MEAN-VARIANCE PORTFOLIO OPTIMIZATION
WHEN MEANS AND COVARIANCES ARE UNKNOWN

By

Tze Leung Lai
Haipeng Xing
Zehao Chen

Technical Report No. 2009-8
May 2009

Department of Statistics
STANFORD UNIVERSITY
Stanford, California 94305-4065
MEAN-VARIANCE PORTFOLIO OPTIMIZATION
WHEN MEANS AND COVARIANCES ARE UNKNOWN

By

Tze Leung Lai
Department of Statistics
Stanford University

Haipeng Xing
Department of Applied Mathematics and Statistics
State University of New York, Stony Brook

Zehao Chen
Bosera Asset Management, P.R. China

Technical Report No. 2009-8
May 2009

This research was supported in part by
National Science Foundation grant DMS 0805879.

Department of Statistics
STANFORD UNIVERSITY
Stanford, California 94305-4065

http://statistics.stanford.edu
Markowitz’s celebrated mean-variance portfolio optimization theory assumes that the means and covariances of the underlying asset returns are known. In practice, they are unknown and have to be estimated from historical data. Plugging the estimates into the efficient frontier that assumes known parameters has led to portfolios that may perform poorly and have counter-intuitive asset allocation weights; this has been referred to as the “Markowitz optimization enigma.” After reviewing different approaches in the literature to address these difficulties, we explain the root cause of the enigma and propose a new approach to resolve it. Not only is the new approach shown to provide substantial improvements over previous methods, but it also allows flexible modeling to incorporate dynamic features and fundamental analysis of the training sample of historical data, as illustrated in simulation and empirical studies.

*Key words and phrases:* Markowitz’s portfolio theory, efficient frontier, empirical Bayes, stochastic optimization.

1. Research supported by the National Science Foundation
1. Introduction. The mean-variance (MV) portfolio optimization theory of Harry Markowitz (1952, 1959), Nobel laureate in economics, is widely regarded as one of the foundational theories in financial economics. It is a single-period theory on the choice of portfolio weights that provide the optimal tradeoff between the mean (as a measure of profit) and the variance (as a measure of risk) of the portfolio return for a future period. For a portfolio consisting of \( m \) assets (e.g., stocks) with expected returns \( \mu_i \), let \( w_i \) be the weight of the portfolio's value invested in asset \( i \) such that \( \sum_{i=1}^{m} w_i = 1 \), and let \( w = (w_1, \ldots, w_m)^T \), \( \mu = (\mu_1, \ldots, \mu_m) \), \( 1 = (1, \ldots, 1)^T \). The portfolio return has mean \( w^T \mu \) and variance \( w^T \Sigma w \), where \( \Sigma \) is the covariance matrix of the asset returns; see Lai and Xing (2008, pp. 67, 69-71). Given a target value \( \mu^* \) for the mean return of a portfolio, Markowitz characterizes an efficient portfolio by its weight vector \( w_{\text{eff}} \) that solves the optimization problem

\[
(1.1) \quad w_{\text{eff}} = \arg \min_{w} w^T \Sigma w \quad \text{subject to} \quad w^T \mu = \mu^*, \, w^T 1 = 1, \, w \geq 0.
\]

When short selling is allowed, the constraint \( w \geq 0 \) (i.e., \( w_i \geq 0 \) for all \( i \)) in (1.1) can be removed, yielding the following problem that has an explicit solution:

\[
(1.2) \quad w_{\text{eff}} = \arg \min_{w,w^T \mu = \mu^*, w^T 1 = 1} w^T \Sigma w \\
= \left\{ B \Sigma^{-1} - A \Sigma^{-1} \mu + \mu^*(C \Sigma^{-1} \mu - A \Sigma^{-1} 1) \right\} / D,
\]

where \( A = \mu^T \Sigma^{-1} 1 = 1^T \Sigma^{-1} \mu, \, B = \mu^T \Sigma^{-1} \mu, \, C = 1^T \Sigma^{-1} 1, \, \text{and} \, D = BC - A^2. \)

Markowitz’s theory assumes known \( \mu \) and \( \Sigma \). Since in practice \( \mu \) and \( \Sigma \) are unknown, a commonly used approach is to estimate \( \mu \) and \( \Sigma \) from historical data, under the assumption that returns are i.i.d. A standard model for the price \( P_{it} \) of the \( i \)th asset at time \( t \) in finance theory is geometric Brownian motion \( dP_{it} / P_{it} = \theta_i dt + \sigma_i dB_{i}^{(i)} \), where \( \{B_{i}^{(i)}, t \geq 0\} \) is standard Brownian motion. The discrete-time analog of this price process has returns \( r_{it} = (P_{it} - P_{i,t-1}) / P_{i,t-1} \), and log returns \( \log(P_{it} / P_{i,t-1}) = \log(1 + r_{it}) \approx r_{it} \) that are i.i.d. \( N(\theta_i - \sigma_i^2/2, \sigma_i^2) \). Under the standard model, maximum likelihood estimates of \( \mu \) and \( \Sigma \) are the sample mean \( \hat{\mu} \) and the sample covariance matrix \( \hat{\Sigma} \), which are also method-of-moments estimates without the assumption of normality and when the i.i.d. assumption is replaced by weak stationarity (i.e., time-invariant means and covariances). It has been found, however, that replacing \( \mu \) and \( \Sigma \) in (1.1) or (1.2) by their sample counterparts \( \hat{\mu} \) and \( \hat{\Sigma} \) may perform poorly and a major direction in the literature is to find other (e.g., Bayes and shrinkage) estimators that yield better portfolios when they are plugged into (1.1) or (1.2). An alternative method, introduced by Michaud (1989) to tackle the “Markowitz optimization
enigma,” is to adjust the plug-in portfolio weights by incorporating sampling variability of $(\hat{\mu}, \hat{\Sigma})$ via the bootstrap. Section 2 gives a brief survey of these approaches.

Let $\mathbf{r}_t = (r_{1t}, \ldots, r_{mt})^T$. Since Markowitz’s theory deals with portfolio returns in a future period, it is more appropriate to use the conditional mean and covariance matrix of the future returns $\mathbf{r}_{n+1}$ given the historical data $\mathbf{r}_n, \mathbf{r}_{n-1}, \ldots$ based on a Bayesian model that forecasts the future from the available data, rather than restricting to an i.i.d. model that relates the future to the past via the unknown parameters $\mu$ and $\Sigma$ for future returns to be estimated from past data. More importantly, this Bayesian formulation paves the way for a new approach that generalizes Markowitz’s portfolio theory to the case where the means and covariances are unknown. When $\mu$ and $\Sigma$ are estimated from data, their uncertainties should be incorporated into the risk; moreover, it is not possible to attain a target level of mean return as in Markowitz’s constraint $\mathbf{w}^T \mu = \mu_*$ since $\mu$ is unknown. To address this root cause of the Markowitz enigma, we introduce in Section 3 a Bayesian approach that assumes a prior distribution for $(\mu, \Sigma)$ and formulates mean-variance portfolio optimization as a stochastic optimization problem. This optimization problem reduces to that of Markowitz when the prior distribution is degenerate. It uses the posterior distribution given current and past observations to incorporate the uncertainties of $\mu$ and $\Sigma$ into the variance of the portfolio return $\mathbf{w}^T \mathbf{r}_{n+1}$, where $\mathbf{w}$ is based on the posterior distribution. The constraint in Markowitz’s mean-variance formulation can be included in the objective function by using a Lagrange multiplier $\lambda^{-1}$ so that the optimization problem is to evaluate the weight vector $\mathbf{w}$ that maximizes $E(\mathbf{w}^T \mathbf{r}_{n+1}) - \lambda \operatorname{Var}(\mathbf{w}^T \mathbf{r}_{n+1})$, for which $\lambda$ can be regarded as a risk aversion coefficient. To compare with previous frequentist approaches that assume i.i.d. returns, Section 4 introduces a variant of the Bayes rule that uses bootstrap resampling to estimate the performance criterion nonparametrically.

To apply this theory in practice, the investor has to figure out his/her risk aversion coefficient, which may be a difficult task. Markowitz’s theory circumvents this by considering the efficient frontier, which is the $(\sigma, \mu)$ curve of efficient portfolios as $\lambda$ varies over all possible values, where $\mu$ is the mean and $\sigma^2$ the variance of the portfolio return. Investors, however, often prefer to use the Sharpe ratio $(\mu - \mu_0)/\sigma$ as a measure of a portfolio’s performance, where $\mu_0$ is the risk-free interest rate or the expected return of a market portfolio (e.g., S&P500). Note that the Sharpe ratio is proportional to $\mu - \mu_0$ and inversely proportional to $\sigma$, and can be regarded as the excess return per unit of risk. In Section 5 we describe how $\lambda$ can be chosen for the rule developed in Section 3 to maximize the Sharpe ratio. Other statistical issues that arise in practice are also considered in Section 5 where they lead to certain
modifications of the basic rule. Among them are estimation of high-dimensional covariance matrices when $m$ (number of assets) is not small relative to $n$ (number of past periods in the training sample) and departures of the historical data from the working assumption of i.i.d. asset returns. Section 6 illustrates these methods in an empirical study in which the rule thus obtained is compared with other rules proposed in the literature. Some concluding remarks are given in Section 7.

2. Using better estimates of $\mu$, $\Sigma$ or $w_{\text{eff}}$ to implement Markowitz’s portfolio optimization theory. Since $\mu$ and $\Sigma$ in Markowitz’s efficient frontier are actually unknown, a natural idea is to replace them by the sample mean vector $\hat{\mu}$ and covariance matrix $\hat{\Sigma}$ of the training sample. However, this plug-in frontier is no longer optimal because $\hat{\mu}$ and $\hat{\Sigma}$ actually differ from $\mu$ and $\Sigma$, but also Frankfurter, Phillips and Seagle (1976) and Jobson and Korkie (1980) have reported that portfolios on the plug-in frontier can perform worse than an equally weighted portfolio that is highly inefficient. Michaud (1989) comments that the minimum variance (MV) portfolio $w_{\text{eff}}$ based on $\hat{\mu}$ and $\hat{\Sigma}$ has serious deficiencies, calling the MV optimizers “estimation-error maximizers”. His argument is reinforced by subsequent studies, e.g., Best and Grauer (1991), Chopra, Hensel and Turner (1993), Canner et al. (1997), Simann (1997), and Britten-Jones (1999). Three approaches have been proposed to address the difficulty during the past two decades. The first approach uses multifactor models to reduce the dimension in estimating $\Sigma$, and the second approach uses Bayes or other shrinkage estimates of $\Sigma$. Both approaches use improved estimates of $\Sigma$ for the plug-in efficient frontier, and can also be modified to provide better estimates of $\mu$ to be used as well. The third approach uses bootstrapping to correct for the bias of $\hat{w}_{\text{eff}}$ as an estimate of $w_{\text{eff}}$.

2.1. Multifactor pricing models. Multifactor pricing models relate the $m$ asset returns $r_i$ to $k$ factors $f_1, \ldots, f_k$ in a regression model of the form

$$r_i = \alpha_i + (f_1, \ldots, f_k)^T \beta_i + \epsilon_i,$$

in which $\alpha_i$ and $\beta_i$ are unknown regression parameters and $\epsilon_i$ is an unobserved random disturbance that has mean 0 and is uncorrelated with $f := (f_1, \ldots, f_k)^T$. The case $k = 1$ is called a single-factor (or single-index) model. Under Sharpe’s (1964) capital asset pricing model (CAPM) which assumes, besides known $\mu$ and $\Sigma$, that the market has a risk-free asset with return $r_f$ (interest rate) and that all investors minimize the variance of their portfolios for their target mean returns, (2.1) holds with $k = 1$, $\alpha_i = r_f$ and $f = r_M - r_f$, where $r_M$ is the return of a hypothetical market portfolio $M$ which can be approximated in practice.
by an index fund such as Standard and Poor’s (S&P) 500 Index. The arbitrage pricing theory (APT), introduced by Ross (1976), involves neither a market portfolio nor a risk-free asset and states that a multifactor model of the form (2.1) should hold approximately in the absence of arbitrage for sufficiently large $m$. The theory, however, does not specify the factors and their number. Methods for choosing factors in (2.1) can be broadly classified as economic and statistical, and commonly used statistical methods include factor analysis and principal component analysis; see Section 3.4 of Lai and Xing (2008).

2.2. Bayes and shrinkage estimators. A popular conjugate family of prior distributions for estimation of covariance matrices from i.i.d. normal random vectors $r_t$ with mean $\mu$ and covariance matrix $\Sigma$ is

\begin{align}
\mu | \Sigma & \sim N(\nu, \Sigma/k), \quad \Sigma \sim IW_m(\Psi, n_0), \\
\end{align}

where $IW_m(\Psi, n_0)$ denotes the inverted Wishart distribution with $n_0$ degrees of freedom and mean $\Psi/(n_0 - m - 1)$. The posterior distribution of $(\mu, \Sigma)$ given $(r_1, \ldots, r_n)$ is also of the same form:

\begin{align}
\mu | \Sigma & \sim N(\hat{\mu}, \Sigma/(n + \kappa)), \quad \Sigma \sim IW_m((n + n_0 - m - 1)\hat{\Sigma}, n + n_0), \\
\end{align}

where $\hat{\mu}$ and $\hat{\Sigma}$ are the Bayes estimators of $\mu$ and $\Sigma$ given by

\begin{align}
\hat{\mu} &= \frac{k}{n + \kappa} \nu + \frac{n}{n + \kappa} \bar{r}, \\
\hat{\Sigma} &= \frac{n_0 - m - 1}{n + n_0 - m - 1} \frac{\Psi}{n_0 - m - 1} + \frac{n}{n + n_0 - m - 1} \left\{ \frac{1}{n} \sum_{i=1}^{n} (r_t - \bar{r})(r_t - \bar{r})^T \\
&\quad + \frac{k}{n + \kappa} (\bar{r} - \nu)(\bar{r} - \nu)^T \right\}. \\
\end{align}

Note that the Bayes estimator $\hat{\Sigma}$ adds to the MLE of $\Sigma$ the covariance matrix $k(\bar{r} - \nu)(\bar{r} - \nu)^T/(n + \kappa)$, which accounts for the uncertainties due to replacing $\mu$ by $\bar{r}$, besides shrinking this adjusted covariance matrix towards the prior mean $\Psi/(n_0 - m - 1)$.

Instead of using directly this Bayes estimator which requires specification of the hyper-parameters $\mu$, $\kappa$, $n_0$ and $\Psi$, Ledoit and Wolf (2003, 2004) propose to estimate $\mu$ simply by $\bar{r}$ and to shrink the MLE of $\Sigma$ towards a structured covariance matrix. Their rationale is that whereas the MLE $S = \sum_{t=1}^{n} (r_t - \bar{r})(r_t - \bar{r})^T/n$ has a large estimation error when $m(m+1)/2$ is comparable with $n$, a structured covariance matrix $F$ has much fewer parameters that can be estimated with smaller variances. They propose to estimate $\Sigma$ by a convex combination of $\hat{F}$ and $S$:

\begin{align}
\hat{\Sigma} = \delta \hat{F} + (1 - \delta)S, \\
\end{align}

4
where \( \hat{\delta} \) is an estimator of the optimal shrinkage constant \( \delta \) used to shrink the MLE toward the estimated structured covariance matrix \( \hat{F} \). Besides the covariance matrix \( F \) associated with a single-factor model, they also suggest using a constant correlation model for \( F \) in which all pairwise correlations are identical, and have found that it gives comparable performance in simulation and empirical studies. They advocate using this shrinkage estimate in lieu of \( S \) in implementing Markowitz’s efficient frontier.

2.3. Bootstrapping and the resampled frontier. To adjust for the bias of \( \hat{w}_{\text{eff}} \) as an estimate of \( w_{\text{eff}} \), Michaud (1989) uses the average of the bootstrap weight vectors:

\[
\bar{w} = B^{-1} \sum_{b=1}^{B} \hat{w}^*_{b},
\]

where \( \hat{w}^*_{b} \) is the estimated optimal portfolio weight vector based on the \( b \)th bootstrap sample \( \{r^*_b, \ldots, r^*_{bn}\} \) drawn with replacement from the observed sample \( \{r_1, \ldots, r_n\} \). Specifically, the \( b \)th bootstrap sample has sample mean vector \( \hat{\mu}^*_b \) and covariance matrix \( \hat{\Sigma}^*_b \), which can be used to replace \( \mu \) and \( \Sigma \) in (1.1) or (1.2), thereby yielding \( \hat{w}^*_b \). Thus, the resampled efficient frontier corresponds to plotting \( \bar{w}^T \hat{\mu} \) versus \( \sqrt{\bar{w}^T \hat{\Sigma} \bar{w}} \) for a fine grid of \( \mu^* \) values, where \( \bar{w} \) is defined by (2.5) in which \( \hat{w}^*_b \) depends on the target level \( \mu^* \).

3. A stochastic optimization approach. The Bayesian and shrinkage methods in Section 2.2 focus primarily on Bayes estimates of \( \mu \) and \( \Sigma \) (with normal and inverted Wishart priors) and shrinkage estimators of \( \Sigma \). However, the construction of efficient portfolios when \( \mu \) and \( \Sigma \) are unknown is more complicated than trying to estimate them as well as possible and then plugging the estimates into (1.1) or (1.2). Note in this connection that (1.2) involves \( \Sigma^{-1} \) instead of \( \Sigma \) and that estimating \( \Sigma \) as well as possible does not imply that \( \Sigma^{-1} \) is reliably estimated. Estimation of a high-dimensional \( m \times m \) covariance matrix when \( m^2 \) is not small compared to \( n \) has been recognized as a difficult statistical problem and attracted much recent attention; see Bickel and Lavina (2008). Some sparsity condition is needed to obtain an estimate that is close to \( \Sigma \) in Frobenius norm, and the conjugate prior family (2.2) that motivates the (linear) shrinkage estimators (2.3) or (2.4) does not reflect such sparsity. For high-dimensional weight vectors \( \hat{w}_{\text{eff}} \), direct application of the bootstrap for bias correction is also problematic.

A major difficulty with the “plug-in” efficient frontier (which uses the MLE \( S \) or (2.4) to estimate \( \Sigma \)) and its “resampled” version is that Markowitz’s idea of using the variance of \( w^T r_{n+1} \) as a measure of the portfolio’s risk cannot be captured simply by the plug-in estimate \( w^T \hat{\Sigma} w \) of \( \text{Var}(w^T r_{n+1}) \). Whereas the problem of minimizing \( \text{Var}(w^T r_{n+1}) \) subject
to a given level $\mu_*$ of the mean return $E(w^T r_{n+1})$ is meaningful in Markowitz’s framework, in which both $E(r_{n+1})$ and $\text{Cov}(r_{n+1})$ are known, the surrogate problem of minimizing $w^T \hat{\Sigma} w$ under the constraint $w^T \hat{\mu} = \mu_*$ ignores the fact both $\hat{\mu}$ and $\hat{\Sigma}$ have inherent errors (risks) themselves. In this section we consider the more fundamental problem

\begin{equation}
(3.1) \quad \max \left\{ E(w^T r_{n+1}) - \lambda \text{Var}(w^T r_{n+1}) \right\}
\end{equation}

when $\mu$ and $\Sigma$ are unknown and treated as state variables whose uncertainties are specified by their posterior distributions given the observations $r_1, \ldots, r_n$ in a Bayesian framework. The weights $w$ in (3.1) are random vectors that depend on $r_1, \ldots, r_n$. Note that if the prior distribution puts all its mass at $(\mu_0, \Sigma_0)$, then the minimization problem (3.1) reduces to Markowitz’s portfolio optimization problem that assumes $\mu_0$ and $\Sigma_0$ are given.

### 3.1. Solution of the optimization problem (3.1)

The problem (3.1) is not a standard stochastic control problem and is difficult to solve directly because of the term $\left[ E(w^T r_{n+1}) \right]^2$ in $\text{Var}(w^T r_{n+1}) = E(w^T r_{n+1})^2 - \left[ E(w^T r_{n+1}) \right]^2$. We first convert it to a standard stochastic control problem by using an additional parameter. Let $W = w^T r_{n+1}$ and note that $E(W) - \lambda \text{Var}(W) = h(EW, EW^2)$, where $h(x, y) = x + \lambda_x^2 - \lambda y$. Let $W_B = w_B^T r_{n+1}$ and $\eta = 1 + 2\lambda E(W_B)$, where $w_B$ is the Bayes weight vector. Then

\[
0 \geq h(EW, EW^2) - h(EW_B, EW_B^2)
= E(W) - E(W_B) - \lambda \{ E(W^2) - E(W_B^2) \} + \lambda \{ (EW)^2 - (EW_B)^2 \}
= \eta \{ E(W) - E(W_B) \} + \lambda \{ E(W_B^2) - E(W^2) \} + \lambda \{ E(W) - E(W_B) \}^2
\geq \{ \lambda E(W_B^2) - \eta E(W_B) \} - \{ \lambda E(W^2) - \eta E(W) \},
\]

Therefore

\begin{equation}
(3.2) \quad \lambda E(W^2) - \eta E(W) \geq \lambda E(W_B^2) - \eta E(W_B),
\end{equation}

with equality if and only if $W$ has the same mean and variance as $W_B$. Hence the stochastic optimization problem (3.1) is equivalent to minimizing $\lambda E[(w^T r_{n+1})^2] - \eta E(w^T r_{n+1})$ over weight vectors $w$ that can depend on $r_1, \ldots, r_n$. Since $\eta$ is a linear function of the solution of (3.1), we cannot apply this equivalence directly to the unknown $\eta$. Instead we solve a family of stochastic optimization problems over $\eta$ and then choose the $\eta$ that maximizes the reward in (3.1). Specifically, we rewrite (3.1) as the following optimization problem:

\begin{equation}
(3.3) \quad \max_{\eta} \left\{ E[w^T(\eta)r_{n+1}] - \lambda \text{Var}[w^T(\eta)r_{n+1}] \right\},
\end{equation}


where
\[ w(\eta) = \arg \min_w \left\{ \lambda E[(w^T r_{n+1})^2] - \eta E(w^T r_{n+1}) \right\}. \]

### 3.2 Computation of the optimal weight vector.
Let \( \mu_n \) and \( V_n \) be the posterior mean and second moment matrix given the set \( R_n \) of current and past returns \( r_1, \ldots, r_n \). Since \( w \) is based on \( R_n \), it follows from \( E(r_{n+1}|R_n) = \mu_n \) and \( E(r_{n+1}^T r_{n+1}|R_n) = V_n \) that
\[
E(w^T r_{n+1}) = E(w^T \mu_n), \quad E[(w^T r_{n+1})^2] = E(w^T V_n w).
\]
Without short selling the weight vector \( w(\eta) \) in (3.3) is given by the following analog of (1.1)
\[
w(\eta) = \arg \min_{w:w^T 1 = 1, w \geq 0} \left\{ \lambda w^T V_n w - \eta w^T \mu_n \right\},
\]
which can be computed by quadratic programming (e.g., by quadprog in MATLAB). When short selling is allowed, the constraint \( w \geq 0 \) can be removed and \( w(\eta) \) in (3.3) is given explicitly by
\[
w(\eta) = \arg \min_{w:w^T 1 = 1} \left\{ \lambda w^T V_n w - \eta w^T \mu_n \right\} = \frac{1}{C_n} V_n^{-1} \left( \mu_n - \frac{A_n}{C_n} 1 \right),
\]
where the second equality can be derived by using a Lagrange multiplier and
\[
A_n = \mu_n^T V_n^{-1} 1 = 1^T V_n^{-1} \mu_n, \quad B_n = \mu_n^T V_n^{-1} 1, \quad C_n = 1^T V_n^{-1} 1.
\]
Note that (3.4) or (3.5) essentially plugs the Bayes estimates of \( \mu \) and \( V := \Sigma + \mu \mu^T \) into the optimal weight vector that assumes \( \mu \) and \( \Sigma \) to be known. However, unlike the “plug-in” efficient frontier described in the first paragraph of Section 2, we have first transformed the original mean-variance portfolio optimization problem into a “mean versus second moment” optimization problem that has an additional parameter \( \eta \).

Putting (3.5) or (3.6) into
\[
C(\eta) := E[w^T(\eta) \mu_n] + \lambda E\left([w^T(\eta) \mu_n]^2\right) - \lambda E\left[w^T(\eta) V_n w(\eta)\right],
\]
which is equal to \( E[w^T(\eta) r] - \lambda \text{Var}[w^T(\eta) r] \) by (3.4), we can use Brent’s method (Press et al., pp. 359-362) to maximize \( C(\eta) \). It should be noted that this argument implicitly assumes that the maximum of (3.1) is attained by some \( w \) and is finite. Whereas this assumption is clearly satisfied when there is no short selling as in (3.5), it may not hold when short selling is allowed. In fact, the explicit formula of \( w(\eta) \) in (3.6) can be used to express (3.8) as a quadratic function of \( \eta \):
\[
C(\eta) = \frac{\eta^2}{4\lambda} E\left(\left(B_n - \frac{A_n^2}{C_n}\right) \left(B_n - \frac{A_n^2}{C_n} - 1\right)\right) + \eta E\left(\left(\frac{1}{2\lambda} + \frac{A_n}{C_n}\right) \left(B_n - \frac{A_n^2}{C_n}\right)\right) + E\left(\frac{A_n}{C_n} + \lambda \frac{A_n^2}{C_n} - C_n\right).
\]
which has a maximum only if

\[(3.9) \quad E\left\{\left(B_n - \frac{A_n^2}{C_n}\right)\left(B_n - \frac{A_n^2}{C_n} - 1\right)\right\} < 0.\]

In the case \(E\left\{\left(B_n - \frac{A_n^2}{C_n}\right)\left(B_n - \frac{A_n^2}{C_n} - 1\right)\right\} > 0\), \(C(\eta)\) has a minimum instead and approaches to \(\infty\) as \(|\eta| \to \infty\). In this case, (3.1) has an infinite value and should be defined as a supremum (which is not attained) instead of a maximum.

\textit{Remark}. Let \(\Sigma_n\) denote the posterior covariance matrix given \(\mathcal{R}_n\). Note that the law of iterated conditional expectations, from which (3.4) follows, has the following analog for \(\text{Var}(W)\):

\[(3.10) \quad \text{Var}(W) = E[\text{Var}(W|\mathcal{R}_n)] + \text{Var}[E(W|\mathcal{R}_n)] = E(w^T \Sigma_n w) + \text{Var}(w^T \mu_n).\]

Using \(\Sigma_n\) to replace \(\Sigma\) in the optimal weight vector that assumes \(\mu\) and \(\Sigma\) to be known, therefore, ignores the variance of \(w^T \mu_n\) in (3.10), and this omission is an important root cause for the Markowitz optimization enigma related to “plug-in” efficient frontiers.

\section*{4. Empirical Bayes, bootstrap approximation and frequentist risk}

For more flexible modeling, one can allow the prior distribution in the preceding Bayesian approach to include unspecified hyperparameters, which can be estimated from the training sample by maximum likelihood, or method of moments or other methods. For example, for the conjugate prior (2.2), we can assume \(\nu\) and \(\Psi\) to be functions of certain hyperparameters that are associated with a multifactor model of the type (2.1). This amounts to using an empirical Bayes model for \((\mu, \Sigma)\) in the stochastic optimization problem (3.1). Besides a prior distribution for \((\mu, \Sigma)\), (3.1) also requires specification of the common distribution of the i.i.d. returns to evaluate \(E_{\mu, \Sigma}(w^T r_{n+1})\) and \(\text{Var}_{\mu, \Sigma}(w^T r_{n+1})\). The bootstrap provides a nonparametric method to evaluate these quantities, as described below.

\subsection*{4.1. Bootstrap estimate of performance}

To begin with, note that we can evaluate the frequentist performance of the Bayes or other asset allocation rules by making use of the bootstrap method. The bootstrap samples \(\{r_{1}^*, \ldots, r_{b}^*\}\) drawn with replacement from the observed sample \(\{r_1, \ldots, r_n\}\), \(1 \leq b \leq B\), can be used to estimate its \(E_{\mu, \Sigma}(w_n^T r_{n+1}) = E_{\mu, \Sigma}(w_n^T \mu)\) and \(\text{Var}_{\mu, \Sigma}(w_n^T r_{n+1}) = E_{\mu, \Sigma}(w_n^T \Sigma w_n) + \text{Var}_{\mu, \Sigma}(w_n^T \mu)\) of various portfolios II whose weight vectors \(w_n\) may depend on \(r_1, \ldots, r_n\), e.g., the Bayes or empirical Bayes weight vector described above. In particular, we can use Bayes or other estimators for \(\mu_n\) and \(V_n\) in (3.5) or (3.6) and then choose \(\eta\) to maximize the bootstrap estimate of
This is tantamount to using the empirical distribution of \( \mathbf{r}_1, \ldots, \mathbf{r}_n \) to be the common distribution of the returns. In particular, using \( \hat{\mathbf{r}} \) for \( \mathbf{r}_n \) in (3.5) and the second moment matrix \( n^{-1} \sum_{t=1}^{n} \mathbf{r}_t \mathbf{r}_t^T \) of the empirical distribution for \( \mathbf{V}_n \) in (3.6) provides a “nonparametric empirical Bayes” implementation of the optimal rule in Section 3.

### 4.2. A simulation study of Bayes and frequentist rewards

The following simulation study assumes i.i.d. annual returns (in %) of \( m = 5 \), \( \kappa \) matrix are generated from the normal and inverted Wishart prior distribution (2.2) with \( \kappa = 5 \), \( n_0 = 10 \), \( \mathbf{v} = (2.48, 2.17, 1.61, 3.42)^T \) and the hyperparameter \( \Psi \) given by

\[
\begin{align*}
\Psi_{11} &= 3.37, \quad \Psi_{22} = 4.22, \quad \Psi_{33} = 2.75, \quad \Psi_{44} = 8.43, \quad \Psi_{12} = 2.04, \\
\Psi_{13} &= 0.32, \quad \Psi_{14} = 1.59, \quad \Psi_{23} = -0.05, \quad \Psi_{24} = 3.02, \quad \Psi_{34} = 1.08.
\end{align*}
\]

We consider four scenarios for the case \( n = 6 \) without short selling. The first scenario assumes this prior distribution and studies the Bayesian reward for \( \lambda = 1.5 \) and 10. The other scenarios consider the frequentist reward at three values of \((\mathbf{\mu}, \Sigma)\) generated from the prior distribution. These values, denoted by Freq 1, Freq 2, Freq3, are:

- **Freq 1**: \( \mathbf{\mu} = (2.42, 1.88, 1.58, 3.47)^T \), \( \Sigma_{11} = 1.17, \Sigma_{22} = 0.82, \Sigma_{33} = 1.37, \Sigma_{44} = 2.86, \Sigma_{12} = 0.79, \Sigma_{13} = 0.84, \Sigma_{14} = 1.61, \Sigma_{23} = 0.61, \Sigma_{24} = 1.23, \Sigma_{34} = 1.35. \)
- **Freq 2**: \( \mathbf{\mu} = (2.59, 2.29, 1.25, 3.13)^T \), \( \Sigma_{11} = 1.32, \Sigma_{22} = 0.67, \Sigma_{33} = 1.43, \Sigma_{44} = 1.03, \Sigma_{12} = 0.75, \Sigma_{13} = 0.85, \Sigma_{14} = 0.68, \Sigma_{23} = 0.32, \Sigma_{24} = 0.44, \Sigma_{34} = 0.61. \)
- **Freq 3**: \( \mathbf{\mu} = (1.91, 1.58, 1.03, 2.76)^T \), \( \Sigma_{11} = 1.00, \Sigma_{22} = 0.83, \Sigma_{33} = 0.35, \Sigma_{44} = 0.62, \Sigma_{12} = 0.73, \Sigma_{13} = 0.26, \Sigma_{14} = 0.36, \Sigma_{23} = 0.16, \Sigma_{24} = 0.50, \Sigma_{34} = 0.14. \)

Table 1 compares the Bayes rule with three other rules: (a) the “oracle” rule that assumes \( \mathbf{\mu} \) and \( \Sigma \) to be known, (b) the plug-in rule that replaces \( \mathbf{\mu} \) and \( \Sigma \) by the sample estimates of \( \mathbf{\mu} \) and \( \Sigma \), and (c) the nonparametric empirical Bayes (NPEB) rule described in Section 4.1. Note that although both (b) and (c) use the sample mean vector and sample covariance (or second moment) matrix, (b) simply plugs the sample estimates into the oracle rule while (c) uses the empirical distribution to replace the common distribution of the returns in the Bayes rule that maximizes (3.1). For the plug-in rule, the quadratic programming procedure may have numerical difficulties if the sample covariance matrix is nearly singular. If it should happen, we use the default option of adding 0.005\( \mathbf{I} \) to the sample covariance matrix. Each result in Table 1 is based on 100 simulations, and the standard errors are given in parentheses. In each scenario, the reward of the NPEB rule is close to that of the Bayes...
rule and somewhat smaller than that of the oracle rule. The plug-in rule has substantially smaller rewards, especially for larger values of $\lambda$.

**4.3. Comparison of the $(\sigma, \mu)$ plots of different portfolios.** The set of points in the $(\sigma, \mu)$ plane that correspond to the returns of portfolios of the $m$ assets is called the feasible region. As $\lambda$ varies over $(0, \infty)$, the $(\sigma, \mu)$ values of the oracle rule correspond to Markowitz’s efficient frontier which assumes known $\mu$ and $\Sigma$ and which is the upper left boundary of the feasible region. For portfolios whose weights do not assume knowledge of $\mu$ and $\Sigma$, the $(\sigma, \mu)$ values lie on the right of Markowitz’s efficient frontier. Figure 1 plots the $(\sigma, \mu)$ values of different portfolios formed from $m = 4$ assets without short selling and a training sample of size $n = 6$ when $(\mu, \Sigma)$ is given by the frequentist scenario Freq 1 above. Markowitz’s efficient frontier is computed analytically by varying $\mu_*$ in (1.1) over a grid of values. The $(\sigma, \mu)$ curves of the plug-in, the Ledoit-Wolf and Michaud’s resampled portfolios are computed by Monte Carlo, using 100 simulated paths, for each value of $\mu_*$ in a grid ranging from 2.0 to 3.47. The $(\sigma, \mu)$ curve of NPEB portfolio is also obtained by Monte Carlo simulations with 100 runs, by using different values of $\lambda > 0$ in a grid. This curve is relatively close to Markowitz’s efficient frontier among the $(\sigma, \mu)$ curves of various portfolios that do not assume knowledge of $\mu$ and $\Sigma$, as shown in Figure 1. For the Ledoit-Wolf portfolio, which is labeled “Shrinkage” in Figures 1 and 2 and also in Table 3, we use a constant correlation model for $\hat{F}$ in (2.4), which can be implemented by their software available at www.ledoit.net. Note that Markowitz’s efficient frontier has $\mu$ values ranging from 2.0 to 3.47, which is the largest component of $\mu$ in Figure 1. The $(\sigma, \mu)$ curve of NPEB lies below the efficient frontier, and further below are the $(\sigma, \mu)$ curves of Michaud’s, shrinkage and plug-in portfolios, in decreasing order.

**INSERT FIGURE 1 ABOUT HERE**

The highest values 3.22, 3.22 and 3.16 of $\mu$ for the plug-in, shrinkage and Michaud’s portfolios in Figure 1 are attained with a target value $\mu_* = 3.47$ and the corresponding values of $\sigma$ are 1.54, 1.54 and 3.16, respectively. Note that without short selling, the constraint $\mathbf{w}^T \hat{\mu} = \mu_*$ used in these portfolios cannot hold if $\max_{1 \leq i \leq 4} \hat{\mu}_i < \mu_*$. We therefore need a default option, such as replacing $\mu_*$ by $\min(\mu_*, \max_{1 \leq i \leq 4} \hat{\mu}_i)$, to implement the optimization procedures for these portfolios. In contrast, the NPEB portfolio can always be implemented for any given value of $\lambda$. In particular, for $\lambda = 0.001$, the NPEB portfolio has
\[ \mu = 3.470 \text{ and } \sigma = 1.691. \]

5. Connecting theory to practice. While Section 4 has considered practical implementation of the theory in Section 3, we develop the methodology further in this section to connect the basic theory to practice.

5.1 The Sharpe ratios and choice of \( \lambda \). As pointed out in Section 1, the \( \lambda \) in Section 3 is related to how risk-averse one is when one tries to maximize the mean return \( \mu \) of a portfolio. It represents a penalty on the risk that is measured by the variance of the portfolio’s return. In practice, it may be difficult to specify an investor’s risk aversion parameter \( \lambda \) that is needed in the theory in Section 3.1. A commonly used performance measure of a portfolio’s performance is the Sharpe ratio \( (\mu - \mu_0)/\sigma \), which is the excess return per unit of risk; the excess is measured by \( \mu - \mu_0 \), where \( \mu_0 \) is a benchmark mean return. We can regard \( \lambda \) as a tuning parameter, and choose it to maximize the Sharpe ratio by modifying the NPEB procedure in Section 3.2, where the bootstrap estimate of \( E_{\mu, \Sigma} [w^T(\eta)r] - \lambda \text{Var}_{\mu, \Sigma} [w^T(\eta)r] \) is used to find the portfolio weight \( w_\lambda \) that solves the optimization problem (3.3). Specifically, we use the bootstrap estimate of the Sharpe ratio

\[
\left\{ E_{\mu, \Sigma}(w_\lambda r) - \mu_0 \right\} / \sqrt{\text{Var}_{\mu, \Sigma}(w_\lambda^T r)}
\]

of \( w_\lambda \), and maximize the estimate Sharpe ratios over \( \lambda \).

5.2. Dimension reduction when \( m \) is not small relative to \( n \). Another statistical issue encountered in practice is the large number \( m \) of assets relative to the number \( n \) of past periods in the training sample, making it difficult to estimate \( \mu \) and \( \Sigma \) satisfactorily. Using factor models that are related to domain knowledge as in Section 2.1 helps reduce the number of parameters to be estimated in an empirical Bayes approach. Another useful way of dimension reduction is to exclude assets with markedly inferior Sharpe ratios from consideration. The only potential advantage of including them in the portfolio is that they may be able to reduce the portfolio variance if they are negatively correlated with the “superior” assets. However, since the correlations are unknown, such advantage is unlikely when they are not estimated well enough.

Suppose we include in the simulation study of Section 4.2 two more assets so that all asset returns are jointly normal. The additional hyperparameters of the normal and inverted Wishart prior distribution (2.2) are \( \nu_5 = -0.014, \nu_6 = -0.064, \Psi_{55} = 2.02, \Psi_{66} = 10.32, \Psi_{56} = 0.90, \Psi_{15} = -0.17, \Psi_{25} = -0.03, \Psi_{45} = -0.91, \Psi_{45} = -0.33, \Psi_{16} = -3.40, \Psi_{26} = -3.99, \Psi_{36} = -0.08 \) and \( \Psi_{46} = -3.58 \). As in Section 4.2, we consider four scenarios
for the case of \( n = 8 \) without short selling, the first of which assumes this prior distribution and studies the Bayesian reward for \( \lambda = 1, 5 \) and 10. Table 2 shows the rewards for the four rules in Section 4.2, and each result is based on 100 simulations. Note that the value of the reward function does not show significant change with the inclusion of two additional stocks, which have negative correlations with the four stocks in Section 4.2 but have low Sharpe ratios.

5.3 Extension to time series models of returns. An important assumption in the modification of Markowitz’s theory in Section 3.2 is that \( r_t \) are i.i.d. with mean \( \mu \) and covariance matrix \( \Sigma \). Diagnostic checks of the extent to which this assumption is violated should be carried out in practice. The stochastic optimization theory in Section 3.1 does not actually need this assumption and only requires the posterior mean and second moment matrix of the return vector for the next period in (3.4). Therefore one can modify the “working i.i.d. model” accordingly when the diagnostic checks reveal such modifications are needed.

A simple method to introduce such modification is to use a stochastic regression model of the form

\[
(5.2) \quad r_{it} = \beta_i^T x_{it} + \epsilon_{it},
\]

where the components of \( x_{it} \) include 1, factor variables such as the return of a market portfolio like S&P500 at time \( t - 1 \), and lagged variables \( r_{i,t-1}, r_{i,t-2}, ... \). The basic idea underlying (5.2) is to introduce covariates (including lagged variables to account for time series effects) so that the errors \( \epsilon_{it} \) can be regarded as i.i.d., as in the working i.i.d. model. The regression parameter \( \beta_i \) can be estimated by the method of moments, which is equivalent to least squares. We can also include heteroskedasticity by assuming that \( \epsilon_{it} = s_{it}(\gamma_i)z_{it} \), where \( z_{it} \) are i.i.d. with mean 0 and variance 1, \( \gamma_i \) is a parameter vector which can be estimated by maximum likelihood or generalized method of moments, and \( s_{it} \) is a given function that depends on \( r_{i,t-1}, r_{i,t-2}, ... \). A well known example is the GARCH(1,1) model

\[
(5.3) \quad \epsilon_{it} = s_{it}z_{it}, \quad s_{it}^2 = \omega_i + a_is_{i,t-1}^2 + b_ir_{i,t-1}^2,
\]

for which \( \gamma_i = (\omega_i, a_i, b_i) \).

Consider the stochastic regression model (5.2). As noted in Section 3.2, a key ingredient in the optimal weight vector that solves the optimization problem (3.1) is \((\mu_n, V_n)\), where
\( \mu_n = E(r_{n+1}|R_n) \) and \( \mathbf{V}_n = E(\mathbf{r}_{n+1}\mathbf{r}_{n+1}^T|R_n) \). Instead of the classical model of i.i.d. returns, one can combine domain knowledge of the \( m \) assets with time series modeling to obtain better predictors of future returns via \( \mu_n \) and \( \mathbf{V}_n \). The regressors \( \mathbf{x}_{it} \) in (5.2) can be chosen to build a combined substantive-empirical model for prediction; see Section 7.5 of Lai and Xing (2008). Since the model (5.2) is intended to produce i.i.d. \( \epsilon_t = (\epsilon_{1t}, \ldots, \epsilon_{mt})^T \), or i.i.d. \( \mathbf{z}_t = (z_{1t}, \ldots, z_{mt})^T \) after adjusting for conditional heteroskedasticity as in (5.3), we can still use the NPEB approach to determine the optimal weight vector, bootstrapping from the estimated common distribution of \( \epsilon_t \) (or \( \mathbf{z}_t \)). Note that (5.2)-(5.3) models the asset returns separately, instead of jointly in a multivariate regression or multivariate GARCH model which has the difficulty of having too many parameters to estimate. While the vectors \( \epsilon_t \) (or \( \mathbf{z}_t \)) are assumed to be i.i.d., (5.2) (or (5.3)) does not assume their components to be uncorrelated since it treats the components separately rather than jointly. The conditional cross-sectional covariance between the returns of assets \( i \) and \( j \) given \( R_n \) is given by

\[
\text{Cov}(r_{i,n+1}, r_{j,n+1}|R_n) = s_{i,n+1}(\gamma_i)s_{j,n+1}(\gamma_j)\text{Cov}(z_{i,n+1}, z_{j,n+1}|R_n),
\]

for the model (5.2)-(5.3). Note that (5.3) determines \( s_{i,n+1}^2 \) recursively from \( R_n \) and that \( \mathbf{z}_{n+1} \) is independent of \( R_n \) and therefore its covariance matrix can be consistently estimated from the residuals \( \mathbf{z}_t \). Under (5.2)-(5.3), the NPEB approach uses the following formulas for \( \mu_n \) and \( \mathbf{V}_n \) in (3.5):

\[
\begin{align*}
\mu_n &= (\hat{\beta}_1^T \mathbf{x}_{1,n+1}, \ldots, \hat{\beta}_m^T \mathbf{x}_{m,n+1})^T, \\
\mathbf{V}_n &= \mu_n\mu_n^T + \left( \hat{s}_{i,n+1}\hat{s}_{j,n+1} \hat{\sigma}_{ij} \right)_{1 \leq i,j \leq n},
\end{align*}
\]

in which \( \hat{\beta}_i \) is the least squares estimate of \( \beta_i \), and \( \hat{s}_{i,n+1} \) and \( \hat{\sigma}_{ij} \) are the usual estimates of \( s_{i,n+1} \) and \( \text{Cov}(z_{i,1}, z_{j,1}) \) based on \( R_n \).

6. An empirical study. In this section we describe an empirical study of the performance of the proposed approach and other methods for mean-variance portfolio optimization when the means and covariances of the underlying asset returns are unknown. The study considers the monthly excess returns of 233 stocks with respect to the S&P500 Index from January 1974 to December 2008. To simplify the study, we work directly with the excess returns \( e_{it} = r_{it} - u_t \) instead of the returns \( r_{it} \) of the \( i \)th stock and \( u_t \) of the S&P500 Index. Such simplification also has the effect of making the time series more stationary, as will be discussed in Section 6.3. The data set and the description of these stocks are available at the Web site http://www.stanford.edu/~xing/Data/data_excess_ret.txt. We study out-of-sample performance of the monthly excess returns of different portfolios of these stocks.
for each month after the first ten years (120 months). Specifically, we use sliding windows of \( n = 120 \) months of training data to construct a portfolio for the subsequent month, allowing no short selling. The excess return \( e_t \) of the portfolio thus formed for investment in month \( t \) gives the realized (out-of-sample) excess return. As \( t \) varies over the monthly test periods from January 1984 to December 2008, we can (i) add up the realized excess returns to give the cumulative realized excess return \( \sum_{t=1}^{T} e_t \) up to time \( t \), and (ii) find the average realized excess return and the standard deviation so that the ratio gives the realized Sharpe ratio \( \sqrt{12} \bar{e}/s_e \), where \( \bar{e} \) is the sample average of the monthly excess returns and \( s \) is the corresponding sample standard deviation, using \( \sqrt{12} \) to annualize the ratio, as in Ledoit and Wolf (2004). Noting that the realized Sharpe ratio is a summary statistic, in the form of mean divided by standard deviation, of the monthly excess returns in the 300 test periods, we find it more informative to supplement this commonly used measure of investment performance with the time series plot of cumulative realized excess returns, from which the realized returns \( e_t \) can be retrieved by differencing.

We use these two performance measures to compare NPEB with the Ledoit-Wolf (labeled “shrinkage”), plug-in and Michaud’s portfolios described in Sections 2 and 4.3. For a given training sample, we first select stocks whose Sharpe ratios are the \( m = 50 \) largest among 233 stocks; see Section 5.2 and Ledoit and Wolf (2004). Then we compute the NPEB, plug-in, Ledoit-Wolf and resampled portfolios. For each training sample, the NPEB procedure first computes a portfolio for each \( \lambda = 2^i, i = -3, -2, \ldots, 11 \) and then chooses the \( \lambda \) that gives the largest bootstrap estimate of the ratio of the mean to the standard deviation of the excess returns; see Section 5.1. Similarly, since the plug-in, shrinkage and Michaud’s portfolios involves a target return \( \mu_* \), we apply these procedures with \( \mu_*^{(i)} = e_{\min} + (e_{\max} - e_{\min})(i/100) \) for \( 1 \leq i \leq 100 \), where \( e_{\max} \) and \( e_{\min} \) are the largest and smallest values, respectively, of the mean excess returns in the training sample, and set \( \mu_* \) equal to the \( \mu_*^{(i)} \) that gives the largest realized Sharpe ratio for the training sample.

6.1 Working model of i.i.d. returns. Table 3 gives the realized Sharpe ratios of the NPEB, plug-in, shrinkage and Michaud’s portfolios described in the preceding paragraph, and Figure 2 plots their cumulative realized returns; see the first paragraph of this section. As in the plug-in, shrinkage and Michaud’s methods, we assume that the returns are i.i.d. for NPEB. Table 3 shows that NPEB has larger realized Sharpe ratios than the plug-in, shrinkage and Michaud’s portfolios.
6.2 Improving NPEB with simple time series models of excess returns. We have performed standard diagnostic checks of the i.i.d. assumption on the excess returns by examining their time series plots and autocorrelation functions; see Section 5.1 of Lai and Xing (2008). In particular, the Ljung-Box test that involves autocorrelations of lags up to 20 months rejects the null hypothesis of i.i.d. excess returns at 5% confidence level for approximately 30% of the 233 stocks; 16 stocks have $p$-value < 0.001, 55 stocks have $p$-values in [0.001, 0.05), 25 stocks have $p$-values in [0.05, 0.1) and 137 stocks have $p$-values $\geq$ 0.1. Figure 3 plots the time series and autocorrelation functions of the 5 stocks that have $p$-value below 0.001 by the Ljung-Box test.

These diagnostic checks suggest that using a time series model as a working model for excess returns may improve the performance of NPEB. The simplest model to try is the AR(1) model $e_{it} = \alpha_i + \gamma_i e_{i,t-1} + \epsilon_{it}$, which includes the i.i.d. model as a special case with $\gamma_i = 0$. Assuming this time series model for the excess returns, we can apply the NPEB procedure in Section 5.3 to the training sample and thereby obtain the NPEB$_{AR}$ portfolio for the test sample. Table 3 and Figure 2 also show corresponding results for NPEB$_{AR}$. The cumulative realized excess returns of NPEB$_{AR}$ are substantially larger than those of NPEB.

6.3 Additional results and discussion. The AR(1) model uses $x_{it} = (1, e_{i,t-1})^T$ as the predictor in a linear regression model for $e_{i,t}$. To improve prediction performance, one can include additional predictor variables, e.g., the return $u_{t-1}$ of the S&P500 Index in the preceding period. Assuming the stochastic regression model $e_{i,t} = (1, e_{i,t-1}, u_{t-1})\beta_i + \epsilon_{it}$, we have also used the training sample to form the NPEB$_{SR}$ portfolio for the test sample. A further extension of the stochastic regression model assumes the GARCH(1,1) model (5.3) for $\epsilon_{i,t}$, which we also use as the working model of the training sample to form the NPEB$_{SRG}$ portfolio for the test sample. Table 3 and Figure 2 show substantial improvements in using these enhancements of the NPEB procedure.

The dataset at the Web site mentioned in the first paragraph of Section 6 also contains $u_t$. From $e_{i,t}$ and $u_t$, one can easily retrieve the returns $r_{it} = e_{it} + u_t$ of the $i$th stock, and one may ask why we have not used $r_{it}$ directly to fit the stochastic regression model (5.2). The main reason is that the model (5.2) is very flexible and should incorporate all important predictors in $x_{it}$ for the stock’s performance at time $t$ whereas our objective is this paper is to introduce a new statistical approach to the Markowitz optimization enigma rather than combining fundamental and empirical analyses, as described in Chapter 11 of Lai.
and Xing (2008), of these stocks. Moreover, in order to compare with previous approaches, the working model of i.i.d. stock returns suffices and we actually began our empirical study with this assumption. However, time series plots of the stock returns and structural changes in the economy and the financial markets during this period show clear departures from this working model for the stock returns. On the other hand, we found the excess returns to be more “stationary” when we used the excess returns $e_{it}$ instead of $r_{it}$, following the empirical study of Ledoit and Wolf (2004). In fact, the realized Sharpe ratio $\sqrt{\frac{\bar{e}}{s_e}}$ based on the excess returns was introduced by them, who called it the \textit{ex post information ratio} instead. Strictly speaking, the denominator in the Sharpe ratio should be the standard deviation of the portfolio return rather than the standard deviation of the excess return, and their terminology “information ratio” avoids this confusion. We still call it the “realized Sharpe ratio” to avoid introducing new terminology for readers who are less familiar with the investment than the statistical background.

As an illustration, the top panel of Figure 4 gives the time series plots of returns of TECO Energy Inc. (TE) and of the S&P500 Index during this period, and the middle panel gives the time series plot of the excess returns. The Ljung-Box test, which involves autocorrelations of lags up to 20 months, of the i.i.d. assumption has $p$-value 0.0444 for the monthly returns of TE and 0.9810 for the excess returns, and therefore rejects the i.i.d. assumption for the actual but not the excess returns. This is also shown graphically by the autocorrelation functions in the bottom panel of Figure 5.

As an illustration, the top panel of Figure 4 gives the time series plots of returns of TECO Energy Inc. (TE) and of the S&P500 Index during this period, and the middle panel gives the time series plot of the excess returns. The Ljung-Box test, which involves autocorrelations of lags up to 20 months, of the i.i.d. assumption has $p$-value 0.0444 for the monthly returns of TE and 0.9810 for the excess returns, and therefore rejects the i.i.d. assumption for the actual but not the excess returns. This is also shown graphically by the autocorrelation functions in the bottom panel of Figure 5.

As an illustration, the top panel of Figure 4 gives the time series plots of returns of TECO Energy Inc. (TE) and of the S&P500 Index during this period, and the middle panel gives the time series plot of the excess returns. The Ljung-Box test, which involves autocorrelations of lags up to 20 months, of the i.i.d. assumption has $p$-value 0.0444 for the monthly returns of TE and 0.9810 for the excess returns, and therefore rejects the i.i.d. assumption for the actual but not the excess returns. This is also shown graphically by the autocorrelation functions in the bottom panel of Figure 5.

In contrast to the simple AR(1) model for excess returns, care must be taken to handle nonstationarity when we build time series models for stock returns. It seems that a regressor such as the return $u_t$ of S&P500 Index should be included to take advantage of the co-movements of $r_{it}$ and $u_t$. However, since $u_t$ is not observed at time $t$, one may need to have good predictors of $u_t$ which should consist not only of the past S&P500 returns but also macroeconomic variables. Of course, stock-specific information such as the firm’s earnings performance and forecast and its sector’s economic outlook should also be considered. This means that thorough fundamental analysis, as carried out by professional stock analysts and economists in investment banks, should be incorporated into the model (5.2). Since this is clearly beyond the scope of the paper, we have focused on simple models to illustrate the benefit of building good models for $r_{n+1}$ in our stochastic optimization approach. Our approach can be very powerful if one can combine domain knowledge with the statistical
modeling that we illustrate here. However, we have not done this in the present empirical study because using our inadequate knowledge of these stocks to specify (5.2) will be a disservice to the power and versatility of the proposed Bayesian or NPEB approach.

7. Concluding remarks. The “Markowitz enigma” has been attributed to (a) sampling variability of the plug-in weights (hence use of resampling to correct for bias due to nonlinearity of the weights as a function of the mean vector and covariance matrix of stocks) or (b) inherent difficulties of estimation of high-dimensional covariance matrices in the plug-in approach. Like the plug-in approach, subsequent refinements that attempt to address (a) or (b) still follow closely Markowitz’s solution for efficient portfolios, constraining the unknown mean to equal to some target returns. This tends to result in relatively low Sharpe ratios when no or limited short selling is allowed, as noted in Sections 4.3 and 6.1. Another difficulty with the plug-in and shrinkage approaches is that their measure of “risk” does not account for the uncertainties in the parameter estimates. Incorporating these uncertainties via a Bayesian approach results in a much harder stochastic optimization problem than Markowitz’s deterministic optimization problem, which we have been able to solve by introducing an additional parameter $\eta$.

Our solution of this stochastic optimization problem opens up new possibilities in extending Markowitz’s mean-variance portfolio optimization theory to the case where the means and covariances of the asset returns for the next investment period are unknown. As pointed out in Section 5.3, our solution only requires the posterior mean and second moment matrix of the return vector for the next period, and one can use fundamental analysis and statistical modeling to develop Bayesian or empirical Bayes models with good predictive properties, e.g., by using (5.2) with suitably chosen $\mathbf{x}_{it}$.

Markowitz’s mean-variance portfolio optimization theory is a single-period theory that does not consider transaction costs. In practice asset allocation by a portfolio manager is multi-period or dynamic and incurs transaction costs; see Section 11.3 of Lai and Xing (2008). The methods and results developed herein to resolve the Markowitz enigma suggest that combining techniques in stochastic control and applied statistics may provide a practical solution to this long-standing problem in financial economics.

REFERENCES


Table 1: Rewards of four portfolios formed from $m = 4$ assets

<table>
<thead>
<tr>
<th>$(\mu, \Sigma)$</th>
<th>Bayes</th>
<th>Plug-in</th>
<th>Oracle</th>
<th>NPEB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\lambda = 1$</td>
<td>Bayes</td>
<td>0.0325 (4.37e-5)</td>
<td>0.0317 (4.35e-5)</td>
<td>0.0328 (3.91e-5)</td>
</tr>
<tr>
<td></td>
<td>Freq 1</td>
<td>0.0332 (6.41e-6)</td>
<td>0.0324 (1.16e-5)</td>
<td>0.0332</td>
</tr>
<tr>
<td></td>
<td>Freq 2</td>
<td>0.0293 (1.56e-5)</td>
<td>0.0282 (1.09e-5)</td>
<td>0.0298</td>
</tr>
<tr>
<td></td>
<td>Freq 3</td>
<td>0.0267 (1.00e-5)</td>
<td>0.0257 (1.20e-5)</td>
<td>0.0268</td>
</tr>
<tr>
<td>$\lambda = 5$</td>
<td>Bayes</td>
<td>0.0263 (4.34e-5)</td>
<td>0.0189 (2.55e-5)</td>
<td>0.0268 (3.71e-5)</td>
</tr>
<tr>
<td></td>
<td>Freq 1</td>
<td>0.0272 (9.28e-6)</td>
<td>0.0182 (1.29e-5)</td>
<td>0.0273</td>
</tr>
<tr>
<td></td>
<td>Freq 2</td>
<td>0.0233 (2.27e-5)</td>
<td>0.0183 (8.69e-6)</td>
<td>0.0240</td>
</tr>
<tr>
<td></td>
<td>Freq 3</td>
<td>0.0235 (1.26e-5)</td>
<td>0.0159 (6.66e-5)</td>
<td>0.0237</td>
</tr>
<tr>
<td>$\lambda = 10$</td>
<td>Bayes</td>
<td>0.0188 (5.00e-5)</td>
<td>0.0067 (1.75e-5)</td>
<td>0.0193 (3.85e-5)</td>
</tr>
<tr>
<td></td>
<td>Freq 1</td>
<td>0.0197 (1.59e-5)</td>
<td>0.0063 (7.40e-6)</td>
<td>0.0200</td>
</tr>
<tr>
<td></td>
<td>Freq 2</td>
<td>0.0157 (2.58e-5)</td>
<td>0.0159 (6.66e-5)</td>
<td>0.0168</td>
</tr>
<tr>
<td></td>
<td>Freq 3</td>
<td>0.0195 (1.29e-5)</td>
<td>0.0083 (3.53e-6)</td>
<td>0.0198</td>
</tr>
</tbody>
</table>

Table 2: Rewards of four portfolios formed from $m = 6$ assets

<table>
<thead>
<tr>
<th>$(\mu, \Sigma)$</th>
<th>Bayes</th>
<th>Plug-in</th>
<th>Oracle</th>
<th>NPEB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\lambda = 1$</td>
<td>Bayes</td>
<td>0.0325 (6.09e-5)</td>
<td>0.0319 (5.72e-6)</td>
<td>0.0331 (5.01e-5)</td>
</tr>
<tr>
<td></td>
<td>Freq 1</td>
<td>0.0290 (4.25e-5)</td>
<td>0.0280 (3.00e-5)</td>
<td>0.0308</td>
</tr>
<tr>
<td></td>
<td>Freq 2</td>
<td>0.0286 (2.04e-5)</td>
<td>0.0277 (1.91e-5)</td>
<td>0.0296</td>
</tr>
<tr>
<td></td>
<td>Freq 3</td>
<td>0.0302 (4.32e-5)</td>
<td>0.0295 (3.21e-5)</td>
<td>0.0329</td>
</tr>
<tr>
<td>$\lambda = 5$</td>
<td>Bayes</td>
<td>0.0256 (5.39e-5)</td>
<td>0.0184 (3.44e-5)</td>
<td>0.0263 (4.28e-5)</td>
</tr>
<tr>
<td></td>
<td>Freq 1</td>
<td>0.0225 (5.43e-5)</td>
<td>0.0150 (1.65e-5)</td>
<td>0.0248</td>
</tr>
<tr>
<td></td>
<td>Freq 2</td>
<td>0.0238 (2.24e-5)</td>
<td>0.0163 (8.70e-6)</td>
<td>0.0242</td>
</tr>
<tr>
<td></td>
<td>Freq 3</td>
<td>0.0208 (4.00e-5)</td>
<td>0.0176 (3.57e-5)</td>
<td>0.0239</td>
</tr>
<tr>
<td>$\lambda = 10$</td>
<td>Bayes</td>
<td>0.0170 (6.74e-5)</td>
<td>0.0386 (4.06e-5)</td>
<td>0.0179 (5.96e-5)</td>
</tr>
<tr>
<td></td>
<td>Freq 1</td>
<td>0.0142 (5.57e-5)</td>
<td>0.0039 (9.66e-6)</td>
<td>0.0174</td>
</tr>
<tr>
<td></td>
<td>Freq 2</td>
<td>0.0175 (2.60e-5)</td>
<td>0.0060 (1.29e-5)</td>
<td>0.0183</td>
</tr>
<tr>
<td></td>
<td>Freq 3</td>
<td>0.0096 (4.26e-5)</td>
<td>0.0024 (3.70e-5)</td>
<td>0.0123</td>
</tr>
</tbody>
</table>

Table 3: Annualized realized Sharpe ratios of different procedures

<table>
<thead>
<tr>
<th></th>
<th>Plug-in</th>
<th>Shrinkage</th>
<th>Michaud</th>
<th>NPEB</th>
<th>NPEB$_{AR}$</th>
<th>NPEB$_{SR}$</th>
<th>NPEB$_{SRG}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-0.1423</td>
<td>-0.1775</td>
<td>-0.1173</td>
<td>0.0626</td>
<td>0.1739</td>
<td>0.2044</td>
<td>0.3330</td>
</tr>
</tbody>
</table>
Figure 1: $(\sigma, \mu)$ curves of different portfolios.

Figure 2: Realized cumulative excess returns over time.
Figure 3: Excess returns and their autocorrelations of three stocks. The Ljung-Box test statistics (and their $p$-values) are 116.69 (1.11e-15), 55.58 (3.36e-5), 49.41 (2.69e-4), 63.11 (2.33e-6), 90.35 (6.43e-11). The dotted lines in the right panel represent rejection boundaries of 5%-level tests of zero autocorrelation at indicated lag.
Figure 4: Comparison of returns and excess returns. Top panel: returns of S&P500 Index (black) and TE (red); Middle panel: excess returns (blue) of TE; Bottom: Autocorrelations of returns (red) and excess returns (blue) of TE; the dotted lines in the right panel represent rejection boundaries of 5%-level tests of zero autocorrelation at indicated lag.