-1}) + \mu' N_{2t}\mu)\exp(-1/2\mu' N_{1t}\mu)dt\,|_{\mu=0}.$$

where $N_{0t} = I+2tM_1$, $N_{1t} = 2tM_1N_{0t}^{-1}$, and $N_{2t} = N_{0t}^{-1}I_{mn}N_{0t}^{-1}$.

$$= \Gamma(1)^{-1}\int_0^\infty |N_{0t}|^{-1/2}\, tr(I_{mn}N_{0t}^{-1})dt.$$

Hence, $b_{ij} = a\sum_{m=1}^{2k}\sum_{n=1}^{2k} M_{im}E_{mn}M_{2jn}$.

(3) By symmetry, $c_{ij} = b_{ji}$.

(4) $E(g(z)g(z)') = E(\left[\frac{\lambda_2 M_2 z}{z' M_1 z}\right]\left[\frac{\lambda_2 M_2 z}{z' M_1 z}\right]') = \lambda_2^2 E\left[\frac{M_2 zz' M_2'}{(z'M_1z)^2}\right] = \lambda_2^2 M_2 E\left[\frac{zz'}{(z'M_1z)^2}\right]M_2'.$

Let F_{mn} be the (m,n) element of $E\left[\frac{zz'}{(z'M_1z)^2}\right]$. Then E_{mn} can be computed using Ullah's lemmas as follows.

$$F_{mn} = E\left[\frac{z'I_{mn}z}{(z'M_1z)^2}\right] \text{ where } I_{mn} \text{ is defined as before.}$$

$$= \Gamma(2)^{-1}\int_0^\infty t|N_{0t}|^{-1/2}(tr(I_{mn}N_{0t}^{-1}) + \mu' N_{2t}\mu)\exp(-1/2\mu' N_{1t}\mu)dt\,|_{\mu=0}.$$

where $N_{0t} = I+2tM_1$, $N_{1t} = 2tM_1N_{0t}^{-1}$, and $N_{2t} = N_{0t}^{-1}I_{mn}N_{0t}^{-1}$.

$$= \Gamma(2)^{-1}\int_0^\infty t|N_{0t}|^{-1/2}\, tr(I_{ij}N_{0t}^{-1})dt.$$

Hence, $d_{ij} = \lambda_2^2\sum_{m=1}^{2k}\sum_{n=1}^{2k} M_{2im}F_{mn}M_{2jn}$. Q.E.D

Reference

Babu and Singh (1983), Inference on Means using the Bootstrap, *Annals of Statistics*, 11, 999-1003

Bates, Charles E. and Halbert White (1993), Determination of Estimators with Minimum Asymptotic Covariance Matrices, *Econometric Theory*, 9, 633-648

Beran, Rudolf (1987), Prepivoting to Reduce Level Error of Confidence Sets, *Biometrika*, 74, 3, 457-468.

Brownstone, D.(1990), Bootstrapping Improved Estimators for Linear Regression Models, *Journal of Econometrics*, 44, 171-187

Delaney, Nancy Jo and Sangit Chatterjee (1986), Ues of the Bootstrap and Cross-Validation in Ridge Regression, *Journal of Business and Economic Statistics*, Vol. 4, No. 2, 255-262.

Efron, Bradley and Robert J. Tibshirani (1993), *An Introduction to the Bootstrap, Monographs on Statistics and Applied Probability 57*, Chapman & Hall.

Hahn, Jinyong (1995), Bootstrapping Quantile Regression Estimators, *Econometric Theory*, 11, 105-121.

Hall, Peter (1987), On the Bootstrap and Likelihood-based Confidence Regions, *Biometrika*, 74, 8, 481-493.

Hall, Peter (1988), Theoretical Comparison of Bootstrap Confidence Intervals, *The Annals of Statistics*, Vol. 16, No. 3, 927-953.

Hall, Peter (1992), *The Bootstrap and Edgeworth Expansion*, Springer-Verlag.

Hartigan, J. A. (1986), Comment on the Paper by Efron and Tibshirani, *Statistical Science*, 1, 75-76.

Politis, Dimitris N. and Joseph P. Romano (1994), The Stationary Bootstrap, *Journal of the American Statistical Association*, Vol. 89, No. 428, 1303-1313.

Schmoyer, R. and S. Arnold (1989), Shrinking Techniques for Robust Regression, *Contributions to Probability and Statistics: Essays in Honor of Ingram Olkin*, 368-384.

Ullah, A. (1990), Contribution to Econometric Theory and Application, *Essays in Honour of A. L. Naga*, 242-292

Vinod, H. D. (1995), Double Bootstrap for Shrinkage Estimators, *Journal of Econometrics*, 68, 287-302

Vinod, H. D. and B. Raj (1988), Economic Issues in Bell System Divesture: A Bootstrap Application, *Applied Statistics*, 37, 251-261

Table 2.5.1 Regression Result

Estimator	Variable	Coefficient	Std. Error	Coeff/Std.
OLS	x_1	0.077618	0.172397	0.450227
	x_2	-0.005561	0.169351	-0.032840
	x_3	0.015524	0.190734	0.081391
LAD	x_1	-0.001885	0.221519	-0.008510
	x_2	0.050000	0.217604	0.229775
	x_3	0.045658	0.245080	0.186299
OWS	x_1	0.072559	0.171099	0.424076
	x_2	-0.002030	0.168950	-0.011990
	x_3	0.017442	0.189935	0.091832

Table 2.5.2 Bootstrap 95% Confidence Intervals

Method		Confidence Lower Bound	Interval Upper Bound	Length	Shape
Percentile-t	x_1	-0.24321	0.434032	0.677245	1.144725
	x_2	-0.33606	0.296353	0.632408	0.893269
	x_3	-0.34916	0.406528	0.755685	1.061344
Naïve Percentile	x_1	-0.25017	0.404900	0.655070	0
	x_2	-0.30556	0.327596	0.633151	0
	x_3	-0.34828	0.358264	0.706539	0
Percentile	x_1	-0.25978	0.395289	0.655070	0.971079
	x_2	-0.33165	0.301504	0.633151	0.920841
	x_3	-0.32338	0.383159	0.706539	1.073044
Normal Approximation	x_1	-0.26280	0.407913	0.670709	1
	x_2	-0.33317	0.329116	0.662283	1
	x_3	-0.35483	0.389714	0.744544	1

Let 'cl' be the lower bound and 'cu' be the upper bound of a confidence interval.

Then 'Length' and 'Shape' are defined as follows.

(1) Length = cu - cl.

(2) Shape = (cu - b)/(b - cl) where b is the JSLAD estimate computed from the original data set. Shape measures how asymmetric the bootstrap confidence interval is around its center (b). If Shape > 1, then (cu - b) > (b - cl). If Shape < 1, then (cu - b) < (b - cl).

Figure 2.5.1 Histogram of Standardized Bootstrap Estimates for β_1

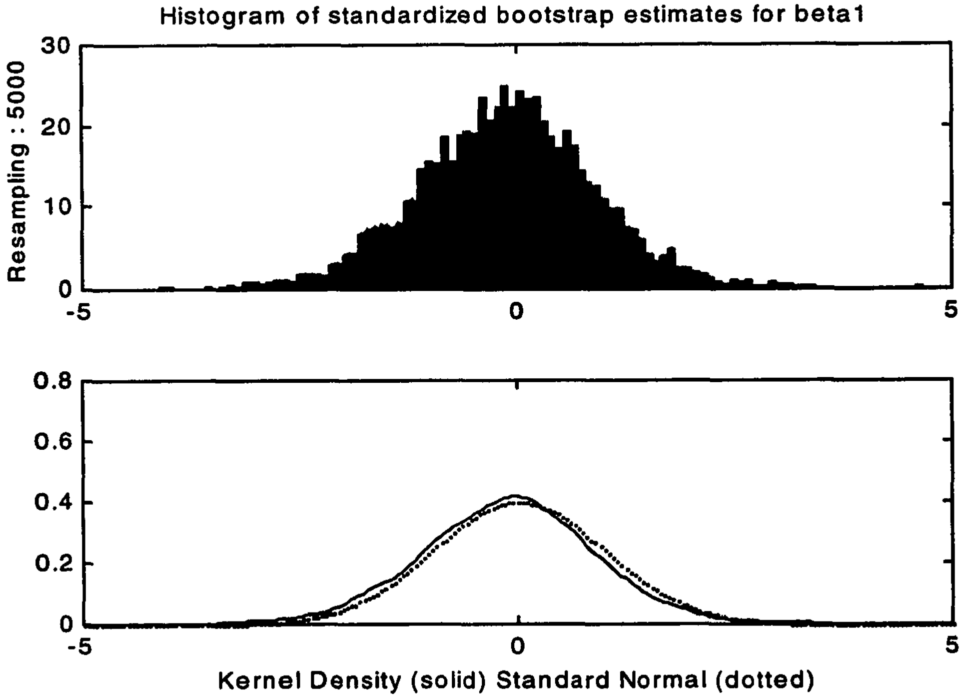


Figure 2.5.2 Histogram of Standardized Bootstrap Estimates for β_2

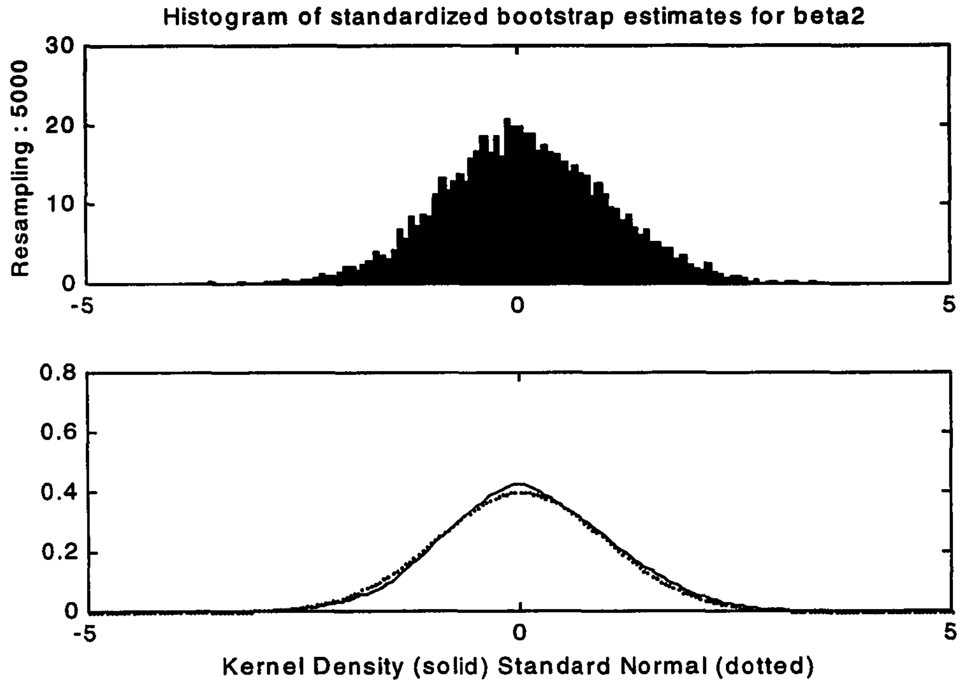


Figure 2.5.3 Histogram of Standardized Bootstrap Estimates for β_2

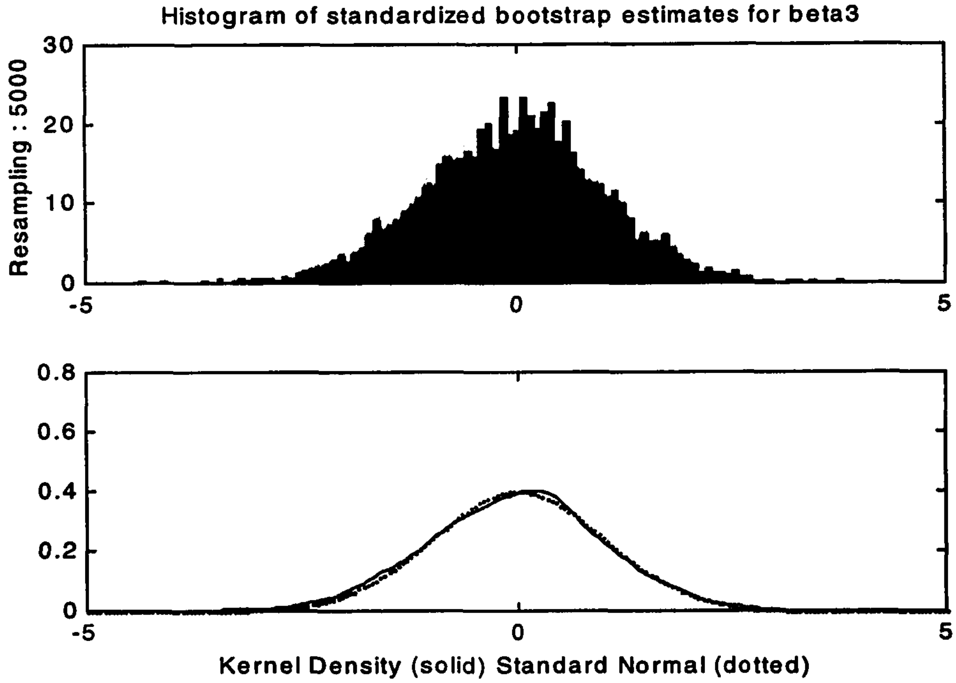


Figure 2.5.4 Integrands in E(m=1,n=1) and F(m=1,n=1)

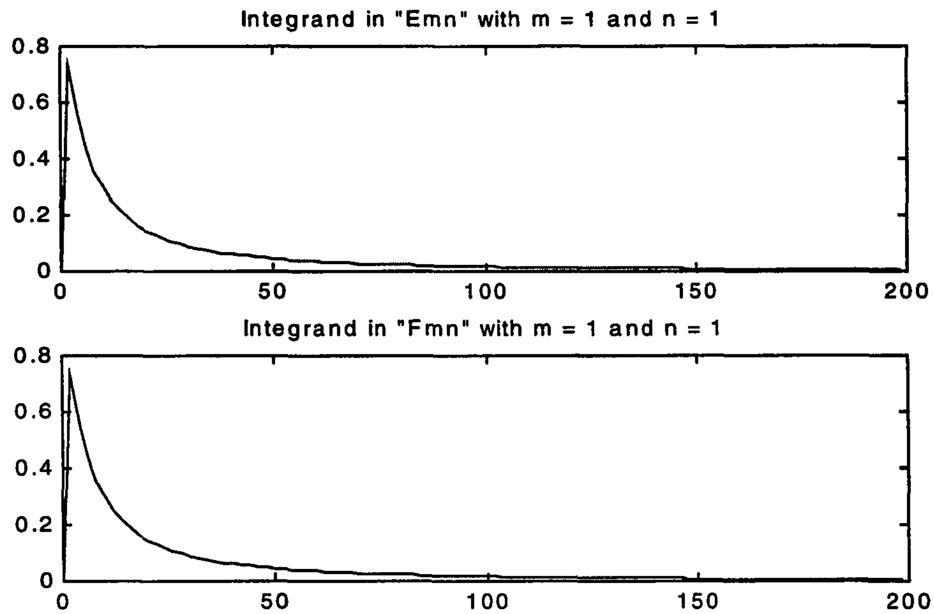
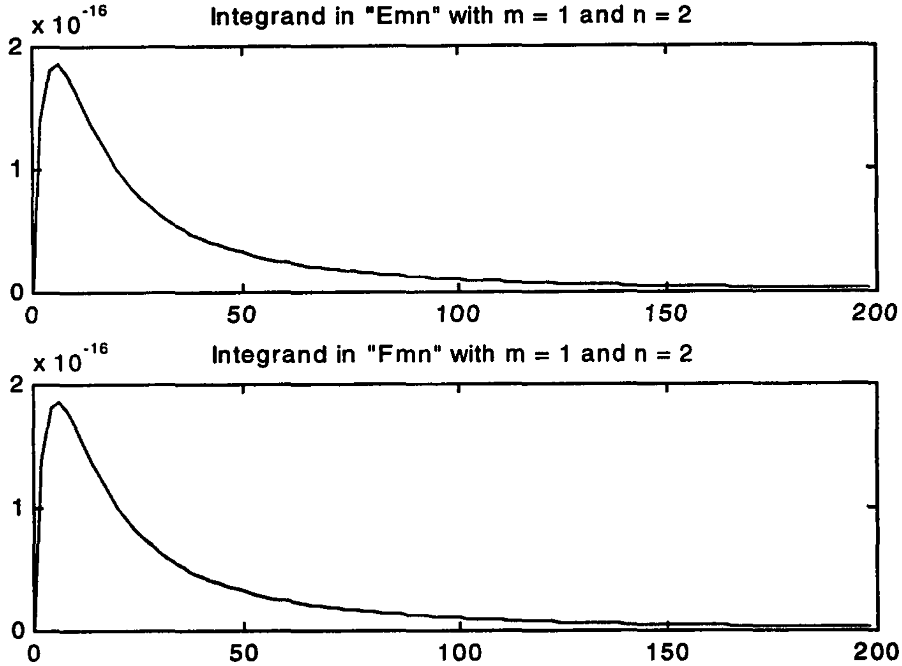


Figure 2.5.5 Integrands in E(m=1,n=2) and F(m=1,n=2)



Chapter 3

Application of Shrinkage LAD Estimation to the Treynor Black Model

3.1 Introduction

We apply various shrinkage estimators developed in the first chapter to the construction of the optimal portfolio as proposed by Treynor and Black (1973) using alpha and beta forecast data obtained from a financial institution. This paper also utilizes an extension of the Treynor-Black model which incorporates off-diagonal terms in the covariance matrix of abnormal returns.

Treynor-Black portfolios are constructed in two stages. First, an active portfolio is constructed from the securities under consideration. The active portfolio is on the efficient frontier based on the abnormal returns of the covered securities. In the second stage, the final, optimized portfolio, which we call the Treynor-Black Portfolio (TBP), is constructed from the Active Portfolio (AP) and the market index (M). The AP is optimally mixed with the market index to improve diversification so as to maximize the Sharpe measure of the TBP. Treynor-Black assume the validity of the Sharpe's Diagonal Model in which securities are correlated only through a common market factor. This is unrealistic, and we apply Theorem 2 in Roll (1977) to derive explicit formulas that delivers the TBP with a general covariance model.

The Treynor-Black model appears to have had little impact despite some early encouraging papers (Ambachtsheer (1974, 1977), Ambachtsheer and Farrell (1979), Black (1973), Ferguson (1975), Hodges and Brealey (1973), Kane and Marcus (1986)). Although theoretically compelling, the model has not been widely adopted by professional managers. It appears that security analysts are reluctant to put themselves to the quantitative test required by the TBP model. On the other hand, many academicians believe that the forecasting ability of most analysts may be below the threshold needed to make the model useful. This paper aims at identifying and lowering this threshold by using several promising statistical methods.

The performance of the TBP model depends critically on 1) the predictive ability contained in the abnormal return forecasts, and 2) application of the statistical properties of the forecasts to the portfolio re-balancing process. It has been intermittently reported in the literature (See Ambachtsheer(1974, 1979) and Black (1973)) that some organizations possess a small yet significant forecasting power, but that they were lacking the efficient technology to use it systematically. We introduce various ways of shrinking a robust estimator toward a data-dependent point and we use these methods to translate the predictive power into the portfolio decision process. Given that abnormal returns tend to exhibit fat-tailed distribution, we choose the Least Absolute Deviation (LAD) estimator as the base estimator.

The quality of the estimates of market betas determines the accuracy of the estimates of ex-post abnormal returns, which, in turn, enables us to measure the bias and precision of alpha forecasts. Because many stocks are traded infrequently, we use Dimson (1979)'s "Aggregate Coefficients" (AC) method to estimate market beta. The analysis of the alpha forecast database suggests (1) the correlation between alpha forecasts and realizations is as low as 0.04. (2) the alpha forecasts are biased and forecast errors are asymmetric. It appears that the analyst predictive ability declines over the sample period. Nevertheless, the application of shrinkage LAD estimation to the full-covariance TBP results in superior performance.

Out-of-sample experiments show that a $1 invested in a properly managed TBP Covariance Model based on forecast database would yield $1.810 with Sharpe Ratio being 1.340 over 3 years. On the other hand, if you invest $1 in the S&P500 index over the same 3 years, the final wealth is $1.259 and the Sharpe Ratio is 0.909. This result shows that a large potential improvement can be obtained, and this can be done without requiring a high threshold for forecasting ability. Nevertheless, the TBP turns out to be unstable in that the weight given to the AP can be large and volatile; as a result the TBP can be "concentrated" rather than diversified. We have also found that managing the market risk (market beta) properly is an important ingredient to obtaining a better TBP. The TBP based on the OLS estimator is always dominated by the TBP based on the LAD and shrinkage LAD estimators.

The paper is organized as follows. First, we restate briefly various methods of shrinking the LAD estimator toward a data-dependent point. A general and detail discussion can be found in the first chapter. Second, we introduce the Treynor-Black

Portfolio Model and derive a closed form solution for a general covariance model using efficient frontier mathematics. Third, we apply shrinkage LAD estimators to construction of the TBP covariance model using actual alpha and beta forecast data.

3.2 Introduction of Shrinkage LAD Estimator

The James-Stein shrinkage estimator can be obtained by shrinking a base estimator (e.g. OLS estimator) toward a fixed point which is usually set to zero. Basically, the risk improvement can be explained by a simple variance-bias trade-off. Shrinking a base estimator makes the variance smaller and the bias greater, which leads to smaller risk under the choice of an optimal degree of shrinkage. One of the problems of the JS estimator is that for large samples, there is no advantage to using the JS estimator as the risk of the JS estimator converges to the risk of the base estimator. This can be understood in terms of a natural relationship between JS-type estimators and Bayesian estimators. The fixed point toward which the base estimator is shrunk can be often viewed as a prior or guess about the true parameter. As the sample size gets larger, the influence of the prior tends to vanish and the base estimator becomes more reliable. As a result, there is less room for the shrinkage technique to play a role, explaining why there is no asymptotic improvement. One way of overcoming this problem is to shrink the base estimator toward a data-dependent point.

Consider $y_t = x_t'\beta^0 + \varepsilon_t$ $t = 1, 2, \ldots, n$ where $\beta^0 \in R^k$ and ε_t is assumed to be identical and independent. We define $X^n = [x_1, x_2, \ldots, x_n]'$. Let b_n be an estimator for β^0. A function $L(b_n,\beta^0)$ is called the loss function if and only if (1) $L(b_n,\beta^0) \geq 0$ for all b_n and all β^0 and (2) $L(b_n,\beta^0) = 0$ if and only if $b_n = \beta^0$. The expectation of the loss function, $E(L(b_n,\beta^0))$ is called the risk, denoted by $R(b_n,\beta^0)$. An example of a loss function is the quadratic loss, $L(b_n,\beta^0) = (b_n-\beta^0)'Q_n(b-\beta^0)$ where Q_n is a symmetric and positive definite matrix. Let $\{b_n\}$ be a sequence of estimators of $\beta^0$and let $\{L(b_n,\beta^0)\}$ be a sequence of loss values. Suppose $L(b_n,\beta^0)$ converges to an integrable random variable Ψ in distribution. The asymptotic risk of $\{b_n\}$ for $\{L(b_n,\beta^0)\}$ is then defined by

$AR(\{b_n\}, \beta^0) = E(\Psi)$.

Let g_n be a data-dependent point, which satisfies the following joint normality condition.

$$\begin{bmatrix} n^{1/2}(b_n - \beta^0) \\ n^{1/2}(g_n - \beta^0) \end{bmatrix} \xrightarrow{d} \begin{bmatrix} U_1 \\ U_2 \end{bmatrix} \sim N\left(\begin{bmatrix} 0_{k\times 1} \\ \theta_{k\times 1} \end{bmatrix}, \begin{bmatrix} A_{k\times k} & \Delta_{k\times k} \\ \Delta'_{k\times k} & B_{k\times k} \end{bmatrix}\right)$$

where Δ is the asymptotic covariance matrix for the base estimator and the guess estimator. We now give a formal definition of the James-Stein Combination Estimator (JSC).

Definition 2-1 James-Stein Combination Estimator

The combination of two estimators using the James-Stein rule,

$$\delta_\lambda^{JS}(b_n, g_n) = \left(1 - \frac{\lambda}{(b_n - g_n)' Q_n (b_n - g_n)}\right)(b_n - g_n) + g_n$$

where λ is a constant, is called the James-Stein Combination Estimator (JSC).

The we can show the asymptotic risk of the JSC estimator is smaller than the asymptotic risk of the base estimator as long as some relative non-efficiency conditions[1]. Note that the shrinkage toward a data-dependent point is equivalent to combining the base estimator and the data-dependent point using a random weight determined by James-Stein rule. We can further extend this approach by developing the Optimal Weighting Scheme (OWS) which includes random weight as well as non-random weight as special cases.

Definition 2-2 Optimal Weighting Scheme Estimator (OWS)

The combination of two estimators defined by

$$\delta_\lambda^{OW}(b_n, g_n) = \left(1 - \lambda_1 - \frac{\lambda_2}{(b_n - g_n)' Q_n (b_n - g_n)}\right)(b_n - g_n) + g_n$$

where $\lambda = [\lambda_1\ \lambda_2]'$, is called the Optimal Weighting Scheme Estimator (OWS).

[1] For a precise statement of this condition, see the first chapter. This condition requires that the base should not be asymptotically efficient.

The following theorem provides the optimal combination weight for the OWS estimator and some conditions under which the asymptotic risk of the OWS estimator is smaller than the asymptotic risk of the base estimator. See Chapter 1 for the proof.

Theorem 2-1 Under the Joint Normality condition and some regularity condtions,

(1) $AR(\{\delta_\lambda^{OW}(b_n,g_n)\},\beta^0)$ is strictly convex in λ.

(2) Let $\lambda^* \in \operatorname{argmin} AR(\{\delta_\lambda^{NR}(b_n,g_n)\},\beta^0)$. Then

$\lambda_1^* = (\alpha\omega-1)^{-1}(\beta\omega-\nu)$ and $\lambda_2^* = (\alpha\omega-1)^{-1}(\alpha\nu-\beta)$.

(3) $AR(\{\delta_{\lambda^*}^{OW}(b_n,g_n)\},\beta^0) = (\alpha\omega-1)^{-2}[-\alpha\beta^2\omega^2-(2\alpha\beta\nu-\alpha^2\nu^2+\beta^2)\omega+(\alpha\nu^2-2\beta\nu)] + \kappa$.

(4) $AR(\{\delta_{\lambda^*}^{OW}(b_n,g_n)\},\beta^0) \leq AR(\{b_n\},\beta^0)$ where the equality holds only when $\beta = 0$ and $\nu = 0$.

where

$$\alpha = E\left[(U_1 - U_2)'Q(U_1 - U_2)\right], \qquad \beta = E\left[U_1'Q(U_1 - U_2)\right].$$

$$\nu = E\left[\frac{U_1'Q(U_1 - U_2)}{(U_1 - U_2)'Q(U_1 - U_2)}\right], \qquad \omega = E\left[\frac{1}{(U_1 - U_2)'Q(U_1 - U_2)}\right].$$

The method to estimate α, β, ν and ω consistently is discussed in detail in the first chapter. In order to compute the combination of the two estimator, we need to estimate the error density evaluated at zero and the covariance matrix between two estimators. We estimate the density using a Kernel method with Gaussian Kernel. See the Appendix for detail discussion. Since both estimators are in the RCASOI (Regular Consistent Asymptotically Second Order Indexed) class, we can exploit the score and Hessian representations of both estimators to compute the covariance matrix. See Bates and White (1993) for a detail discussion

It is well known that even though the OLS estimator is BLUE, it is sensitive to outliers and is not stable in that a small change in the data can cause a big change in the estimation result. Considering that the ex-post abnormal return process is the dependent variable in our application and that this has a fat-tailed distribution, we proceed using the Least Absolute Deviations (LAD) estimator. This estimator has the attractive property that it is robust to the outliers in the dependent variable. If we choose the LAD and OLS estimators to shrink toward each other, the theory developed so far give the following 3 optimal shrinkage LAD estimators.

$$b^{NRLAD} = (1-w_1)(b^{LAD} - b^{LS}) + b^{LS}$$

$$b^{JSLAD} = \left(1 - \frac{w_2}{(b^{LAD} - b^{LS})'Q(b^{LAD} - b^{LS})}\right)(b^{LAD} - b^{LS}) + b^{LS}$$

$$b^{OWLAD} = \left(1 - \lambda_1 - \frac{\lambda_2}{(b^{LAD} - b^{LS})'Q(b^{LAD} - b^{LS})}\right)(b^{LAD} - b^{LS}) + b^{LS}$$

where $w_1 = \beta/\alpha$, $w_2 = \nu/\omega$, $\lambda_1 = (\alpha\omega-1)^{-1}(\beta\omega-\nu)$ and $\lambda_2 = (\alpha\omega-1)^{-1}(\alpha\nu-\beta)$. The JSLAD is better than the LAD estimator and the NRLAD and OWLAD are better than both the LAD estimator and the OLS estimator in terms of asymptotic risk when the relative non-efficiency condition and some regularity conditions are satified.

3.3 Construction of the Treynor-Black Portfolio

In this section, we explain the original Treynor-Black diagonal model and make an extension of the model which incorporates off-diagonal terms in the covariance matrix of abnormal returns. The Treynor-Black model starts with the single-index model,

(1) $r^s_i = r_f + \beta_i(r^s_m - r_f) + z_i \qquad i = 1, 2, \ldots, N.$

where

r^s_i = return on the *i*th security,

r_f = riskless rate of return,

β_i = market sensitivity of the *i*th security,

r^s_m = return on the market,

z_i = abnormal return on the *i*th security,

N = number of securities.

Each return is decomposed into three pieces: a riskless return component [r_f], a systematic (normal) return component [$\beta_i(r^s_m - r_f)$] and an independent (abnormal) return component

$[z_i]$. Denote the excess return on the *i*th security and the excess return on the market by r_i and r_M respectively: $r_i = r^s_i - r_f$ and $r_m = r^s_m - r_f$. Then

(2) $r_i = \beta_i r_m + z_i$.

Treynor and Black impose the following error structure on the abnormal returns z_i.

TBA 1. $E(z_i) = \alpha_i$

TBA 2. $\text{Cov}(z_i,z_j) = \begin{cases} \sigma_i^2 & \text{if } i = j \\ 0 & \text{otherwise.} \end{cases}$

TBA 3. $\text{Cov}(z_i,r_m) = 0$.

In addition to these error structure assumptions, Treynor and Black assume that there are no restrictions on borrowing and short sales and that there are no taxes. The α_i can be interpreted as a measure of how far the excess return on the *i*th security is away from its 'equilibrium value'. If $\alpha_i = 0$ for all i, then the market is in equilibrium. By allowing α_i to be non-zero, the model provides a way of forming an optimal portfolio by exploiting the security analyst's findings about the α_i. The second assumption means that the only source of the co-movement between two excess returns is the common market return component. We think that this assumption may be unrealistic[2] and we drop it. We therefore impose.

Assumption 1 $E(z_i) = \alpha_i$.

Assumption 2 $\text{Cov}(z_i,r_m) = 0$.

We define $\varepsilon_i \equiv z_i - E(z_i|I)$. By using the abnormal return decomposition, the basic model (2) can be rewritten as follows.

(3) $r_i = \alpha_i + \beta_i r_m + \varepsilon_i$

where $E(\varepsilon) = 0$ and $\text{Var}(\varepsilon) = \Omega$ with (i,j)th element $= \sigma_{ij}$.

For the moment, we assume an ideal situation where we know all the true values $\{\alpha_i, \sigma_{ij}, \beta_i, \mu_m, \sigma_m : i, j = 1, 2, \ldots, M\}$, where $\mu_m = E(r_m)$, $\sigma_m = \text{Var}(r_m)$ and $M \leq N$.[3] In order to make notation simple, we assume that M = N. See Bodie, Kane and Marcus (1996) for an illustrative numerical and graphical example for this two step approach.

---

[2] See Ferguson (1975) pp 72.

Step One: Active Portfolio Construction

Definition 3-1 Active Efficient Frontier (AEF)

The efficient frontier traced by solving the following optimization programming based on the abnormal returns is called the 'Active Efficient Frontier'.

$$\underset{h}{\text{Min}} \quad \sigma(h)^2 \equiv h'\Omega h$$

$$\text{subject to } h'\alpha = r_a \text{ and } h'\iota = 1$$

where h is the N×1 vector of portfolio weights, α is the N×1 vector with elements α_i, and r_a is a pre-specified abnormal return level.

Definition 3-2 Active Sharpe Ratio (ASR)

Let h be on the AEF. Then $\frac{h'\alpha}{\sqrt{h'\Omega h}}$ is called the 'Active Sharpe Ratio'.

Definition 3-3 Active Portfolio (AP) Weight

The portfolio weight derived by the following ASR maximization problem is called the 'Active Portfolio Weight'.

$$h^* \in \arg\max_{h} \frac{h'\alpha}{\sqrt{h'\Omega h}} \quad \text{subject to } h'\iota = 1.$$

Theorem 3-1 Given the true values $\{\alpha, \Omega\}$, the AP weight (h^*) is given by

$$h^* = [\alpha'\Omega^{-1}\iota]^{-1}\Omega^{-1}\alpha.$$

Proof: See Appendix 1.

[3] In order to make the 'Active Portfolio' defined later, we do not have to investigate all securities in the market.

Note that if all off-diagonal terms are zero, then the AP weight (h^*) simplifies to the following.[4]

$$h_i^* = \left[\sum_{j=1}^{N} \frac{\alpha_j}{\sigma_j^2}\right]^{-1} \frac{\alpha_i}{\sigma_j^2}.$$

The return on the AP is

$$r_A = \sum_{i=1}^{N} h^*_i r_i$$

$$= \sum_{i=1}^{N} h^*_i \alpha_i + \sum_{i=1}^{N} h^*_i \beta_i r_m + \sum_{i=1}^{N} h^*_i \varepsilon_i .$$

Define $\alpha_A \equiv \sum_{i=1}^{N} h^*_i \alpha_i$, $\beta_A \equiv \sum_{i=1}^{N} h^*_i \beta_i$ and $\varepsilon_A \equiv \sum_{i=1}^{N} h^*_i \varepsilon_i$. Then $r_A = \alpha_A + \beta_A r_m + \varepsilon_A$. We can compute the expectation and variance of the AP using Assumption 1 and Assumption 2 as follows.[5]

$$\mu_A \equiv E(r_A) = \alpha_A + \beta_A \mu_m.$$

$$\sigma_A^2 \equiv Var(r_A) = \beta_A^2 \sigma_m^2 + \sigma(\varepsilon_A)^2$$

where $\sigma(\varepsilon_A)^2 = h^{*\prime}\Omega h^*$.

Definition 3-4 Appraisal Ratio (AR)

The quantity $\alpha'\Omega^{-1}\alpha$ is called the 'Appraisal Ratio'.

The AR is a distance of the mean vector of the abnormal returns from the origin (the market equilibrium) weighted by the inverse of the variance-covariance matrix. It measures how far we deviate from the market equilibrium or how much of a contribution

---

[4] If Ω is diagonal, then $b = \sum_{j=1}^{N} \frac{\alpha_j}{\sigma_j^2}$ and $\Omega^{-1}\alpha = \left[\frac{\alpha_1}{\sigma_1^2}, \ldots, \frac{\alpha_N}{\sigma_N^2}\right]'$.

[5] Note that $Cov(\varepsilon_A, r_m) = E[\varepsilon_A(r_m - \mu_m)] = \sum_{i=1}^{N} h_i^* E(\varepsilon_i r_m) = 0$ because $Cov(\varepsilon_i, r_m) = 0$ by assumption 2.

security analysis can make. As the following corollary shows, the AR is an evaluation measure of the AP.

Corollary 3-2 The square of the maximized ASR is equal to the AR. In other words,[6]

$$\left[\frac{h^{*\prime}\alpha}{\sqrt{h^{*\prime}\Omega h^{*}}}\right]^2 = \alpha'\Omega^{-1}\alpha.$$

If all off-diagonal terms are zero, then the maximized ASR simplifies to[7]

$$\left[\frac{h^{*\prime}\alpha}{\sqrt{h^{*\prime}\Omega h^{*}}}\right]^2 = \sum_{i=1}^{N}\frac{\alpha_i^2}{\sigma_i^2}.$$

Step Two: TB-Portfolio Construction

We then mix the AP with a market portfolio. Define $r_p(w)$ to be the mixture portfolio with weight w: $r_p(w) = wr_A + (1-w)r_m$. Accordingly we can calculate the mean and variance[8] of the mixture portfolio as functions of w.

$$\mu_p(w) \equiv E(r_p(w)) = w\mu_A + (1-w)\mu_m$$

$$\sigma_p(w)^2 \equiv Var(r_p(w)) = w^2\sigma_A^2 + (1-w)^2\sigma_m^2 + 2w(1-w)\sigma_{Am}$$

$$= w^2(\beta_A\sigma_m^2 + \sigma(\varepsilon_A)^2) + (1-w)^2\sigma_m^2 + 2w(1-w)\beta_A\sigma_m^2.$$

---

[6] Note that (1) $h^{*\prime}\alpha = [\alpha'\Omega^{-1}\iota]^{-1}\alpha'\Omega^{-1}\alpha$ and (2) $h^{*\prime}\Omega h^{*} = [\alpha'\Omega^{-1}\iota]^{-1}\alpha'\Omega^{-1}\Omega[\alpha'\Omega^{-1}\iota]^{-1}\Omega^{-1}\alpha = [\alpha'\Omega^{-1}\iota]^{-2}\alpha'\Omega^{-1}\alpha$.

Therefore, $\left[\frac{h^{*\prime}\alpha}{\sqrt{h^{*\prime}\Omega h^{*}}}\right]^2 = \alpha'\Omega^{-1}\alpha$.

[7] $$\left[\frac{h^{*\prime}\alpha}{\sqrt{h^{*\prime}\Omega h^{*}}}\right]^2 = \frac{(\sum_{i=1}^{N} h^*_i\alpha_i)^2}{\sum_{i=1}^{N} h^{*2}_i\sigma_i^2} = \frac{(\sum_{i=1}^{N}\frac{\alpha_i^2}{\sigma_i^2})^2}{(\sum_{i=1}^{N}\frac{\alpha_i}{\sigma_i^2})^2}\frac{(\sum_{i=1}^{N}\frac{\alpha_i}{\sigma_i^2})^2}{(\sum_{i=1}^{N}\frac{\alpha_i}{\sigma_i^2})} = \sum_{i=1}^{N}\frac{\alpha_i^2}{\sigma_i^2}.$$

[8] $$\begin{aligned}Cov(r_A,r_m) &= E[(r_A-\mu_A)(r_m-\mu_m)]\\ &= E[(\alpha_A+\beta_A r_m+\varepsilon_A-\alpha_A-\beta_A\mu_m)(r_m-\mu_m)]\\ &= E[\beta_A(r_m-\mu_m)^2] + E[\varepsilon_A(r_m-\mu_m)]\\ &= \beta_A E(r_m-\mu_m)^2 \quad (E(\varepsilon_A r_m) = E(\varepsilon_A) = 0 \text{ by Assumption 1, 2})\\ &= \beta_A\sigma_m^2\end{aligned}$$

Definition 3-5 Treynor-Black Portfolio (TBP)[9] *Weight*

The portfolio weight derived by maximizing the following Sharpe Ratio is called the Treynor-Black Portfolio Weight.

$$w^* \in \arg\max_{w} S_p(w) \equiv \frac{\mu_p(w)}{\sigma_p(w)}.$$

Theorem 3-3 Given the true values $\{\alpha, \Omega, \beta, \mu_m, \sigma_m\}$, the TBP weight ($w^*$) is given by

$$w^* = \frac{\alpha_A \sigma_m^2}{(1-\beta_A)\alpha_A \sigma_m^2 + \mu_m h^{*\prime} \Omega h^*}$$

Proof: See Appendix 1.

Note that the TBP weight (w^*) can be simplified to

$$w^* = \frac{w_0}{1+(1-\beta_A)w_0} \text{ where } w_0 = \frac{\dfrac{\alpha_A}{h^{*\prime}\Omega h^*}}{\dfrac{\mu_m}{\sigma_m^2}} = \frac{\dfrac{\textit{alpha of the AP}}{\textit{residual variance}}}{\dfrac{\textit{market mean}}{\textit{market variance}}}.$$

The TBP return is given by

$$r_p(w^*) = w^* r_A^* + (1-w)^* r_m$$

$$= \sum_{i=1}^{N} w^* h_i^* r_i + (1-w^*) r_m$$

using the definition of $r_A^* \equiv \sum_{i=1}^{N} h_i^* r_i$. Hence we have N+1 securities including the market index. The optimal weight to the *i*th security is $w^* h_i^*$ and the optimal weight to the market index is $(1-w^*)$. The square of the maximized Sharpe Ratio can be decomposed into the square of the market portfolio's Sharpe Ratio and the AR, i.e.

$$\left[\frac{\mu_p(w^*)}{\sigma_p(w^*)}\right]^2 = \mu_m \sigma_m^{-2} \mu_m + \alpha' \Omega^{-1} \alpha.$$

[9] Usually a tangential portfolio is computed with respect to the riskless rate of return. Since we have already subtracted the riskless rate of return from returns, the tangential portfolio should be computed with respect to the origin.

This clearly shows that the market Sharpe Ratio is the contribution of a market index and that the AR is the contribution of the security analysis to the TBP.

3.4 Forecast Data Description

We use alpha and beta forecasts provided by a financial investment firm, Advanced Investment Technology (AIT). These are 12 week (approximately 1 quarter) ahead forecasts. The firm generates alpha and beta forecasts every month.[10] The data set spans 3 years from December 1992 through December 1995, which gives us 37 observations in the time dimension. December 31, 1992 the firm generated alpha and beta forecasts for 711 securities. Each month it has added several new securities into the forecast data base, ending up with 771 securities on December 29, 1995.

Let $\{(\hat{\alpha}_{i\tau}, \hat{\beta}_{i\tau}) \mid i = 1,2, \ldots, N,\ \tau =$ December 31, 1992, January 29, 1993, . . . , November 24, 1995, December 29, 1995$\}$ represent the forecast data set. The definition of $\hat{\alpha}_{i(\tau=December 31,1992)}$ is that it is the forecast of the ith security's abnormal return over the time period between December 31, 1992 and March 26, 1993. We call τ the 'Date Forecasts Made (DFM)'. Likewise, $\hat{\beta}_{i(\tau=December 31,1992)}$ is the forecast of the ith security's market beta 12 weeks later.[11]

There are 37 dates when the alpha and beta forecasts were made, which will be indexed by $\tau \in \text{DFM} \equiv \{\tau_i : i = 1, 2, \ldots, 36, 37\}$, where τ_1 = December 31, 1992, τ_2 = January 29, 1993, . . ., τ_{37} = December 29, 1995. See Table 3.4.1 for the list of exact dates of the forecasts. Also, we define the Prediction Index Set $S \equiv \{1, 2, \ldots 36, 37\}$.

After deleting any security having at least one missing value, we have alpha and beta forecasts for 646 securities over the sample period. Figure 3.4.1 and Figure 3.4.2 are

---

[10] Roughly 6-8 years of history data (including both technical and fundamental information) are used to generate alpha forecasts.

[11] The firm generates alpha and beta forecasts the last Friday every month. There are two exceptions: March 1993 forecasts and June 1994 forecasts have been made on December 31, 1992 (Thursday) and March 31, 1994 (Thursday) respectively.

distributions of average alpha and beta forecasts over 646 securities. The plots are drawn by taking the average over the sampling period for each security and then forming histograms. The distribution of alpha forecasts is skewed to the right. The distribution of beta forecasts is centered around one as expected, but has a local maximum in the downside tail. In many cases, financial firms produce alpha forecasts in the form of a ranking variable. In such cases, the conversion of the ranking variable to some comparably scaled variable can be done using the 'IC (Information Correlation) Adjustment' proposed by Ambachtsheer (1977). In our case, the alpha forecast is not solely a ranking variable, even though it takes only integer values between -12 and 14 which is due to rounding made by the institution, i.e. the integer values are point forecasts of returns. Table 3.4.2 gives summary statistics for the location and dispersion of the average alpha and beta distributions.

3.5 Stratified Random Sampling

In this work, we confine our study to a smaller subset of the full 646 stock population because we want to reduce the computational burden[12] and increase the precision of the residual covariance matrix. Accordingly, we have chosen to work with a subset of 105 randomly selected stocks. We explain how to select 105 stocks in the following.

Rosenberg, Reid, and Lanstein (1985) and Fama and French (1992) provide empirical evidence that the book-to-market equity (BE/ME) ratio and market capitalization have significant explanatory power for average stock returns.[13] Accordingly, we want to preserve the distribution of the BE/ME ratio and capitalization distribution during the sampling process.

[12] Charnes et al. (1954) showed that the LAD estimator can be obtained by simplex linear programming methods. This method is not efficient, in that the parameter space grows along with the number of observations and as a result, it requires a long search time. Barrodale and Roberts (1974) proposed a modified version of the simplex algorithm referred as the BR-L_1 Algorithm. This algorithm is much more efficient than the simplex method and greatly reduces the computation time. Nevertheless, even using the BR-L_1 Algorithm, it takes considerable computation time to analyze all 646 stocks.

For this, we have obtained the following annual data over the 3 year sample periods from the COMPUSTAT database: book value (common equity, COMPUSTAT item 60,unit = millions of dollars), common shares outstanding (COMPUSTAT item 25, unit = millions) and closing price (COMPUSTAT item 24,unit = dollars and cents). Since there are 41 missing observations for book value in 1995 and 45 missing observations for common shares outstanding in 1995, we have used only the middle year (1994) data to generate the following variables[14].

(1) Size = Common Shares Outstanding × Closing Price.

(2) BE/ME = Book Value/Size.

First we divide the market capitalization distribution of the population into 7 classes such that each contains the same number of stocks. Then we divide the BE/ME population distribution in the same manner. Thus the population is divided into 49 strata and each strata has approximately 13 observations. Define the following variables.

N = total number of stocks in population (646).

N_i = total number of stocks in *i*th stratum (about 13).

n = desired sample size (105).

n_i = number of stocks drawn from *i*th stratum.

We choose n_i such that $n_i/n = N_i/N$ and then draw n_i (typically 2 or 3) samples from the *i*th stratum randomly. This procedure gives us a stratified random sample of 105 observations. We will analyze this sample in the subsequent sections. Figure 3.5.1 and Figure 3.5.2 show population and sample histogram for capitalization as well as BE/ME ratio.

3.6 Estimation of Market Beta and Ex-post Abnormal Returns

[13] Average stock returns in the United States are negatively related to both market to book value and market capitalization. Fama & French argue that these might serve as proxies for a portion of risk premium which is not captured by market beta.

[14] The Sampling distribution of size and BE/ME in 1994 is almost identical to the sampling distribution of size and BE/ME averaged over 1993 and 1994.

When we are given alpha forecasts, we have two choices of how to use them in the TB analysis. One simple way is to use the alpha forecasts directly, without any modification, so that

$$E(z_t|I_t) = \hat{\alpha}_t .$$

Nevertheless, we typically do not want to use alpha forecasts in this way because we are not certain of the quality predictive power of such alpha forecasts. Instead, we want to make an appropriate transformation of the alpha forecasts, using information available up to the time at which the forecast is made. In this case, we have

$$E(z_t|I_t) = f(\hat{\alpha}_t, I_t)$$

where the function f(.) embodies our confidence in the forecasts and discounts the forecasts accordingly. We call f(.) the 'discount function'. If we have enough past data on ex-post abnormal return (z_t), then we can use the historical correlation between ex-post abnormal returns and alpha forecasts to form the discount function for alpha forecasts. For example, if we regress z_t on a constant and $\hat{\alpha}_t$, then the discount function is

$$f(\hat{\alpha}_t, I_t) = a + b\hat{\alpha}_t$$

where (1) $b = \rho \dfrac{\sigma_z}{\sigma_\alpha}$

(2) $a = E(z_t) - bE(\hat{\alpha}_t)$

(3) $\rho = Cor(z_t, \hat{\alpha}_t)$.

In order to estimate the correlation, we need to know the actual abnormal return process $\{z_t\}$, which requires us to know excess returns (r_{it}), market excess return (r_{mt}) and true market beta (β_i); $z_{it} \equiv r_{it} - \beta_i r_{mt}$. But it is not possible to obtain historical values for abnormal returns because we do not know the true market beta. Therefore the quality of estimated ex-post abnormal returns critically depends on the quality of the estimated market beta. Let $\hat{\beta}_i$ be an estimate of the market beta for the *i*th security. Then the estimated ex-post abnormal return[15] is $\hat{z}_{it} \equiv r_{it} - \hat{\beta}_i r_{mt}$.

One distinguishing characteristic of the alpha forecasts we have obtained is that the forecasting horizon is about 3 months, but new forecasts arrive every month. Therefore, we can investigate such interesting issues as how to update old forecasts when new forecast information arrives within the old forecast's horizon. However, this kind of overlapping structure introduces certain serial correlation problems and makes out-of-sample portfolio analysis and interpretation much more complicated. We leave the topic of updating forecasts for further research and seek to make our analysis as simple as possible. Accordingly we divide the alpha forecasts by 3 and treat them as one month abnormal return forecasts. We also assume that the imaginary portfolio manager acts as if the forecasts are made on the last trading day of each month, and he makes all portfolio decisions and buy and sell executions on the first trading day of the next month. Therefore, whenever we mention "Date Forecasts Made", it means the last trading day of the month. Table 3.4.1 shows that the number of days between actual DFM and the buy and sell execution date is very small, which justifies our simplifying assumption.

One final point should be emphasized. Using our simple transformation of the alpha forecasts means that we are using only part of the information in the alpha forecasts. If this partial information is useful in TB portfolio analysis, it is then plausible to make additional improvements using the full information in the alpha forecasts.

---

[15] We have errors in variable problem, but in the dependent variable, which is easier to handle than the errors in independent variable. Note that

$$\hat{z}_t \equiv r_{it} - \hat{\beta}_i r_{mt} = \beta_i r_{mt} + z_t - \hat{\beta}_i r_{mt}$$

$$\equiv z_t + \eta_t \text{ where } \eta_t \equiv (\beta_i - \hat{\beta}_i) r_{mt}.$$

Assume that the model

$$z_t = a + b\hat{\alpha}_t + v_t$$

satisfies relevant classical regression assumptions. Since we observe only $\hat{z}_t$, the regression we run is

$$\hat{z}_t = a + b\hat{\alpha}_t + v_t$$

which means

$$z_t = a + b\hat{\alpha}_t + v^*_t$$

where $v^*_t \equiv v_t - \eta_t$. We assume that $\text{cov}(\hat{\alpha}_t, \eta_t) = 0$.

(1) Data

We have obtained daily closing prices (p_{it}) for the 105 securities, the S&P500 index (p_{mt}) and annualized US treasury Bills - 3 month (r_t^f) from the DATASTREAM database. The sample period covers 1/1/90 through 3/31/96 which gives us daily 1630 observations in the time dimension. We transform daily prices into effective daily holding period returns as

$$r_{it} = r_{it}^s - r_t^f \times (1/260)$$

$$r_{mt} = r_{mt}^s - r_t^f \times (1/260)$$

where

$$(1) \quad r_{it}^s = \frac{p_{i(t+1)} - p_{it}}{p_{it}} \times 100$$

$$(2) \quad r_{mt}^s = \frac{p_{m(t+1)} - p_{mt}}{p_{mt}} \times 100.$$

We add daily returns in the month to obtain monthly returns which we will denote as

$${}^m r_{it} = \sum_{k=t+1}^{n} r_{ik}$$

$${}^m r_{mt} = \sum_{k=t+1}^{n} r_{mk}$$

where n is the number of days in that month.[16]

(2) Beta Estimation

Estimating market beta correctly is critical to the TB portfolio analysis for the following two reasons. Firstly, better beta estimates lead to better ex-post abnormal returns, which enables us to correctly measure the predictive ability of the forecasts. Secondly, as seen in the closed form solution for the TBP weight, the estimated market beta or beta forecast has a direct impact on the TBP weight. One simple way of

---

[16] Since daily returns are discretely compounded returns, the exact monthly return is $\Pi(1+r_t) - 1$. It can be shown by Taylor expansion that $\Pi(1+r_t) - 1 \approx \Sigma r_t$.

estimating market beta is to use the beta forecasts we have obtained from the financial institution (AIT); another is to use the OLS regression of excess return on a constant and excess market return. Figure 3.6.1 shows the distribution of beta forecasts.[17] However, the OLS beta estimates are not satisfactory when securities are traded infrequently because beta estimates are biased downward.[18] Figure 3.6.2 shows the distribution of the OLS beta estimates for daily returns. The mean is smaller than one and the distribution is skewed to the right indicating that the OLS beta estimates are possibly biased downward.

Dimson (1979) proposes a procedure called the "Aggregate Coefficients (AC)" method to estimate the market beta when securities are traded infrequently. The AC method uses lagged, contemporary, and leading market return as independent variables in the beta equation as follows:

$$r_{it} = a_i + \sum_{k=-n}^{n} b_{ik}\, r_{m\,t+k} + \varepsilon_{it} \quad t = 1, , \ldots, T.$$

The beta estimate ($b_{i\tau}$) is then defined as the sum of all coefficients.

$$b_{i\tau} = \sum_{k=-n}^{n} b_{i\tau k}\,.$$

Dimson shows that the AC method corrects the downward bias problem. The intuition is that lagged and leading market returns capture the serial correlation induced by infrequent trading. However there is no obvious optimal rule for selecting the number of lags and leads (n). It is plausible that if the correct number of lags and leads is selected, the regression of the out-of-sample residual[19] on a constant and excess market return should have zero slope coefficient. Table 3.6.1 shows the regression result for different number of lags and leads up to 5. Only when the number of lags and leads equal 0 or 1, we fail to reject the hypothesis that this slope parameter is equal to zero. Once the number of lags and leads is greater than 1, the estimated beta is inclined to be biased upward. We choose

[17] We take time average for each 105 stocks and use the 105 time averages to draw a histogram. Then the histogram is smoothed by using Gaussian Kernel.

[18] Under infrequent trading, current security price tends to reflect earlier prices which introduces positive serial correlation into its return and the market index. Positive serial correlation in the market index causes estimate of both its variance and covariance with individual stock return to be biased downward. In general, the downward bias in covariance estimate is greater, which means that beta estimate is biased downward.

the number of lags and leads to be one even though we fail to reject the null when n = 0 as well as n=1, because our prior is that when n = 0, we still have the infrequent trading problem. Table 3.6.1 also shows the grand mean of the estimated beta for each n where the grand mean is increasing with the number of lags and leads. The AC method shifts the distribution of beta estimates upward, possibly correcting the downward bias problem (see Figure 3.6.2).

Vasicek (1973) emphasizes the importance of using prior information when estimating market beta, and proposes a Bayesian type estimation method which is

$$b_s = wb + (1-w)b', \qquad w = \frac{1/v_b^2}{1/v_{b'}^2 + 1/v_b^2}.$$

where b = estimated market beta,

v_b^2 = estimate of variance of b,

b' = mean of prior distribution of market beta,

$v_{b'}^2$ = variance of prior distribution of market beta.

This method shrinks the sample estimate, b, toward the mean of the prior distribution. The degree of shrinkage depends on the variance of both the sample estimator and the prior distribution. For example, large v_b^2 leads to small w. This estimator requires us to specify the mean as well as the variance of the prior distribution. As Vasicek (1973) suggests, we can use cross-sectional information (mean and variance) to choose the mean and variance of the prior distribution. Our view is that this shrinkage method is not a substitute for the AC method, but a complementary method. Thus once we obtain an unbiased beta estimate using the AC method, then we apply the shrinkage method to the AC beta estimates. Figure 3.6.2 shows the distribution of beta estimates based on Vasicek's method applied to AC beta. First of all, we notice that all extreme values (maximum as well as minimum) are shrunk substantially toward the cross section mean. Vasicek's method shrinks the tail of the distribution and leaves the central part almost unchanged.

---

[19] See Appendix 2 for a detailed procedure to select the number of lags and leads.

On the other hand, a modified James-Stein shrinkage method shrinks the central part of distribution and leaves the tail part unchanged. This estimator takes the following form.

$$b_s = w^*b + (1-w)^*g, \qquad w = \left[1 - \frac{h}{(b-g)\mathrm{Var}(b)^{-1}(b-g)}\right]^+$$

where g is the mean of prior distribution. Note that $(b\text{-}g)\mathrm{Var}(b)^{-1}(b\text{-}g)$ is an F-statistic with degrees of freedom (n-1,1) or approximately a χ^2-statistic with one degree of freedom for the implicit hypothesis, H_0: b = g. We can write the weight as $\left[1-\frac{h}{F}\right]^+$.

When F is large (that is, we are likely to reject the implicit null hypothesis), w is large (or close to one), which means that we do not shrink. Here we do not have to specify the variance of the prior distribution. Instead, the shrinkage factor (h) needs to be specified because the number of regressors is 2. If we set h = 1 as in Wonnocott (1981), then P[F<h] = 0.6827 which is a big loss of continuity. We choose h = 0.4549 which make P[F<h] equal to 0.5. (See Table 3.6.2 for the relation between h and P[F<h]) Figure 3.6.2 shows the distribution of beta estimates based on the JS technique combined with the AC method. As expected, this method shrinks the central part of the distribution and the tail part is not changed.

(3) Ex-post Abnormal Return Estimation

Given five methods (See Table 3.6.3 for the list of five methods) available for us to estimate the market beta, we estimate ex-post abnormal returns as follows. For each time $\tau \in$ DFM, let $b_{i\tau}$ be one of 5 market beta estimates from the above regression. By defining $\hat{z}_{i\tau} \equiv {}^m r_{i\tau} - b_{i\tau}\, {}^m r_{m\tau}$, we can generate $\{\hat{z}_{i\tau}: i \in S, \tau \in \mathrm{DFM}\}$, which we call "ex-post abnormal returns". These are monthly abnormal returns up to the surprise in $b_{i\tau}$.

3.7 Calibration of the Ex-post Predictive Ability in Alpha Forecasts

We want to investigate how much of the ex-post abnormal returns can be explained by alpha forecasts. This explanatory power is well summarized by the usual R^2 statistic. However we can increase the R^2 simply by adding extra variables. In order to avoid overfitting the data, we use the adjusted R^2 (R^2_a) as our measurement for the predictive ability of alpha forecasts.

One problem in measuring R^2_a using the data set is that we have only 37 observations in the time dimension for each stock. Because we may not obtain a statistically meaningful measurement if we try to compute R^2_a for each stock, we assume[20] that all securities have the same coefficients and we pool the data. We denote the ex-post abnormal return by z_t instead of $\hat{z}_t$ whenever there is no ambiguity. If we assume for the moment the joint normality of z_t and $\hat{\alpha}_t$ as in Treynor and Black (1973), then

$$z_{i\tau_k} = a + b\hat{\alpha}_{i\tau_k} + \varepsilon_{i\tau_k} \quad i = 1, 2, \ldots, 105 \quad k = 1, 2, \ldots, 37 \quad \tau_k \in \text{DFM}.$$

Table 3.7.1 and Figure 3.7.2 show the simple regression results based on pooling 3885 observations. The estimation results are almost identical for the various beta estimation methods. Both the constant and the alpha term are significant and the sign of the alpha term is positive as one would expect if one believed that the security analysis has value. As indicated in the scatter diagram (Figure 3.7.1) of ex-post abnormal returns against alpha forecasts, the dispersion of ex-post abnormal returns is varying as the alpha forecasts change. White's Heteroscedasticity test (1980) confirms that we have enough evidence to reject the homoscedasticity assumption for the error term. When heteroscedasticity is present, the usual OLS covariance matrix is incorrect, which leads to incorrect inferences on coefficient parameters. Accordingly, we use the Heteroscedasticity Consistent Covariance Matrix Estimator[21] (HCCME) proposed by

[20] If each security or a group of securities is followed by individual analysts and they generate alpha and beta forecasts, then this assumption is hard to justify. In our case, however, the alpha and beta forecasts are generated by a computerized system, which supports our assumption.

[21] There are several ways of estimating the HCCME. The HCCME we estimate is

$$(X'X)^{-1}X'\hat{\Omega}X(X'X)^{-1}$$

where (1) $\hat{\Omega} = \text{Diag}[e_t^2/(1-h_t)]$

(2) e_t = regression residual

White (1980). It is noticeable that for every beta estimation, R^2_a is surprisingly small, all less than 0.0012.

When we add a quadratic term in the regression, the adjusted R^2 is increased for all beta estimation methods even though the coefficient itself is not significant (See Table 3.7.2). Adding a quadratic term has an impact on how one should act on negative alpha forecasts. When alpha is smaller than -1, we use the signal in the alpha forecasts in the reverse way (See Figure 3.7.2); we take a long position instead of a short position. In order to investigate possible asymmetric effects, we use a dummy variable that takes 1 when the alpha forecast is greater than zero and 0 otherwise. It turns out that using only positive alphas gives the best fit in terms of R^2_a. Both $R^2$and R^2_a are now higher than in the other two specifications. This specification implies that we take a long position proportional to the signal when the signal is positive and very small fixed long or short positions (depending on the sign of the constant estimate) when the signal is negative. What is interesting about these specifications is that they differ only for negative alpha forecasts (See Figure 3.7.2). That is to say, the linear and parabolic specifications exploit the information in negative alpha forecasts, but in an opposite way, while the kinked line specification roughly ignores the signal contained in negative alpha forecasts.

So far we have focused on the adjusted R^2 using all observation available up to the last date in DFM, that is $\tau = \tau_{37}$. If we vary τ from τ_1 to τ_{37}, then we can investigate the dynamics of the predictive ability. For each h $\in$ {1, 2, ..., 37}, we estimate the regression

$$z_{i\tau_k} = E(z_{i\tau_k} | I_{\tau_k}) + \varepsilon_{i\tau_k} \quad i = 1, 2, \dots, 105 \quad k = 1, 2, \dots, h$$

and compute $R^2_a(h)$. By plotting $R^2_a(h)$ against h, we can see how the predictive ability in alpha forecasts is changing over time. We display the results only for the case where the market beta is estimated using the Vasicek's Bayesian method combined with the AC

---

(3) $h_t = t^{th}$ element of the hat matrix $(X(X'X)^{-1}X)$.
this method has better finite-sample performance than simply using e_t or $\frac{n}{n-k} e_t$ as the diagonal elements in $\hat{\Omega}$. Even though MacKinnon and White (1985) find that "Jackknife" method is best, we do not attempt this in the benefit of computation time.

method. The results are almost identical to those for other beta estimation methods. Figure 3.7.3 clearly shows that the predictive power is declining over time.

We can summarize our findings concerning the alpha forecasts as follows. First, the predictive power in alpha forecasts is small, as indicated by the adjusted R^2. Secondly, both the quadratic effect and asymmetric effect increase the predictive power. Third, the noise level in negative alpha forecasts is higher than in positive alpha forecasts so that the use of the signal in the negative alpha forecasts depends on the conditional mean specification. Fourth, the predictive ability is declining over time.

3.8 Out-of-sample Analysis of the Treynor-Black Portfolio

In order to evaluate the TBP, we perform an out-of-sample analysis instead of in-sample analysis, for the following reasons. First, we want to avoid "overfitting the data" or "data mining" which is more likely to occur in the in-sample analysis. We think that out-of-sample analysis can reduce the likelihood of this happening. Second, we think the TBP framework is interesting not only in a theoretical viewpoint but also in a practical viewpoint, especially from the portfolio manager's viewpoint. An out-of-sample experiment would tell us how the TBP framework can add value in a practical manner. It should be made clear that when we generate out-of-sample performance starting at some point (the second DFM, τ_2 = 1/29/93) and rolling into the future, we use the only past and current information up to the time of making forecasts. The key parameters to be predicted during the out-of-sample experiment are listed below.

- Market beta
- Conditional mean of abnormal return
- Conditional variance of abnormal return
- Conditional covariance of abnormal returns
- Conditional mean of market return
- Conditional variance of market return

(1) Forecasting Market Beta

Beta forecasts have two effects on the out-of-sample performance of the TBP. In the first stage, the estimated "past" abnormal returns depend on beta forecasts. In the second stage, it determines the sign and magnitude of the TBP weight. We use 5 market beta estimation methods described earlier and daily returns with 3 year expanding estimation window in order to forecast market beta for each security.

(2) Forecasting Mean of Abnormal Return

In previous sections, we estimate ex-post abnormal returns using future information in order to measure the "ex-post" predictive ability in alpha forecasts. In the out-of-sample experiment, we are not allowed to use future information. However we still need to estimate "past" abnormal returns to investigate how past abnormal returns are correlated to past alpha forecasts.

For each prediction time, k = 2,3, ... , 37, we forecast market beta for each security using information available up to the time of estimation. Let $b_{i\tau_k}$ be one of 5 market beta forecasts.

We define "past" abnormal return $\hat{z}_{i\tau_{k-1}} = {}^m r_{i\tau_{k-1}} - \hat{b}_{i\tau_k} \, {}^m r_{i\tau_{k-1}}$. Note that for each k, the prediction time is τ_k and the information set at time τ_k (I_{τ_k}) is

$$I_{\tau_k} = \{ (\hat{z}_{i\tau_h}, \hat{\alpha}_{i\tau_h}) : i = 1, 2, \dots, 105 \quad h = 1, 2, \dots, k\text{-}1 \} \cup \{ \hat{\alpha}_{i\tau_k} \}$$

For each prediction index k = 2, 3, . . ., 37, we estimate

$$\hat{z}_{i\tau_h} = f(\hat{\alpha}_{i\tau_h}) + \varepsilon_{i\tau_h} \qquad h = 1, 2, \dots, k\text{-}1$$

using 3 specifications for the conditional mean (Line, Parabola and Kinked Line) and 5 estimation methods (OLS, NRLAD, JSLAD, OWLAD, LAD). Note that by assuming that all securities have the same coefficients, we have 105 observations for estimation at the prediction index k = 2, and 210 observations for k = 3, and so on. Once the coefficients are estimated, the forecast for the mean of abnormal return is given by

$$\hat{E}(\hat{z}_{i\tau_k} | I_{\tau_k}) = \hat{f}(\hat{\alpha}_{i\tau_k}).$$

This is our alpha forecast obtaining by discounting the raw, unadjusted alpha forecast from security analysis by its past performance.

(3) Forecasting Variance and Covariance of Abnormal Returns

In addition to predicting the conditional mean of abnormal returns, we need to make forecasts of the conditional variance of abnormal returns for each security and the conditional covariance of abnormal returns for all pairs of securities. We take the simplest specification for the conditional variance and covariance specification: the historical sample variance and covariance.

For each prediction time, k = 2,3, ... , 37, we estimate daily abnormal returns for each security using daily returns and 5 market beta estimates. We use 3 year expanding window. The daily abnormal return is defined by

$$e_{it} = r_{it} - b_{i\tau_k} r_{mt}$$

where $b_{i\tau_k}$ is one of market beta estimate and r_{it}, r_{mt} are daily returns available up to the time of prediction. We take the sample variance and sample covariance multiplied by the number of days in that month as our forecasts for the variance and covariance of monthly abnormal returns. We think that if we use better methods such as GARCH to forecast the conditional variance and covariance, we might make potential improvement. We leave this issue for future empirical work.

(4) Forecasting Market: Mean and Volatility

The square of the maximized S-Ratio of the TBP can be decomposed into to the square of the market's S-Ratio and the appraisal ratio. Under the Diagonal Model assumption,

$$\left[\frac{\mu_p(w^*)}{\sigma_p(w^*)}\right]^2 = \frac{\mu_m^2}{\sigma_m^2} + \sum_{i=1}^{N} \frac{\alpha_i^2}{\sigma_i^2}.$$

From this expression, Ferguson (1975) argues that the contribution of a market forecast is equivalent to that of a single security and "if market forecasting is no more or less difficult than security analysis, then the effort put into market analysis should be about equal to that put into following a single security". We think the contribution of a market forecast is more than that of a single security. Firstly, Ferguson's argument is based on the assumption that the optimal portfolio weight for both the AP and TBP is known. Secondly, the appearance of making the contribution of a market forecast equivalent to a single security disappears when off-diagonal terms are not equal to zero. Third, it turns out to be very crucial to obtain a good quality market forecast at the second stage of the TBP construction where we mix the AP with a market index.

There are few papers (See Merton (1980)) on forecasting market return while lots of research have been done for the market volatility. We use AR(0)-GARCH(1,1) specification as in Engle, Kane and Noh (1993) as follows.

$$r_t = \mu + \varepsilon_t$$

$$h_t = w + \alpha\varepsilon_{t-1}^2 + \beta h_{t-1} \qquad t = 1, 2, \ldots, T$$

Daily returns and 3 year rolling estimation window are used to forecast monthly return and volatility of the S&P500 index over February 1993 through January 1996. Once an estimation is done, the k-step ahead volatility prediction ($h_{T,T+k}$) is generated by

$$h_{T,T+1} = w + \alpha\varepsilon_T + \beta h_T$$

$$h_{T,T+k} = w + (\alpha + \beta)h_{T,T+k-1} \qquad k = 1, 2, \ldots, m$$

where m is the number of days in the target month. The forecast of that month is then defined by the sum of forward daily volatility forecasts as

$$\sum_{k=1}^{m} h_{T,T+k} .$$

The same method applies to the forecast of market return. The monthly return and volatility forecasts of the S&P500 index is given in Figure 3.8.1. Prediction Root MSE is 2.2032.

(5) Evaluation

The algorithm generates the TBP weight process $\{ w_{i\tau_k} \mid k = 2, 3, \ldots, 37\}$ and the realized monthly TBP return process $\{ r_{P\tau_k} \mid k = 2, 3, \ldots, 37\}$. We evaluate the out-of-sample performance of the market index and the TBP by the Sharpe Ratio (SR) defined as

$$SR = \frac{M}{SD}$$

where M and SD are ex-post sample mean and sample standard deviation respectively. We have assumed that the imaginary portfolio manager makes portfolio decision and executes buy and sell orders at the first date of the month. He is assumed to keep the same weight during the entire month. Therefore we can obtain daily TBP returns (783 observations) as well as monthly TBP returns (36 observations). We use daily returns to compute the ex-post mean and standard deviation of the TBP and the market index.[22]

Even though the Sharpe Ratio is extensively used in finance literature to evaluate a portfolio's performance, the meaning of "return per one unit of risk" is not clear. The

[22] The sample mean is same whether we use the monthly TBP return or the daily TBP return. However, the sample standard deviation depends on which frequency to use. We tend to observe that the sample standard deviation is smaller when we use monthly return for the S&P500 index. The sample standard deviation (annualized) is 7.8056 over 3 years from 2/1/93 to 1/31/96 when monthly returns are used. It is however 8.9802 when daily returns are used. Since this difference indicates possible serial correlation, we have tried other methods to compute the standard deviation.

(1) Spectral Density Approach

If daily returns (r_t) are serially correlated, then it can be expressed by $r_r = C(L)\varepsilon_t$. By exploiting the central limit theorem for serially dependent process, we can approximate the annualized variance of daily returns as

$$Var(\sum_{t=1}^{260} r_t) \approx 260C(1)^2\sigma^2$$

where C(1) is estimated by AR(p) approximation and the number of AR lags (p) is determined by some information criterion such as BIC. We have tried two BIC lag search methods: sequential simple to general method and general to simple method. The sequential simple to general method picks no lags so that the estimate of annualized standard deviation (8.99023) is same as the sample variance. However the general to simple method picks the 19th lag, which gives smaller estimate (8.0081).

(2) AR(0)-GARCH(1,1) Approach

Let $\{h_{t,t+1}\}$ be one day ahead variance forecast from the AR(0)-GARCH(1,1) specification over the 3 year prediction period. Then the annualized standard deviation can be estimated by $\sqrt{(\sum_{t=1}^{783} h_{t,t+1})/3}$. The estimated annualized standard deviation is 9.0862, which is approximately same as the sample standard deviation.

We think that the spectral density method using general to simple BIC lag search can be overfitting the data. Since other methods give reasonably similar estimate, we simply use sample standard deviation. On the other hand, we have found some evidence that the daily returns of the TBP shows stronger "serial correlation" depending on the beta estimation method, conditional mean specification, and estimation method. We are not certain what creates such serial correlation. Given that we re-balance the TBP every month, it is possible that daily portfolio returns could have different regime for some periods if the TBP weight is not stable. This kind of structural breaks can cause a "spurious serial correlation". In order to check this intuition, we have performed a small simulation generating two kinks of series. One is a random walk interrupted by multiple structural breaks with different mean and variance and the other one is a random walk interrupted by one time big shock. They show strong and consistent serial correlation which is not real. We do not attempt to develop a method to estimate the standard deviation in this situation. We leave it for further research.

M^2-measure used among practitioners is a modified version of the Sharpe Ratio. It is defined by

Portfolio's Sharpe Ratio × Market Risk - Market Return

which[23] measures the expected excess return of a portfolio over the benchmark (market) when its risk is the same as the market risk. We report both the annualized Sharpe Ratio and the annualized M^2-measure.

(6) Out-of-Sample Experiment and Discussion

We begin our experiment using the Diagonal Model and imposing no restriction on the TBP weight. Table 3.8.1 shows the Sharpe Ratio and M^2-measure for the Diagonal Model where we have 45 different TBPs (3 beta estimation methods × 3 conditional mean specifications × 5 estimation methods).[24] By looking at the sign of the M^2-measure in Table 3.8.1, we can see how the 45 TBPs perform relative to the S&P500 index. TBP is better than the S&P500 index in 22 out of 45 cases (about 50%). Table 3.8.2 indicates that both the return and the risk (std. dev.) of the TBP are large. The main reason seems that the ratio of the alpha of the AP to residual variance is over-estimated (See the formula for the TBP weight and its simplification in Theorem 3-3) and as a result the TBP weight assigned to the AP is large and volatile which makes the return and standard deviation of the TBP return large. In two cases (based on the OLS estimation), the Sharpe Ratio is even negative. The reason we want to mix the AP with a market index is

[23] Let r_p = excess return of a portfolio
σ_p= risk of a portfolio
r_m= excess return of a benchmark
σ_m= risk of a benchmark.

In the (r,σ) space, (r_p,σ_p) is a point and r_p/σ_p is the slope of the line, $r = (r_p/\sigma_p)\sigma$, passing through the point and the origin. This line can be interpreted as all return and risk combination which the portfolio can generate assuming the return and risk trade-off relationship is linear. When $\sigma = \sigma_m$, $r = (r_p/\sigma_p)\sigma_m$ which tells us how much return the portfolio can generate when its risk same as the market's. If we subtract the market return (r_m) from $(r_p/\sigma_p)\sigma_m$, it is the expected excess return of the portfolio over the market when its risk is same as the market risk. When we use the Sharpe Ratio ($S_p = r_p/\sigma_p$, $S_m = r_m/\sigma_m$), we are interested in whether S_p-$S_m > 0$. On the other hand, when we use the M^2-measure (M^2=$S_p\sigma_m$-r_m), we are interested in $M^2 > 0$. The following relationship can be proven.

$$S_p\text{-}S_m \geq 0 \Leftrightarrow M^2 \geq 0.$$

[24] We report only three beta estimation methods (Beta Forecasts, Vasicek's method applied AC beta, JS method applied to AC betas) which are supposed to be superior to the other two methods (OLS, AC).

for the benefit of diversification. By taking a large short position on the market, we have "concentration" rather than diversification. However, when we restrict the TBP weight to be between [0,1], the performance improves and the TBP is better than the market in many cases. This restriction makes the TBP weight stable and reduces the risk of the portfolio. When the [0,1] restriction is imposed, the TBP weight tends to be equal to 1. This means that the TBP is same as the AP. From this observation, we think that the TBP model and alpha forecasts are capable of generating the AP better than the market. The problem seems to be not in the predictive ability but in the improper mix of the AP and the market.

In the Diagonal Model without restriction on the TBP weight, the OLS estimator is always dominated by the LAD estimator and shrinkage LAD estimators (See symbol * in Table 3.8.1). Given that the OLS estimator is not stable and is sensitive to outliers and the LAD estimator is robust to outliers in the dependent variable, this result is not surprising. Interestingly, the JSLAD estimator tends to achieve the best performance among shrinkage LAD estimators even though the HCLAD is theoretically best. In many cases the JSLAD is better than both estimators. This can be explained by looking at the return and risk table (See Table 3.8.2). In most cases the return of the JSLAD estimator is a convex combination and is smaller than one of its components. The same thing happens to the risk, but risks tend to be much smaller which leads to the improved Sharpe Ratio.

The effect of the quadratic term is mixed. However the kinked line specification is uniformly better than the other specifications no matter what the [0,1] restriction is imposed on the TBP weight. Therefore, in the Diagonal Mode, using the signal from positive alpha forecasts and ignoring the signal from negative alpha forecasts gives the best performance. Given that the adjusted R^2 is highest in the kinked line specification, the adjusted R^2 is an important criterion to measure the predictive ability of alpha forecasts.

Comparing the result with Table 3.8.5 shows that managing the market risk (market beta) properly is an important ingredient in obtaining better TBP. The TBP based on the beta estimates without any adjustment is not better in most cases than the TBP based on the market beta adjusted by AC method along with Bayesian and JS shrinkage. Those managed market beta estimates also generate higher Sharpe Ratios than the beta forecasts itself when no [0,1] restriction is imposed.

The out-of-sample performance for the Covariance Model is shown in Table 3.8.3 and Table 3.8.4. The effect of using off-diagonal terms[25] improves the S-Ratio and M^2-measure in many cases. The Covariance Model is better than the Diagonal Model in 39 out of 45 cases when no restriction is imposed on the TBP weight (Count the number of the symbol (+) in Table 3.8.3). Using off-diagonal terms favors the LAD estimator, the JSLAD estimator and the kinked line specification. The other properties in the Diagonal Model discussed earlier are fairly preserved in the Covariance Model. First, the OLS estimator is dominated by the LAD estimator and shrinkage LAD estimators. Second, the effect of using the quadratic term is mixed and the kinked line specification is best. As far as the comparison of estimators is concerned, the JSLAD is best in 7 out of 9 cases when there is no restriction on the TBP weight. When the [0,1] restriction is imposed, the LAD estimator and the JSLAD estimator are better than the other estimators mostly, and the OWLAD is best in two cases where the performance of the OLS estimator is not much worse than the LAD estimator. What is surprising is that even though R^2 and adjusted R^2 are so small (about 0.001-0.002 as seen in section 3.7), there is a potential gain of using the properly managed Covariance TBP.

Theoretically we prefer the Covariance Model to the Diagonal Model because we use all information available. We also prefer the Vasicek's beta estimation to the Beta Forecast and the James-Stein beta estimation because the Beta Forecast is not on our control and the shrinkage factor in the James-Stein beta estimation is arbitrary. In this theoretically preferred block, we can compare the performance of each estimator. The M^2-measure is 2.948, 4.910, 5.958, 4.868 and 5.948 for the OLS, NRLAD, JSLAD, OWLAD and LAD estimator respectively. The maximum is given by the JSLAD

[25] In this experiment, we use all off-diagonal without testing whether or not they are significant. For one run (Covariance Model, No [0,1] restriction, Vasicek's beta estimation, Kinked Line specification and

estimator. Even though this is not the global maximum (which is 6.758 given by the Covariance Model, Beta Forecast, Kinked Line, and the JSLAD estimator), we discuss this portfolio in detail.

Figure 3.8.2 shows how the TBP weight for the 3 estimators is changing over time. Note that the weight given to the AP is very large and volatile. The closed form solution for the TBP weight is given by *Theorem 3-3* in Section 3.3. Since β_A is a weighted average of cross-sectional beta forecasts which is close to one, we assume that $\beta_A = 1$ for now. Then the TBP weight (w) is simplified as

$$w = \frac{\dfrac{\alpha_A}{h^{*\prime}\Omega h^{*}}}{\dfrac{\mu_m}{\sigma_m^2}}.$$

The numerator is the ratio[26] of the alpha of the AP and the residual variance and the denominator is the ratio of the mean and variance of the market index. Figure 3.8.1 and Figure 3.8.3 clearly show that over the prediction period the ratio in the numerator is much greater than the ratio in the denominator. If we are willing to assume that our market forecast is reasonable, then the source of the over-estimation of the TBP weight comes from the over-estimation of the mean-variance ratio in the numerator.[27] The effect of the deviation of β_A from 1 is that as the distance between β_A and zero is smaller, it makes the TBP weight smaller.[28] By comparing Figure 3.8.2 and Figure 3.8.4, we notice that the high peaks in the TBP weight are corresponding to the high peaks in the forecast of the beta of the AP.[29] Figure 3.8.5 shows the time path of each estimate in the preferred block. As expected, shrinkage estimates are located between the OLS estimate and the

---

JSLAD estimation), a simple test shows that 3170 off-diagonal terms (out of 5460) are significant on average at 10% level and 2273 terms at 5% level.

[26] Note that this ratio is equivalent to what we maximize in order to obtain the AP weight.

[27] Figure 8-3 reveals that the forecast of the residual variance is surprisingly flat over the entire sample period while the actual residual return is volatile. The initial estimation window is 3 years and expansion window is used to increase the precision of the sample covariance matrix. Estimating unconditional moments instead of conditional moments might possibly be a source of the flat volatility.

[28] It can be shown that $\Delta w^*/\Delta\beta_A > 0$ if the mean of the Minimum Variance Portfolio (MVP) on the frontier generated by the AP and the market index is positive; $(1-\beta_A)\alpha_A\sigma_m^2+\mu_m h^{*\prime}\Omega h^* > 0$.

LAD estimate. The OLS estimate is greater than the other estimates which leads to more over-estimated alpha of the AP and more over-estimated TBP weight. This might be one of reasons why the OLS estimator is always dominated by the LAD estimator and shrinkage LAD estimators.

If the market is in equilibrium all the time, then the alpha for any security is zero. In this case we have zero AP and the TBP is same as the market portfolio. In this sense, market equilibrium is our default position or "the mean of the prior distribution of the TBP". Shrinkage of coefficients toward zero using Ridge estimation allows us to embody this prior information in the process of estimating the conditional mean of abnormal returns. Furthermore, shrinkage toward zero could stabilize the TBP weight.[30] The zero point serves as the mean of the prior distribution and we choose the OLS covariance matrix different only up to a scale (λ_0^2;shrinkage factor) for the prior covariance.[31] We determine the optimal value for the shrinkage factor (λ_0^2) by minimizing the Prediction Mean Squared Errors (PMSE). Table 3.8.6 shows the results of the Ridge estimation applied to the theoretically preferred block. The Sharpe Ratio and M^2 are both decreased, but return and risk are stabilized. The TBP weight is also stabilized shown in Figure 3.8.6. The JSLAD estimator still gives the maximum M^2-measure in this block. We compute the Wealth Index using this TBP. The Wealth Index for the TBP, the AP and the S&P500 index is in Figure 3.8.7. The AP is moderately outperforms the market. On the other hand, the TBP generated by putting a large weight on the AP displays a huge return as well as risk. The superiority of the TBP mainly comes from the early period. This observation can be well explained by the dynamics of predictive ability we have

---

[29] The definition of β_A is $h^{*\prime}\beta$ where h^* is the forecast of the AP. We observe that the sample correlation between the forecast of the AP weight and beta forecast tends to be negative. When it is positive occasionally, we have high peaks in the forecast of β_A.

[30] Note that $\Delta w^*/\Delta\alpha_A > 0$ if the mean of the MVP is positive; $(1-\beta_A)\alpha_A\sigma_m^2+\mu_m h^{*\prime}\Omega h^* > 0$. Hence, if we shrink coefficients, then we decrease α_A which in turn decrease w^*. However, if $(1-\beta_A)\alpha_A\sigma_m^2+\mu_m h^{*\prime}\Omega h^* < 0$, then we take long position on the risk-free asset and short position on the TBP. In other words, we reverse the sign of the TBP weight. In this case, $\Delta w^*/\Delta\alpha_A < 0$ and shrinkage toward zero can increase the instability in the TBP weight.

[31] Suppose $y = X\beta + \varepsilon$ where $\varepsilon \sim N(0,\sigma^2 I)$. The MLE is given by $b = (X'X)^{-1}X'y$. In this case, our assumption on the prior distribution is $\beta \sim N(0,\lambda_0^2(X'X)^{-1})$. Then the ridge estimator $b_R(\lambda)$ is given by $b_R(\lambda) = \lambda b$ where $\lambda = (1/\sigma^2)/[(1/\sigma^2)+(1/\lambda_0^2)]$. The same estimated ridge parameter applies to the LAD and JSLAD estimator.

investigated in section 3.7 (See Figure 3.7.3) where we have found that the predictive power is very good in early period and is declining over time. Out-of-sample experiment shows that a $1 invested in a properly managed TBP Covariance Model would yield $1.810 with the Sharpe Ratio being 1.340 over 3 years. On the other hand, if you invest $1 in the S&P500 index over 3 years, the final wealth is $1.259 and the Sharpe Ratio is 0.909. This result shows that a large potential value can be obtained and it can be done without requiring a big threshold of forecasting ability.

3.9 Conclusion

Treynor-Black Portfolio model has a potential to be a valuable method for active portfolio management when a great deal of effort to refine abnormal return forecasts and market risk is taken. In the paper, we have taken several specifications for the conditional mean of abnormal returns and used various estimation methods in order to refine alpha forecasts. The adjusted R^2 turns out to be a good measure for the predictive ability contained in alpha forecasts. In order to carry market risk management, we have used Dimson's Aggregate Coefficient method coupled with Bayesian and James-Stein shrinkage. Out-of-sample experiments show that the conditional mean specification, estimation method and market risk management all together play a significant role in obtaining better performance of the TBP model.

The Covariance Model shows better performance than the Diagonal Model. We have used simply sample residual covariance matrix, but more advanced econometric tools to forecast the covariance matrix seems necessary. The use of the newly developed shrinkage LAD estimators as a tool to extract predictive ability from the raw alpha forecast turns out to be useful. Out-of-sample experiments show that even though the predictive power measured by the adjusted R^2 is as low as 0.0015, the potential value of using the TBP Model is large, which indicates that the minimum threshold of forecasting power may be lower than many academicians currently think. However, it turns out that despite its superiority in terms of both the Sharpe Ratio and the Wealth Index, the TBP weight is not stable, which we conjecture might be caused by the over-estimation of the ratio the alpha of the AP to residual variance. Ridge estimation is helpful to reduce the instability in the TBP weight. When we impose the [0,1] restriction on the portfolio weight, the TBP is stabilized and still better than the market index. We have found that the OLS estimator is always dominated by the LAD estimator and shrinkage LAD estimators.

In this study, we have focused on the stratified random sample only. We need to extend the same study for the total 600 stocks. We have not performed a formal statistical procedure to test whether the difference between the best TBP's and the market's Sharpe Ratio is statistically greater than zero. We believe that White's (1996) bootstrap method can be used for this purpose. This paper has not addressed transaction costs and management fees which are important issues in an active portfolio management. We think we need to incorporate these issues in the TBP framework. In order to measure what is the minimum required predictive power in the TBP Model to achieve a given level of Sharpe Ratio, a simulation study like Hodges & Brealey (1973) is needed. In this paper, we divided the 3 month ahead alpha forecast by 3 and treated it as 1 month ahead forecast for simplicity. If we want to use the full information, we might need to investigate some interesting issues such as how to update old forecasts when new forecast information arrives within the old forecast's horizon because new forecasts arrive every month. We leave these issues for future research.

Appendix 1

Proof of Theorem 3-1

According Theorem 2 in Roll (1977) the weight vector (h) is given by

$$h = \sigma_0^2 \Omega^{-1} \left[\frac{\alpha - r^0 \iota}{\mu_0 - r^0} \right]$$

where μ_0 and σ_0^2 are the mean return and variance of the Minimum Variance Portfolio (MVP) on the Active Efficient Portfolio. r^0 is the reference rate of return (usually riskless rate of return) with respect to which the tangential portfolio is computed. In our case, r^0 = 0. Corollary 2 in Roll (1977) also gives us the formulas for the mean return and variance of the MVP as follows.

$\mu_0 = b/c, \ \sigma_0^2 = 1/c$

where b, c are elements of the Information Matrix (A) defined as

$$A \equiv [\alpha\ \iota]'\Omega^{-1}[\alpha\ \iota] = \begin{bmatrix} \alpha'\Omega^{-1}\alpha & \alpha'\Omega^{-1}\iota \\ \alpha'\Omega^{-1}\iota & \iota'\Omega^{-1}\iota \end{bmatrix} \equiv \begin{bmatrix} a & b \\ b & c \end{bmatrix}.$$

Hence $h = b^{-1}\Omega^{-1}\alpha = [\alpha'\Omega^{-1}\iota]^{-1}\Omega^{-1}\alpha$ which completes the proof. Q.E.D.

Proof of Theorem 3-3

Considering that we have only two assets to combine, we define their mean vector and co-variance matrix as follows.

$$\mu \equiv \begin{bmatrix} \mu_A \\ \mu_m \end{bmatrix} = \begin{bmatrix} \alpha_A + \beta_A\mu_m \\ \mu_m \end{bmatrix}.$$

$$\Sigma \equiv \begin{bmatrix} \sigma_A^2 & \sigma_{Am} \\ \sigma_{Am} & \sigma_m^2 \end{bmatrix} = \begin{bmatrix} \beta_A^2\mu_m^2 + h^{*\prime}\Omega h^* & \beta_A\sigma_m^2 \\ \beta_A\sigma_m^2 & \sigma_m^2 \end{bmatrix}.$$

Then the TBP is the solution to the following maximization problem.

$$\underset{w}{Max}\ \frac{w'\mu}{\sqrt{w'\Sigma w}} \quad \text{subject to } w'\iota = 1$$

where w is 2×1 vector. This is the tangential portfolio on the efficient frontier constructed based on the AP and the market portfolio. By the same reasoning in the proof of Theorem 3-1, the optimal weight is given by

$$w^* = (\mu'\Sigma^{-1}\iota)^{-1}\Sigma^{-1}\mu$$

$$= \frac{\mu_A\sigma_m^2 - \mu_m\sigma_{Am}}{\mu_A\sigma_m^2 - \mu_m\sigma_{Am} + \mu_m\sigma_A^2 - \mu_A\sigma_{Am}}.$$

Note that

(1) $\mu'\Sigma^{-1}\iota = (\sigma_m^2 h^{*\prime}\Sigma h^*)^{-1}[(1-\beta_A)\alpha_A\sigma_m^2 + \mu_m h^{*\prime}\Sigma h^*]$.

(2) $\Sigma^{-1}\mu = (\sigma_m^2 h^{*\prime}\Omega h^*)^{-1}\begin{bmatrix} \alpha_A\sigma_m^2 \\ -\alpha_A\beta_A\sigma_m^2 + \mu_m h^{*\prime}\Omega h^* \end{bmatrix}$.

Therefore,

$$w_1^* = \frac{\alpha_A\sigma_m^2}{(1-\beta_A)\alpha_A\sigma_m^2 + \mu_m h^{*\prime}\Omega h^*}$$

$$w_2^* = \frac{-\alpha_A\beta_A\sigma_m^2 + \mu_m h^{*\prime}\Omega h^*}{(1-\beta_A)\alpha_A\sigma_m^2 + \mu_m h^{*\prime}\Omega h^*}. \quad \text{Q.E.D.}$$

Appendix 2: How to Select AC Lags and Leads: Out-of-sample Correlation Between Ex-post Abnormal returns and Excess Market Return

For each prediction time $\tau \in DFM \equiv \{\tau_i : i = 1, 2, \dots, 37\}$, we estimate the following CAPM equation using daily returns.

$$r_{it} = a_i + \sum_{k=-n}^{n} b_{ik}\, r_{m\,t+k} + \varepsilon_{it} \qquad t = 1, 2, \dots, \tau \quad i = 1, 2, \dots, N$$

The beta estimate (b_i) is then defined as the sum of all coefficients.

$$\hat{b}_{i\tau}(n) = \sum_{k=-n}^{n} \hat{b}_{ik}(n).$$

Then we generate ex-post abnormal returns using the predicted beta and monthly excess return, market return as follows.

$$\hat{z}_{i\tau}(n) \equiv {}^{m}r_{i\tau} - \hat{b}_{i\tau}(n)\, {}^{m}r_{m\tau}.$$

We should expect that if the beta estimate is unbiased, then the correlation between out-of-sample ex-post abnormal returns and the excess market return should be zero. The deviation from zero indicates the bias in beta estimate. We regress the deviation of ex-post abnormal returns from unadjusted alpha forecasts on a constant and the excess market return.

$$\hat{z}_{i\tau}(n) - \hat{\alpha}_{i\tau} = c_{0i} + c_{1i}\, {}^{m}r_{m\tau} + \varepsilon_{i\tau} \quad i = 1, 2, \dots, N \qquad \tau \in \text{DFM}$$

Since we expect $c_{0i} = c_{1i} = 0$ for all i, we pool the data and estimate

$$\hat{z}_{i\tau}(n) - \hat{\alpha}_{i\tau} = c_0 + c_1\, {}^{m}r_{m\tau} + \varepsilon_{i\tau} \quad i = 1, 2, \dots, N \qquad \tau \in \text{DFM}$$

If the slope coefficient is significantly positive, it indicates that the estimated beta is downward biased. If negative, it means upward bias. On the other hand, the deviation of the intercept coefficient from zero indicates a bias in the unadjusted alpha forecast, but only when there is no bias in beta estimate.

Reference

Ambachtsheer, K. (1974), Profit Potential in an Almost Efficient Market, *Journal of Portfolio Management*, Fall, 84-87.

Ambachtsheer, K. (1977), Where Are the Customers' Alphas?, *Journal of Portfolio Management*, Fall, 53-56.

Ambachtsheer, K. and James L. Farrell, Jr. (1979), Can Active Management Add Value?, *Financial Analysts Journal*, November-December, 39-47.

Bates, Charles E. and Halbert White (1993), Determination of Estimators with Minimum Asymptotic Covariance Matrices, *Econometric Theory*, 9, 633-648

Barrodale, I. and Roberts, F.D.K. (1974), Algorithm 478: Solution of an Over-determined System of Equations in the L_1 Norm, *Communications of the Association for Computing Machinery*, 17, 319-320.

Black, F. (1973), Yes, Virginia, There Is Hope: Tests of the Value Line Ranking System. *Financial Analysts Journal*, September-October, 10-14.

Bodie, Zvi, Alex Kane and Alan J. Marcus (1996), *Investment*, Third Edition, Irwin.

Charnes, A. and C. E. Lemke (1954), Computational Theory of Linear Programming: The Bounded Variables Problem, Graduate School of Ind. Administration, Carnegie Inst. Of Technol., Pittsburgh, Pennsylvania.

Dimson, E., (1979), Risk Measurement When Shares Are Subject to Infrequent Trading, *Journal of Financial Economics* 7, 197-226.

Engle, F. Robert, Alex Kane and Jaesun Noh, (1993), Index-Option Pricing with Stochastic Volatility and The Value of Accurate Variance Forecasts, *NBER Working Paper*, No 4519.

Fama, Eugene F. and Kenneth R. French, (1992), The Cross-Section of Expected Stock Returns, *The Journal of Finance*, Vol. XLVII, No. 2, June 1992, 427-465.

Ferguson, Robert, (1975), Active Portfolio Management: How to Beat The Index Funds, *Financial Analysts Journal*, May-June, 63-72.

Green, Edwin J. and Strawderman William E. (1991), James-Stein Type Estimator for Combining Unbiased and Possibly biased Estimators, *Journal of the American Statistical Association,* Vol 86, No 416, 1001-1006.

Grossman, S. J., and L. E. Stiglitz (1980), On the Impossibility of Informationally Efficient Markets, *American Economic Review*, 70, 393-408.

Hodges, S. D. and R. A. Brealey, (1973), Portfolio Selection in a Dynamic and Uncertain World, *Financial Analysts Journal*, March-April, 50-65.

Kane, Alex and Alan J. Marcus, (1986), The Valuation of Security Analysis, *NBER Working Paper*, No 1958.

MacKinnon, J. G., and H. White (1985), Some Heteroskedasticity Consistent Covariance Matrix Estimators with Improved Finite Sample Properties, *Journal of Econometrics*, 29, 305-325.

Merton, Robert C., (1980), On Estimating The Expected Return on The Market: An Exploratory Investigation, *NBER Working Paper Series*, No 444.

Roll, Richard (1977), A Critique of the Asset Pricing Theory's Tests, *Journal of Financial Economics*, 4, 129-176.

Rosenberg, Barr, Kenneth Reid, and Ronald Lanstein (1985), Persuasive Evidence of Market Inefficiency, *Journal of Portfolio Management* 11, 9-17.

Schmoyer, R. and Arnold, S. (1989), Shrinking Techniques for Robust Regression, *Contributions to Probability and Statistics: Essays in Honor of Ingram Olkin*, 368-384.

She, P. (1992), Another Solution, *Bulletin of the Institute of Mathematical Statistics*, 21, 234.

Treynor, Jack L. and Fisher Black (1973), How to Use Security Analysis to Improve Portfolio Selection, *The Journal of Business*, Vol. 46, No. 1, 66-86.

Vasicek, O. A., (1973), A Note on Using Cross-Sectional Information in Bayesian Estimation of Security Betas, *The Journal of Finance*, Vol 28, 1233-1239.

White, H (1997), Various Methods for Measuring Value at Risk, manuscript, University of California, San Diego.

White, H (1996), Data Snooping with Confidence, *manuscript*, University of California, San Diego.

White, H (1993), Estimation, Inference and Specification Testing,

White, H. (1980), A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity, *Econometrica*, 48, 847-838.

Wonnacott, T. H. and R. J. Wonnacott (1981), Regression: A Second Course in Statistics, *New York: Wiley*.

Table 3.4.1 Date Forecasts Made

Year	Month	Date	Prediction Index	Number of days between DFM & Buy and Sell Execution Date
1992	12	31	1	0
1993	01	29	2	0
1993	02	26	3	0
1993	03	26	4	3
1993	04	30	5	0
1993	05	28	6	1
1993	06	25	7	3
1993	07	30	8	0
1993	08	27	9	2
1993	09	24	10	4
1993	10	29	11	0
1993	11	26	12	2
1993	12	31	13	0
1994	01	28	14	1
1994	02	25	15	1
1994	03	31	16	0
1994	04	29	17	0
1994	05	27	18	2
1994	06	24	19	4
1994	07	29	20	0
1994	08	26	21	3
1994	09	30	22	0
1994	10	28	23	1
1994	11	25	24	3
1994	12	30	25	0
1995	01	27	26	2
1995	02	24	27	2
1995	03	31	28	0
1995	04	28	29	0
1995	05	26	30	3
1995	06	30	31	0
1995	07	28	32	1
1995	08	25	33	4
1995	09	29	34	0
1995	10	27	35	2
1995	11	24	36	4
1995	12	29	37	0

Note: Buy and Sell Execution Date = Buy and Sell Decision Date = The First Business Date in The Month

Table 3.4.2 Summary Statistics for Alpha & Beta Forecasts

	Mean	Std.	Minimum	25%	50%	75%	Maximum
Alpha	-1.355	2.052	-5.404	-2.783	-1.742	-0.269	8.135
Beta	0.981	0.266	0.208	0.847	0.986	1.141	1.703

Table 3.6.1 Lag Selection for the Aggregate Coefficient Method

# Lags	Intercept	Slope	R^2	R^2_a	Prob(F)	Mean of Beta
0	1.5364004	0.0522124	0.0001944	-0.0000631	0.3849372	0.8810
	(0.1383879)	(0.0600879)				
1	1.5367453	-0.0672165	0.0003213	0.0000639	0.2639868	0.9995
	(0.1385682)	(0.0601662)				
2	1.5381416	-0.1120830	0.0008924	0.0006351	0.0626334	1.0447
	(0.1386105)	(0.0601846)				
3	1.5381416	-0.1120830	0.0008924	0.0006351	0.0626334	1.0447
	(0.1386105)	(0.0601846)				
4	1.5387283	-0.1704644	0.0020579	0.0018009	0.0046828	1.1029
	(0.1387398)	(0.0602407)				
5	1.5402138	-0.1736378	0.0021271	0.0018701	0.0040367	1.1056
	(0.1390003)	(0.0603538)				

Note: Standard errors are in parenthesis.

Table 3.6.2 Relationship between Degree of Discontinuity and Shrinkage Factor

P[F<h]	0.1	0.2	0.3	0.4	0.5	0.6	0.7
h	0.0158	0.0642	0.1485	0.2750	0.4549	0.7083	1.00742

Table 3.6.3 Estimation Methods for Market Beta

Method	Description
1	OLS estimation with no restriction
2	Institution's Beta Forecasts
3	AC method with n = 1
4	Vasicek's Bayesian estimation applied to AC beta
5	James-Stein Shrinkage estimation applied to AC beta

Table 3.7.1 Linear Specification

Beta Estimation	Constant	Alpha	R^2	R^2_a	White's Heteroscedasticity Test
1	0.3970 (0.1333)	0.1909 (0.0944)	0.001341	0.001084	92.6821 (0.0000)
2	0.3371 (0.1330)	0.1940 (0.0942)	0.001391	0.001134	92.7363 (0.0000)
3	0.3122 (0.1336)	0.1872 (0.0950)	0.001286	0.001028	94.9599 (0.0000)
4	0.3210 (0.1334)	0.1911 (0.0947)	0.001342	0.001085	93.4382 (0.0000)
5	0.3154 (0.1336)	0.1892 (0.0950)	0.001314	0.001057	94.9233 (0.0000)

Note: (1) Heteroscedasticity Consistent Covariance Matrix used for the standard errors

(2) P-value in parenthesis for White's Heteroscedasticity Test

Table 3.7.2 Parabolic Specification

Beta	Constant	Alpha	$Alpha^2$	R^2	R^2_a	White's Heteroscedasticity Test
1	0.2808 (0.1523)	0.1914 (0.0946)	0.0554 (0.0455)	0.001841	0.001327	98.1026 (0.0000)
2	0.2257 (0.1521)	0.1945 (0.0943)	0.0531 (0.0453)	0.001851	0.001337	98.4132 (0.0000)
3	0.1936 (0.1526)	0.1877 (0.0952)	0.0565 (0.0462)	0.001804	0.001290	99.5293 (0.0000)
4	0.2037 (0.1525)	0.1916 (0.0948)	0.0559 (0.0458)	0.001851	0.001337	98.4014 (0.0000)
5	0.1975 (0.1526)	0.1898 (0.0952)	0.0562 (0.0462)	0.001827	0.001313	99.6631 (0.0000)

Note: (1) Heteroscedasticity Consistent Covariance Matrix used for the standard errors

(2) P-value in parenthesis for White's Heteroscedasticity Test

Table 3.7.3 Kinked Linear Specification

Beta	Constant	Alpha*(Alpha>0)	R^2	R^2_a	White's Heteroscedasticity Test
1	0.1823 (0.1198)	0.3939 (0.1960)	0.001793	0.001536	92.2511 (0.0000)
2	0.1225 (0.1194)	0.3896 (0.1951)	0.001762	0.001505	92.3440 (0.0000)
3	0.1000 (0.1200)	0.3910 (0.1979)	0.001762	0.001505	94.4778 (0.0000)
4	0.1057 (0.1198)	0.3952 (0.1967)	0.001804	0.001547	92.9745 (0.0000)
5	0.1016 (0.1199)	0.3930 (0.1978)	0.001781	0.001524	94.4409 (0.0000)

Note: (1) Heteroscedasticity Consistent Covariance Matrix used for the standard errors

(2) P-value in parenthesis for White's Heteroscedasticity Test

Table 3.8.1 Sharpe Ratio and M^2-measure: Diagonal Model (S&P500 Sharpe Ratio = 0.909)

		Sharpe Ratio					M^2-measure				
		OLS	NRLAD	JSLAD	OWLAD	LAD	OLS	NRLAD	JSLAD	OWLAD	LAD
	No Restriction imposed on TBP weight										
Beta	Line	-0.360	0.169	0.824	0.144	0.823	-11.400	-6.650	-0.769 *	-6.874	-0.773
	Parabola	-0.043	0.181	0.943	0.182	0.942	-8.549	-6.543	0.302 *	-6.528	0.294
	Kinked	1.096	1.218	1.481	1.218	1.481	1.680	2.770	5.132 *	2.771	5.130
Vasicek	Line	0.406	0.843	0.638	0.827	0.637	-4.523	-0.595 *	-2.439	-0.742	-2.443
	Parabola	0.021	0.585	0.720	0.583	0.719	-7.981	-2.914	-1.701 *	-2.926	-1.711 *
	Kinked	1.052	1.243	1.293	1.238	1.292	1.284	2.995	3.448 *	2.953	3.436
JS	Line	0.327	0.783	1.005	0.774	1.005	-5.232	-1.136	0.861 *	0.774	0.855
	Parabola	0.264	0.505	0.660	0.504	0.659	-5.796	-3.629	-2.240 *	-3.642	-2.250
	Kinked	0.973	1.165	1.216	1.164	1.215	0.575	2.295	2.755 *	2.283	2.745
	[0, 1] Restriction imposed on TBP weight										
Beta	Line	0.895	0.922	1.033	0.921	1.025	-0.131	0.112	1.033	0.107	1.036 *
	Parabola	0.934	0.951	0.999	0.952	0.999	0.218	0.374	0.802 *	0.384	0.802 *
	Kinked	1.076	1.095	1.235	1.095	1.235	1.501	1.667	2.920	1.668	2.921 *
Vasicek	Line	0.905	0.957	1.058	0.955	1.058	-0.037	0.427	1.337 *	0.413	1.337 *
	Parabola	0.956	0.989	0.963	0.989	0.963	0.420	0.713	0.485	0.714 *	0.480
	Kinked	1.094	1.119	1.146	1.118	1.146	1.654	1.886	2.124 *	1,878	2.123
JS	Line	0.904	0.965	0.941	0.963	0.940	-0.050	0.497 *	0.287	0.481	0.279
	Parabola	0.957	0.986	0.861	0.987	0.859	0.430	0.691	-0.433	0.696 *	-0.451
	Kinked	1.095	1.120	1.145	1.120	1.145	1.665	1.894	2.114 *	1.890	2.112

Note: * indicates the best estimator in the row.

Table 3.8.2 TBP Return and Risk: Diagonal Model (S&P500 Return = 8.166: S&P500 Risk = 8.980)

		Sharpe Ratio					M^2-measure				
		OLS	NRLAD	JSLAD	OWLAD	LAD	OLS	NRLAD	JSLAD	OWLAD	LAD
	No Restriction imposed on TBP weight										
Beta	Line	-39.403	14.396	70.700	12.180	70.498	109.423	85.295	85.831	84.649	85.629
	Parabola	-7.474	97.310	24.341	94.105	24.350	175.282	538.260	25.812	515.963	25.847
	Kinked	34.356	34.664	43.248	34.743	43.247	31.334	28.465	29.204	28.527	29.208
Vasicek	Line	83.036	77.618	45.198	74.916	45.210	204.691	92.058	70.872	90.615	70.937
	Parabola	1.052	17.222	14.879	17.142	14.848	51.008	29.445	20.669	29.380	20.657
	Kinked	35.983	32.839	24.499	32.742	24.458	34.193	26.422	18.943	26.442	18.930
JS	Line	95.752	58.326	49.408	47.888	49.360	293.107	74.500	49.151	74.827	49.119
	Parabola	11.937	15.613	14.948	15.526	14.915	45.220	30.900	22.650	30.817	22.641
	Kinked	34.250	30.611	22.722	30.615	22.691	35.186	26.278	18.684	26.311	18.675
	[0, 1] Restriction imposed on TBP weight										
Beta	Line	8.799	9.286	10.706	9.262	10.710	9.833	10.072	10.451	10.054	10.452
	Parabola	8.647	8.749	8.957	8.757	8.956	9.261	9.199	8.969	9.198	8.968
	Kinked	9.539	9.581	10.794	9.585	10.796	8.861	8.750	8.744	8.752	8.744
Vasicek	Line	8.872	9.294	11.637	9.277	11.641	9.800	9.713	10.996	9.712	11.000
	Parabola	8.860	8.749	9.023	8.748	9.024	9.267	8.849	9.367	8.846	9.373
	Kinked	9.725	9.623	9.403	9.618	9.400	8.893	8.597	8.206	8.600	8.205
JS	Line	8.899	9.323	11.577	9.307	11.583	9.846	9.664	12.299	9.666	12.317
	Parabola	8.895	8.746	9.429	8.746	9.435	9.292	8.867	10.950	8.863	10.983
	Kinked	9.758	9.601	9.323	9.601	9.320	8.914	8.570	8.144	8.574	8.143

Table 3.8.3 Sharpe Ratio and M^2-measure: Covariance Model (S&P500 Sharpe Ratio = 0.909)

		Sharpe Ratio					M^2-measure				
		OLS	NRLAD	JSLAD	OWLAD	LAD	OLS	NRLAD	JSLAD	OWLAD	LAD
	No Restriction imposed on TBP weight										
Beta	Line	0.049	0.344	0.845	0.336	0.845	-7.730 +	-5.074 +	-0.579 +	-5.148+	-0.579 +
	Parabola	0.723	0.974	1.176	0.974	1.174	-1.673 +	0.576 +	2.398 +	0.579 +	2.380 +
	Kinked	1.531	1.600	1.662	1.599	1.661	5.586 +	6.205 +	6.758 +	6.196	6.751 +
Vasicek	Line	-0.264	-0.119	0.995	-0.162	0.995	-10.540	-9.238	0.773 +	-9.625	0.771 +
	Parabola	0.462	0.846	0.963	0.846	0.962	-4.016 +	-0.572 +	0.484 +	-0.571 +	0.471 +
	Kinked	1.238	1.456	1.573	1.451	1.572	2.948 +	4.910 +	5.958 +	4.868 +	5.948 +
JS	Line	-0.150	-0.396	1.027	-0.438	1.027	-9.512	-11.720	1.057 +	-12.099 +	1.058 +
	Parabola	0.297	0.797	0.949	0.798	0.948	-5.497 +	-1.008 +	0.358 +	-0.998 +	0.345 +
	Kinked	1.203	1.412	1.517	1.409	1.516	2.637 +	4.510 +	5.456 +	4.487 +	5.447 +
	[0, 1] Restriction imposed on TBP weight										
Beta	Line	0.765	0.744	1.019	0.747	1.019	-1.293	-1.482	0.980	-1.460	0.984
	Parabola	0.942	0.929	1.063	0.931	1.063	0.294 +	0.180	1.378 +	0.197	1.377 +
	Kinked	1.205	1.232	1.435	1.232	1.435	2.659 +	2.897 +	4.717 +	2.898 +	4.718 +
Vasicek	Line	0.792	0.867	1.077	0.865	1.077	-1.054	-0.378	1.504 +	-0.399 +	1.507 +
	Parabola	0.979	1.033	0.994	1.033	0.994	0.626 +	1.108 +	0.760 +	1.112 +	0.757 +
	Kinked	1.210	1.244	1.261	1.242	1.261	2.703 +	3.003 +	3.158 +	2.985 +	3.157 +
JS	Line	0.789	0.866	1.015	0.864	1.015	-1.082	-0.388	0.948 +	-0.407	0.951 +
	Parabola	0.980	1.027	0.987	1.028	0.987	0.636 +	1.057 +	0.696 +	1.065 +	0.693 +
	Kinked	1.214	1.248	1.267	1.247	1.267	2.732 +	3.042 +	3.216 +	3.030 +	3.214 +

Note: + indicates the M^2-measure is greater than the M^2-measure in the Diagonal Model.

Table 3.8.4 TBP Return and Risk: Covariance Model (S&P500 Return = 8.166: S&P500 Risk = 8.980)

		Sharpe Ratio					M^2-measure				
		OLS	NRLAD	JSLAD	OWLAD	LAD	OLS	NRLAD	JSLAD	OWLAD	LAD
	No Restriction imposed on TBP weight										
Beta	Line	4.034	15.868	30.256	15.639	30.245	83.054	46.080	35.798	46.533	35.800
	Parabola	36.106	37.503	35.948	37.478	35.882	49.933	38.522	30.559	38.487	30.553
	Kinked	68.657	62.694	51.146	62.875	51.098	44.833	39.176	30.775	39.315	30.761
Vasicek	Line	-21.434	-8.644	92.562	-12.346	92.746	81.075	72.399	92.987	75.986	93.190
	Parabola	20.664	26.476	21.943	26.411	21.895	44.744	31.310	22.782	31.225	22.765
	Kinked	52.436	45.674	35.696	45.674	35.657	42.368	31.443	22.696	31.468	22.687
JS	Line	-9.729	-38.617	40.545	-46.863	40.558	64.901	97.588	39.475	106.987	39.485
	Parabola	13.608	24.074	20.648	24.032	20.602	45.783	30.201	21.752	30.107	21.737
	Kinked	48.660	41.968	32.493	41.926	32.463	40.449	29.732	21.419	29.755	21.414
	[0, 1] Restriction imposed on TBP weight										
Beta	Line	7.958	8.170	11.972	8.162	11.983	10.398	10.976	11.754	10.930	11.760
	Parabola	9.532	9.535	10.591	9.551	10.590	10.118	10.259	9.965	10.256	9.965
	Kinked	11.532	11.632	13.827	11.638	13.829	9.567	9.442	9.638	9.445	9.639
Vasicek	Line	8.163	8.818	12.456	8.794	12.464	10.307	10.168	11.568	10.167	11.571
	Parabola	9.809	9.923	9.892	9.923	9.888	10.019	9.608	9.952	9.604	9.951
	Kinked	11.526	11.390	10.978	11.378	10.976	9.523	9.158	8.706	9.162	8.705
JS	Line	8.137	8.811	11.054	8.788	11.057	10.315	10.173	10.891	10.171	10.892
	Parabola	9.821	9.867	9.843	9.870	9.839	10.020	9.607	9.974	9.602	9.973
	Kinked	11.550	11.365	10.927	11.358	10.925	9.517	9.106	8.621	9.110	8.621

Table 3.8.5 M^2-measure for Unmanaged Beta Estimates

		OLS	NRLAD	JSLAD	OWLAD	LAD
Diagonal Model						
No Restriction						
Beta Estimates	Line	-8.681	-8.169	-5.891	-8.199	-5.888
	Parabola	-4.090	-3.228	-4.433	-3.232	-4.456
	Kinked Line	1.004	0.522	-1.853	0.542	-1.863
[0,1] Restriction						
Beta Estimates	Line	-0.027	0.029	0.164	0.027	0.165
	Parabola	0.356	0.398	0.266	0.397	0.267
	Kinked Line	1.497	1.440	1.159	1.441	1.159
Covariance Model						
No Restriction						
Beta Estimates	Line	-8.481	-6.945	-3.869	-6.997	-3.867
	Parabola	-2.087	-1.160	-1.961	-1.165	-1.977
	Kinked Line	3.460	3.768	2.192	3.785	2.184
[0,1] Restriction						
Beta Estimates	Line	-0.742	-0.293	0.282	-0.305	0.283
	Parabola	0.725	1.029	1.034	1.026	1.032
	Kinked Line	2.588	2.586	2.225	2.585	2.225

Table 3.8.6 Ridge Estimation

	Ridge-OLS	Ridge-NRLAD	Ridge-JSLAD	Ridge-OWLAD	Ridge-LAD
Statistics					
Sharpe Ratio	1.183	1.296	1.340	1.291	1.339
M^2-measure	2.457	3.475	3.864	3.430	3.861
Return	27.600	25.535	21.035	25.461	21.041
Std. Dev.	23.332	19.698	15.702	19.718	15.710
Final Wealth	2.106	2.028	1.810	2.023	1.810

Figure 3.4.1 Histogram of Alpha Forecasts

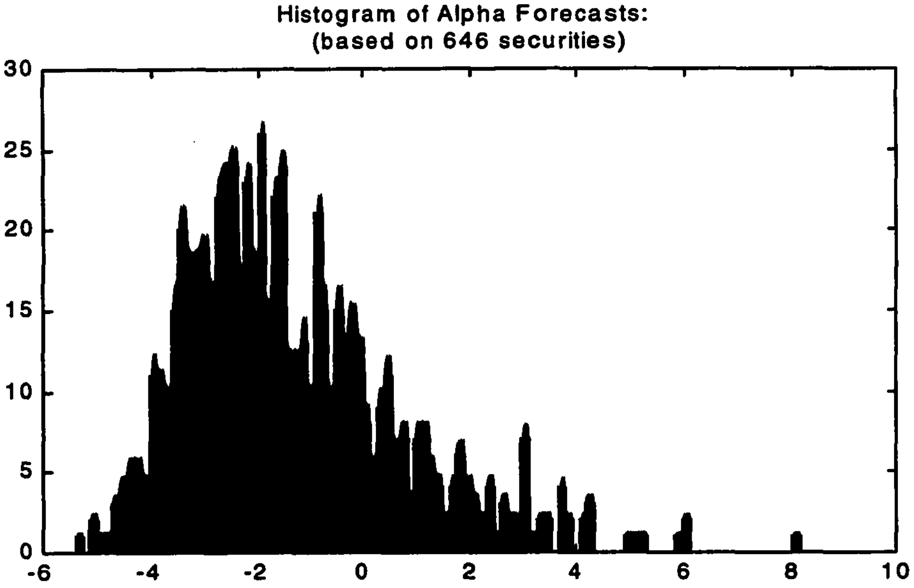


Figure 3.4.2 Histogram of Beta Forecasts

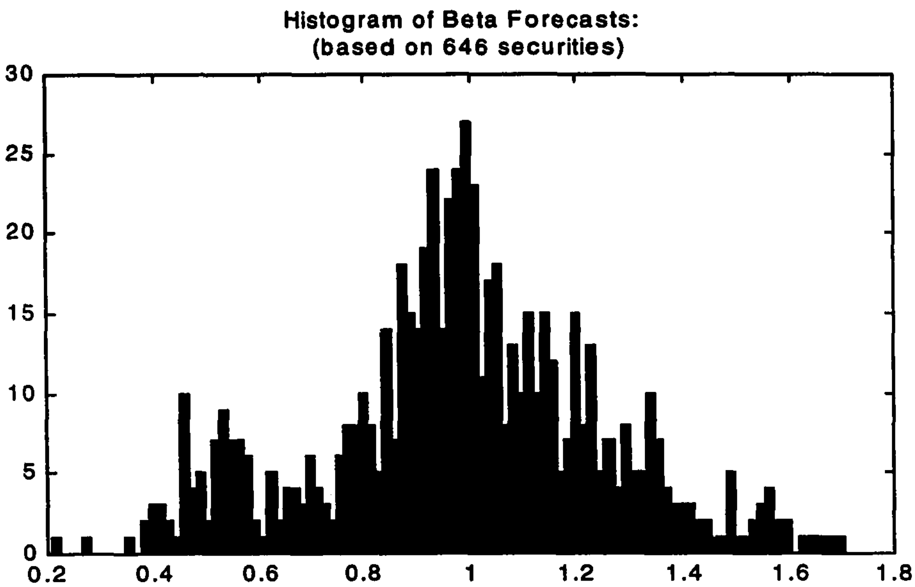


Figure 3.5.1 Market Value: Population vs. Sample

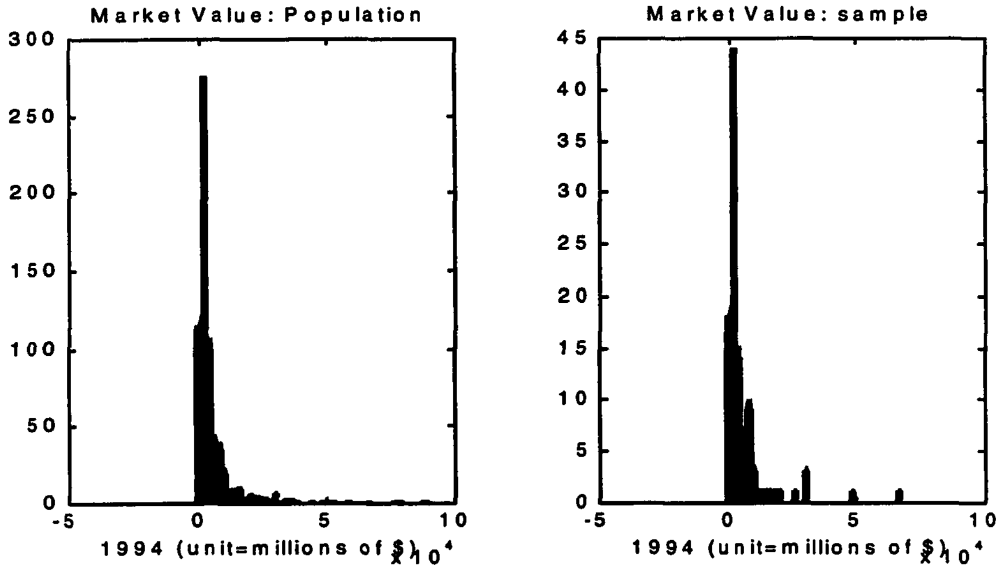


Figure 3.5.2 Book/Market Value: Population vs. Sample

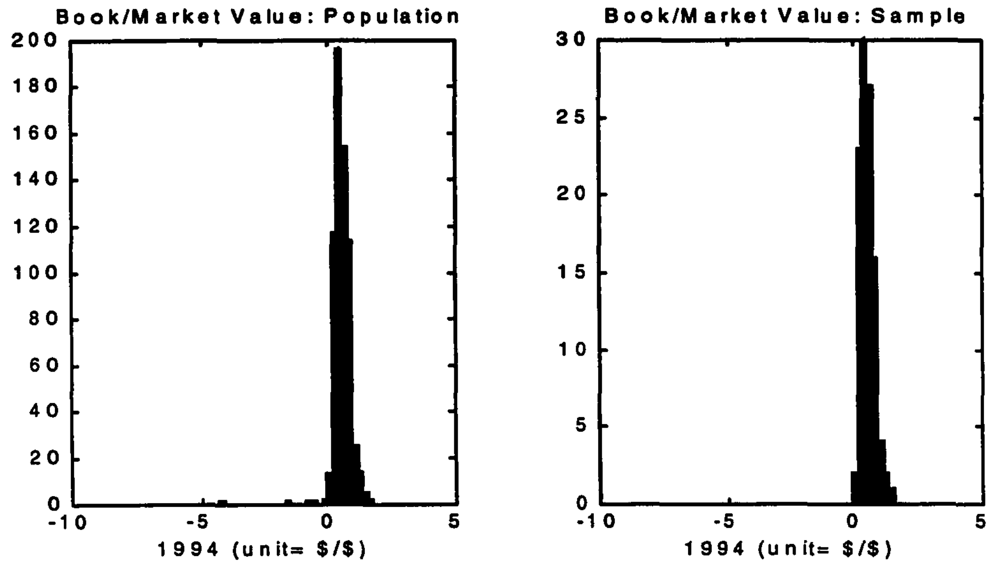


Figure 3.6.1 Distribution of Beta Forecasts

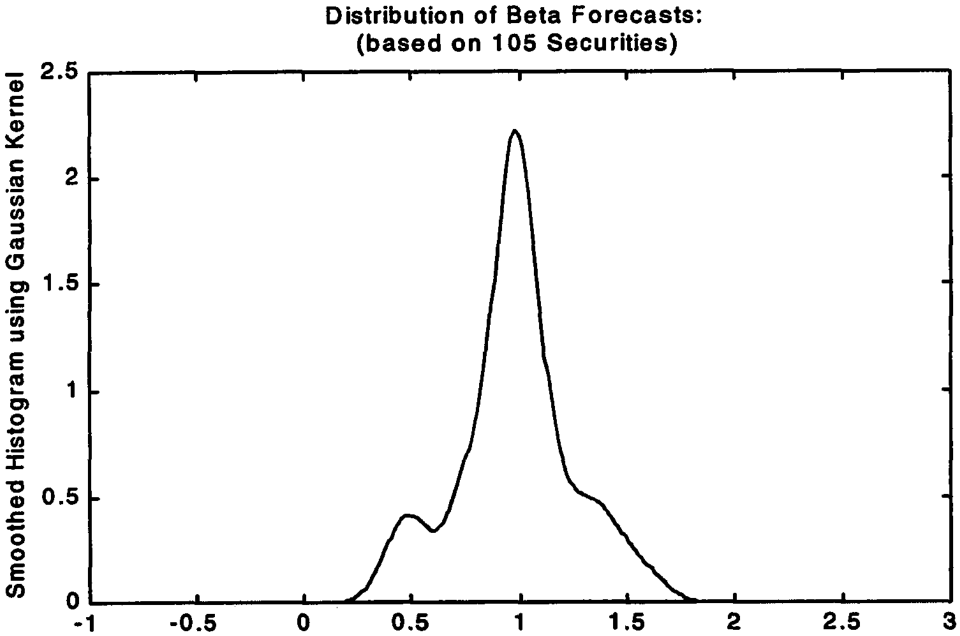


Figure 3.6.2 Distribution of Beta Estimates Based on 4 Methods.

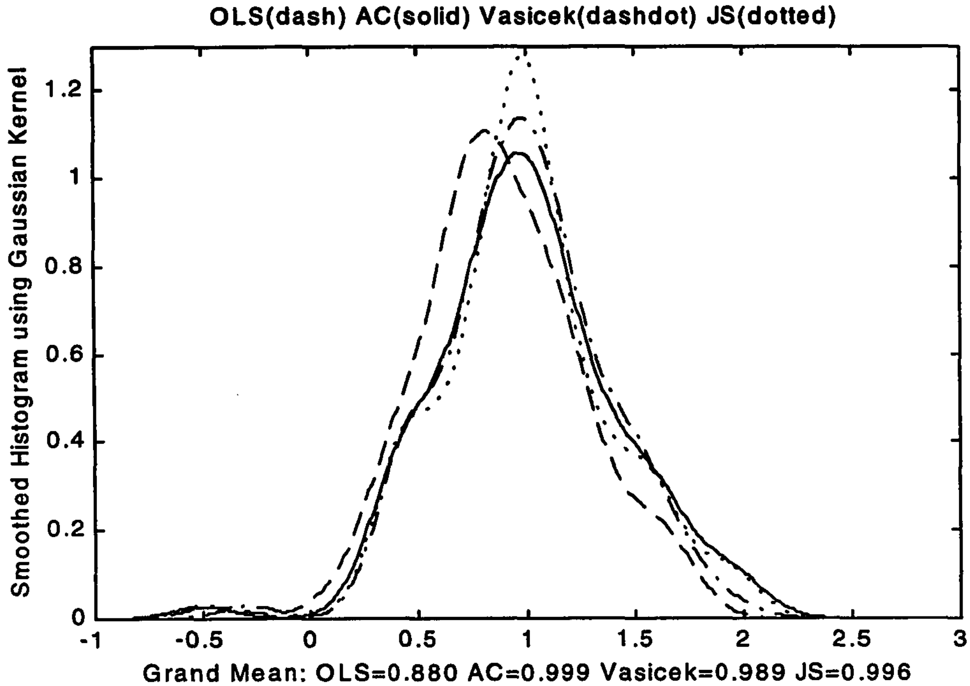


Figure 3.7.1 Scatter Diagram of Ex-Post Abnormal Returns vs. Alpha Forecasts

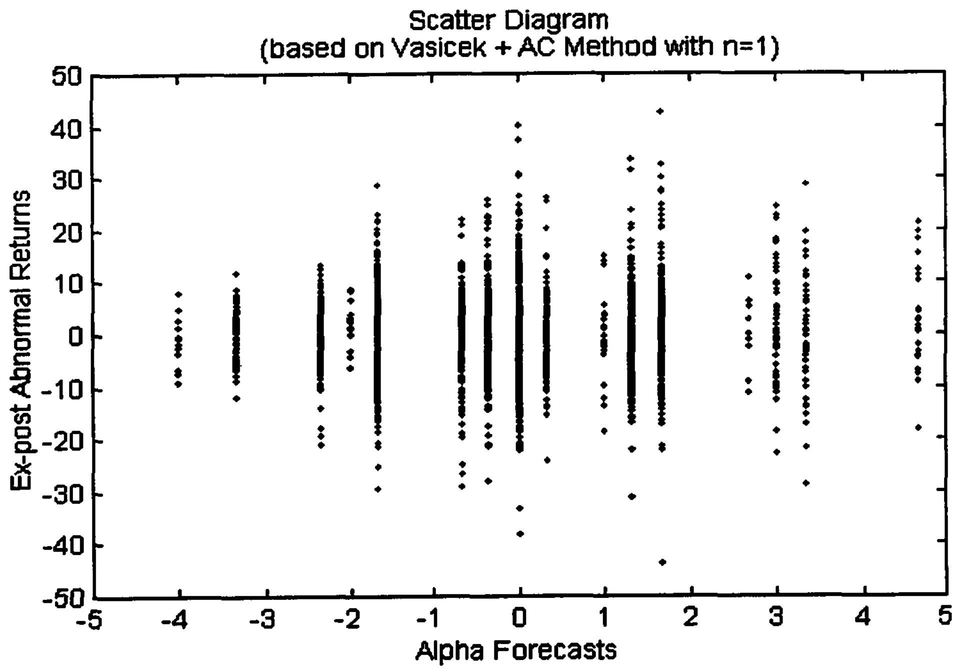


Figure 3.7.2 Fitted Lines Based on 3 Specifications

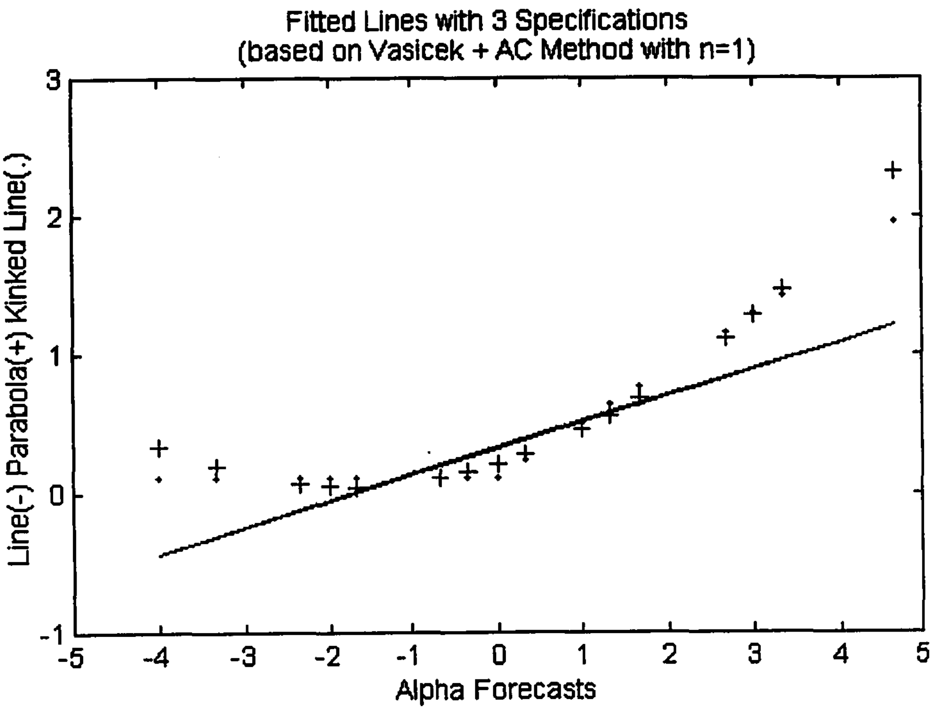


Figure 3.7.3 Predictive Power (adjusted R^2)

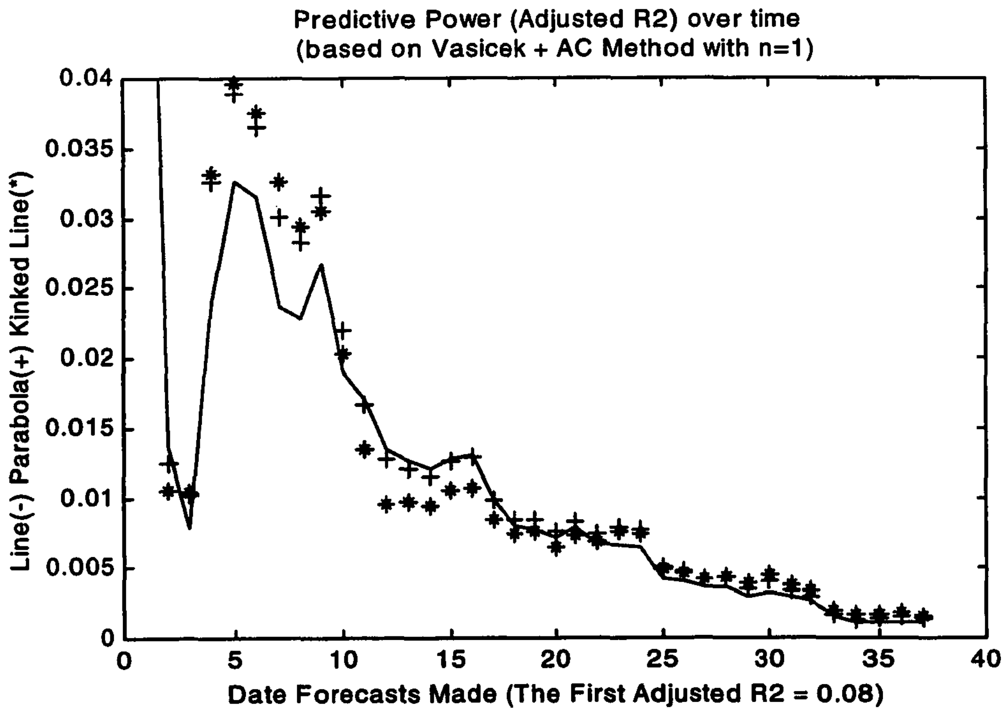


Figure 3.8.1 Forecasting Monthly S&P500 Index

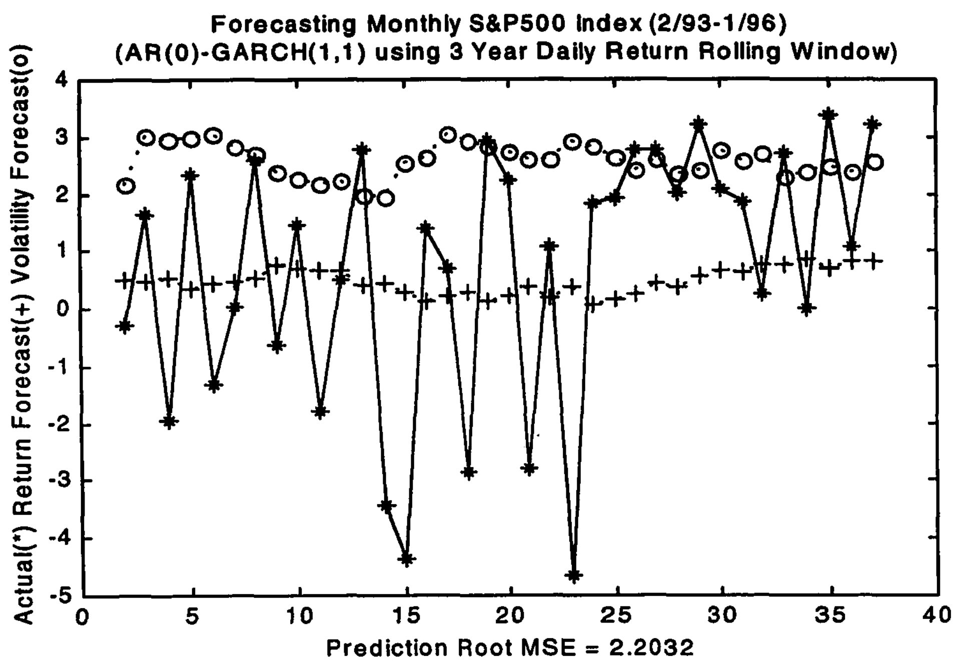


Figure 3.8.2 Trynor-Black Portfolio Weight

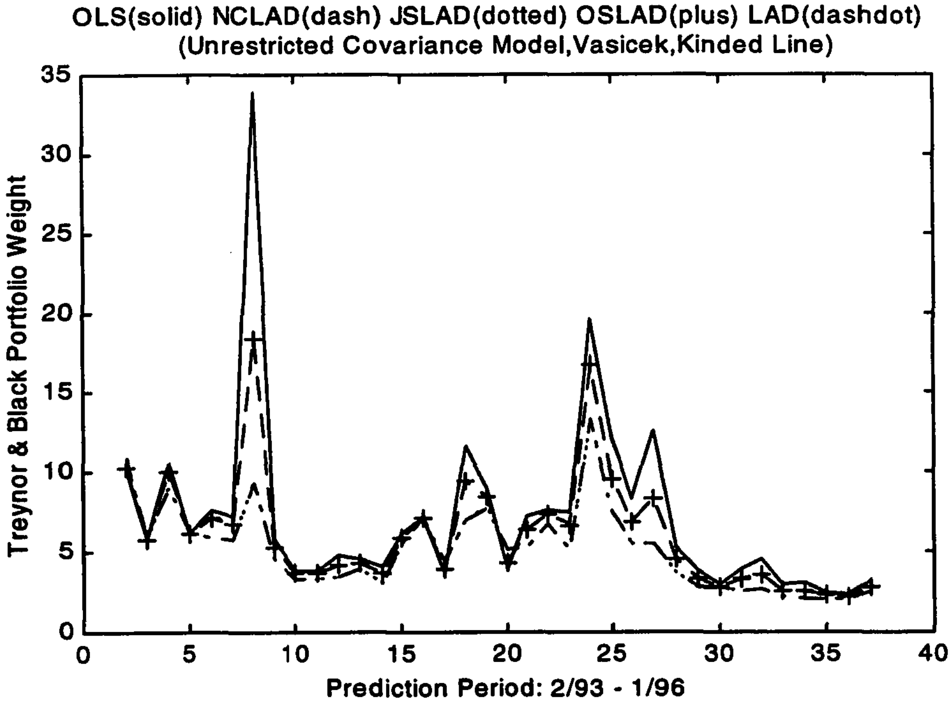


Figure 3.8.3 Forecasting Residual Return

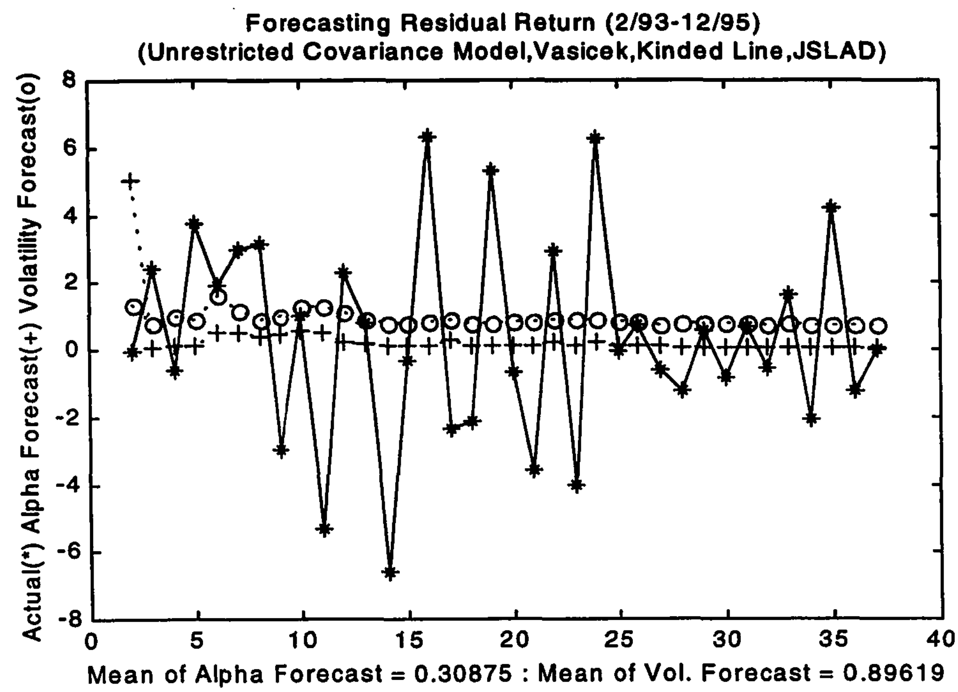


Figure 3.8.4 Forecasting the Beta of the Active Portfolio

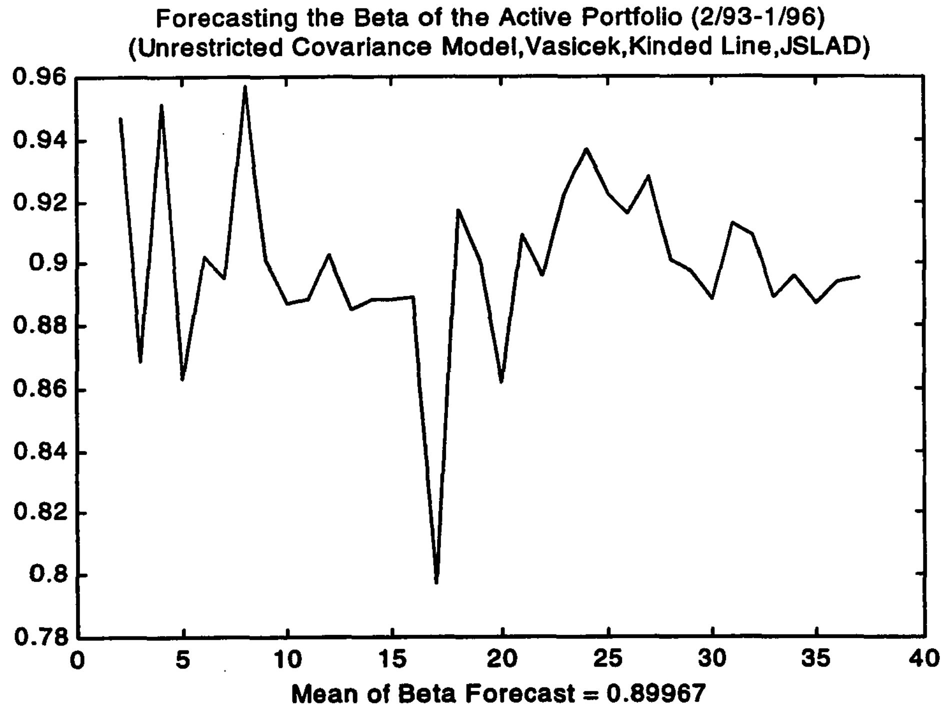


Figure 3.8.5 Estimate of Slope Coefficient

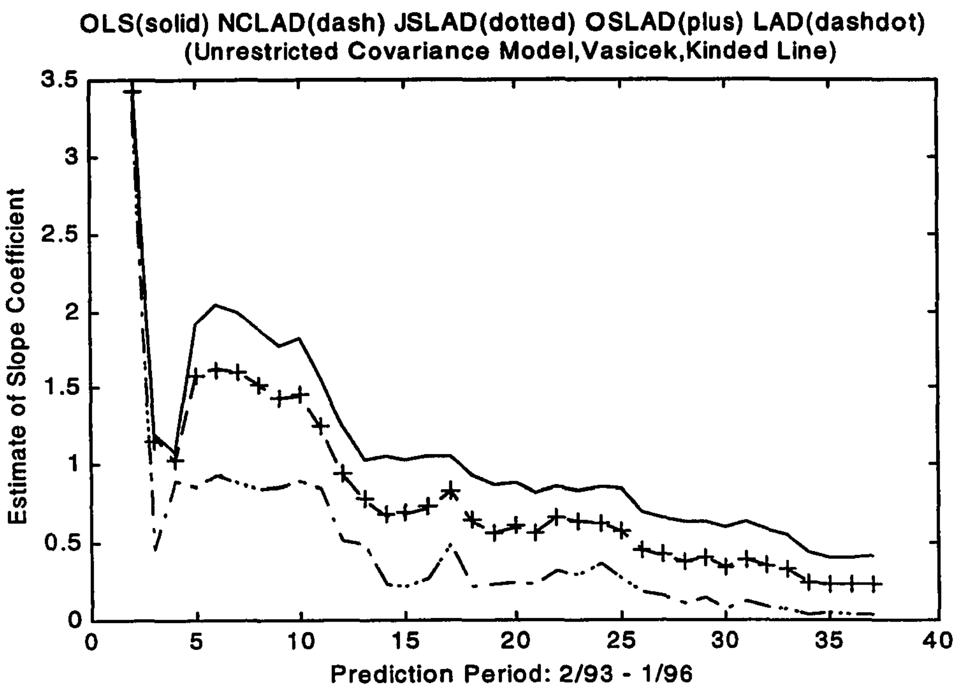


Figure 3.8.6 Treynor-Black Portfolio Weight (Stabilized)

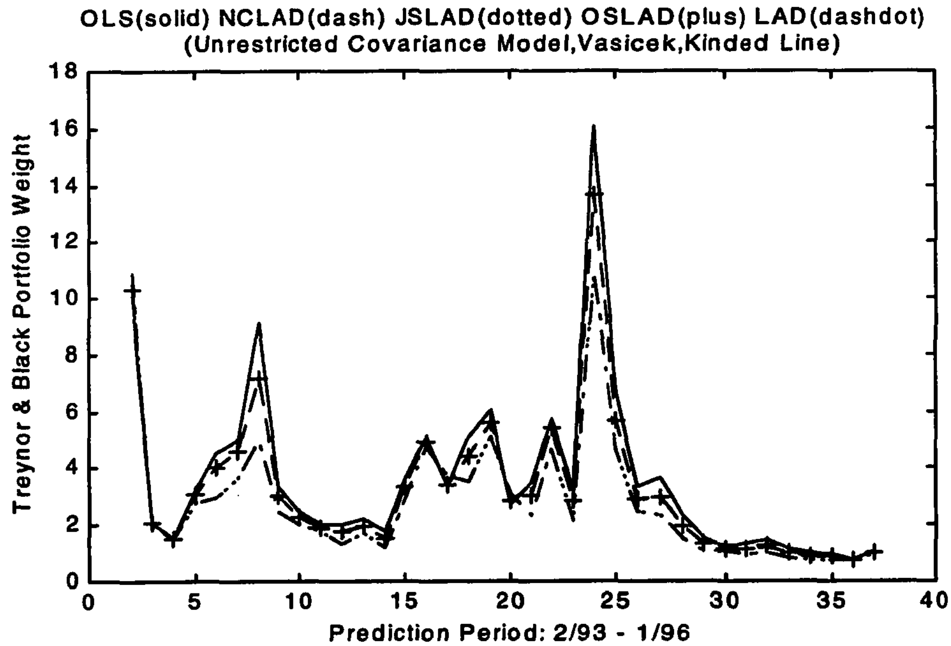


Figure 3.8.7 Wealth Index

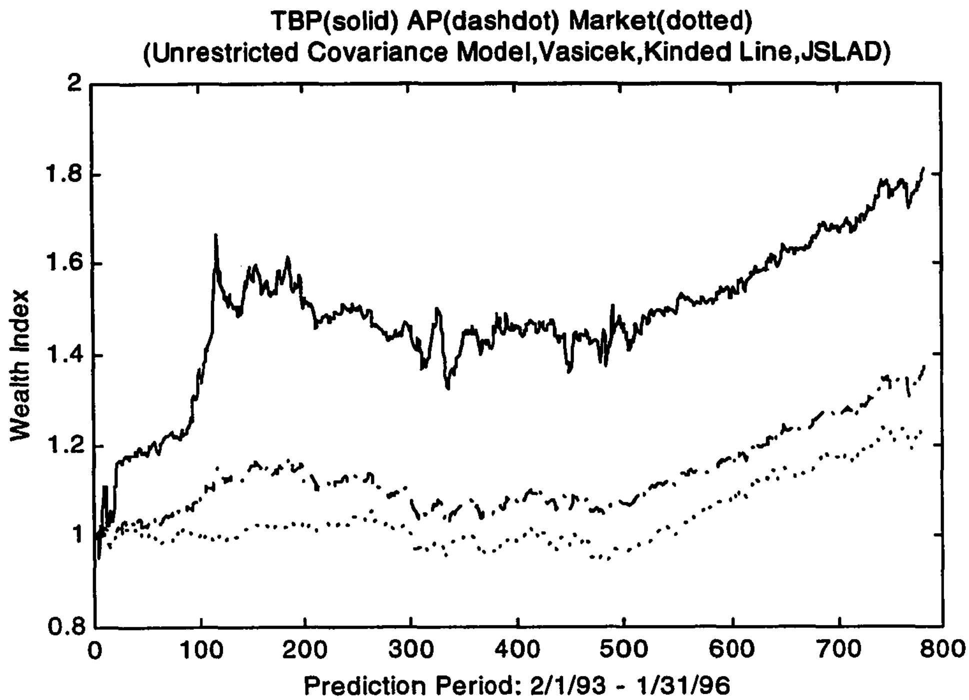
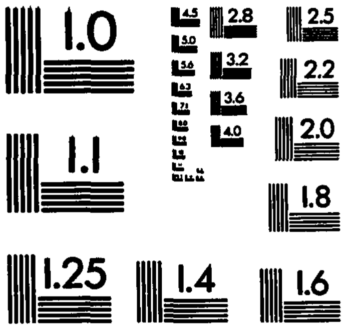


IMAGE EVALUATION TEST TARGET (QA-3)



150mm

6"

APPLIED IMAGE, Inc

1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989