Two Modifications of Mean-Variance Portfolio Theory

Thomas J. Sargent and John Stachurski

March 15, 2020

1 Contents

• Overview 2
• Appendix 3

2 Overview

2.1 Remarks About Estimating Means and Variances

The famous Black-Litterman (1992) [1] portfolio choice model that we describe in this lecture is motivated by the finding that with high or moderate frequency data, means are more difficult to estimate than variances.

A model of robust portfolio choice that we’ll describe also begins from the same starting point.

To begin, we’ll take for granted that means are more difficult to estimate that covariances and will focus on how Black and Litterman, on the one hand, an robust control theorists, on the other, would recommend modifying the mean-variance portfolio choice model to take that into account.

At the end of this lecture, we shall use some rates of convergence results and some simulations to verify how means are more difficult to estimate than variances.

Among the ideas in play in this lecture will be

• Mean-variance portfolio theory
• Bayesian approaches to estimating linear regressions
• A risk-sensitivity operator and its connection to robust control theory

Let’s start with some imports:

In [1]: import numpy as np
import scipy as sp
import scipy.stats as stat
import matplotlib.pyplot as plt
%matplotlib inline
from ipywidgets import interact, FloatSlider
2.2 Adjusting Mean-variance Portfolio Choice Theory for Distrust of Mean Excess Returns

This lecture describes two lines of thought that modify the classic mean-variance portfolio choice model in ways designed to make its recommendations more plausible.

As we mentioned above, the two approaches build on a common and widespread hunch – that because it is much easier statistically to estimate covariances of excess returns than it is to estimate their means, it makes sense to contemplated the consequences of adjusting investors’ subjective beliefs about mean returns in order to render more sensible decisions.

Both of the adjustments that we describe are designed to confront a widely recognized embarrassment to mean-variance portfolio theory, namely, that it usually implies taking very extreme long-short portfolio positions.

2.3 Mean-variance Portfolio Choice

A risk-free security earns one-period net return $r_f$.

An $n \times 1$ vector of risky securities earns an $n \times 1$ vector $\bar{r} - r_f \mathbf{1}$ of excess returns, where $\mathbf{1}$ is an $n \times 1$ vector of ones.

The excess return vector is multivariate normal with mean $\mu$ and covariance matrix $\Sigma$, which we express either as

$$\bar{r} - r_f \mathbf{1} \sim \mathcal{N}(\mu, \Sigma)$$

or

$$\bar{r} - r_f \mathbf{1} = \mu + C \epsilon$$

where $\epsilon \sim \mathcal{N}(0, I)$ is an $n \times 1$ random vector.

Let $w$ be an $n \times 1$ vector of portfolio weights.

A portfolio consisting $w$ earns returns

$$w'(\bar{r} - r_f \mathbf{1}) \sim \mathcal{N}(w'\mu, w'\Sigma w)$$

The **mean-variance portfolio choice problem** is to choose $w$ to maximize

$$U(\mu, \Sigma; w) = w'\mu - \frac{\delta}{2} w'\Sigma w$$

(1)

where $\delta > 0$ is a risk-aversion parameter. The first-order condition for maximizing (1) with respect to the vector $w$ is

$$\mu = \delta \Sigma w$$

which implies the following design of a risky portfolio:

$$w = (\delta \Sigma)^{-1} \mu$$

(2)
2.4 Estimating the Mean and Variance

The key inputs into the portfolio choice model (2) are

- estimates of the parameters $\mu, \Sigma$ of the random excess return vector $(\bar{r} - r_f \mathbf{1})$
- the risk-aversion parameter $\delta$

A standard way of estimating $\mu$ is maximum-likelihood or least squares; that amounts to estimating $\mu$ by a sample mean of excess returns and estimating $\Sigma$ by a sample covariance matrix.

2.5 The Black-Litterman Starting Point

When estimates of $\mu$ and $\Sigma$ from historical sample means and covariances have been combined with reasonable values of the risk-aversion parameter $\delta$ to compute an optimal portfolio from formula (2), a typical outcome has been $w$’s with extreme long and short positions.

A common reaction to these outcomes is that they are so unreasonable that a portfolio manager cannot recommend them to a customer.

In [2]: np.random.seed(12)

    N = 10  # Number of assets
    T = 200  # Sample size

    # Random market portfolio (sum is normalized to 1)
    w_m = np.random.rand(N)
    w_m = w_m / (w_m.sum())

    # True risk premia and variance of excess return (constructed
    # so that the Sharpe ratio is 1)
    mu = (np.random.randn(N) + 5) / 100  # Mean excess return (risk premium)
    S = np.random.randn(N, N)  # Random matrix for the covariance matrix
    V = S @ S.T  # Turn the random matrix into symmetric psd
    # Make sure that the Sharpe ratio is one
    Sigma = V * (w_m @ mu)**2 / (w_m @ V @ w_m)

    # Risk aversion of market portfolio holder
    delta = 1 / np.sqrt(w_m @ Sigma @ w_m)

    # Generate a sample of excess returns
    excess_return = stats.multivariate_normal(mu, Sigma)
    sample = excess_return.rvs(T)

    # Estimate $\mu$ and $\Sigma$
    mu_est = sample.mean(0).reshape(N, 1)
    Sigma_est = np.cov(sample.T)

    w = np.linalg.solve(delta * Sigma_est, mu_est)

    fig, ax = plt.subplots(figsize=(8, 5))
    ax.set_title('Mean-variance portfolio weights recommendation \ and the market portfolio')
    ax.plot(np.arange(N)+1, w, 'o', c='k', label='$w$ (mean-variance)'
    ax.plot(np.arange(N)+1, w_m, 'o', c='r', label='$w_m$ (market portfolio)')
Black and Litterman’s responded to this situation in the following way:

- They continue to accept (2) as a good model for choosing an optimal portfolio $w$.
- They want to continue to allow the customer to express his or her risk tolerance by setting $\delta$.
- Leaving $\Sigma$ at its maximum-likelihood value, they push $\mu$ away from its maximum value in a way designed to make portfolio choices that are more plausible in terms of conforming to what most people actually do.

In particular, given $\Sigma$ and a reasonable value of $\delta$, Black and Litterman reverse engineered a vector $\mu_{BL}$ of mean excess returns that makes the $w$ implied by formula (2) equal the actual market portfolio $w_m$, so that

$$w_m = (\delta \Sigma)^{-1} \mu_{BL}$$

### 2.6 Details

Let’s define

$$w'_m \mu = (r_m - r_f)$$
as the (scalar) excess return on the market portfolio \( w_m \).

Define

\[
\sigma^2 = w_m' \Sigma w_m
\]
as the variance of the excess return on the market portfolio \( w_m \).

Define

\[
SR_m = \frac{r_m - r_f}{\sigma}
\]
as the **Sharpe-ratio** on the market portfolio \( w_m \).

Let \( \delta_m \) be the value of the risk aversion parameter that induces an investor to hold the market portfolio in light of the optimal portfolio choice rule (2).

Evidently, portfolio rule (2) then implies that \( r_m - r_f = \delta_m \sigma^2 \) or

\[
\delta_m = \frac{r_m - r_f}{\sigma^2}
\]
or

\[
\delta_m = \frac{SR_m}{\sigma}
\]

Following the Black-Litterman philosophy, our first step will be to back a value of \( \delta_m \) from

- an estimate of the Sharpe-ratio, and
- our maximum likelihood estimate of \( \sigma \) drawn from our estimates or \( w_m \) and \( \Sigma \)

The second key Black-Litterman step is then to use this value of \( \delta \) together with the maximum likelihood estimate of \( \Sigma \) to deduce a \( \mu_{BL} \) that verifies portfolio rule (2) at the market portfolio \( w = w_m \)

\[
\mu_m = \delta_m \Sigma w_m
\]

The starting point of the Black-Litterman portfolio choice model is thus a pair \((\delta_m, \mu_m)\) that tells the customer to hold the market portfolio.

---

In [3]:

- **Observed mean excess market return**
  \[
  r_m = w_m \odot \mu_{est}
  \]

- **Estimated variance of the market portfolio**
  \[
  \sigma_m = w_m \odot \Sigma_{est} \odot w_m
  \]

- **Sharpe-ratio**
  \[
  SR_m = \frac{r_m - r_f}{\sigma_m}
  \]

- **Risk aversion of market portfolio holder**
  \[
  d_m = \frac{r_m}{\sigma_m}
  \]

- **Derive "view" which would induce the market portfolio**
  \[
  \mu_m = (d_m \ast \Sigma_{est} \odot w_m).reshape(N, 1)
  \]
2.7 Adding Views

Black and Litterman start with a baseline customer who asserts that he or she shares the market’s views, which means that he or she believes that excess returns are governed by

\[
\tilde{r} - r_f 1 \sim N(\mu_{BL}, \Sigma)
\]  

Black and Litterman would advise that customer to hold the market portfolio of risky securities.

Black and Litterman then imagine a consumer who would like to express a view that differs from the market’s.

The consumer wants appropriately to mix his view with the market’s before using (2) to choose a portfolio.
Suppose that the customer’s view is expressed by a hunch that rather than (3), excess returns are governed by

$$
\tilde{r} - r_f \mathbf{1} \sim \mathcal{N}(\hat{\mu}, \tau \Sigma)
$$

where \( \tau > 0 \) is a scalar parameter that determines how the decision maker wants to mix his view \( \hat{\mu} \) with the market’s view \( \mu_{BL} \).

Black and Litterman would then use a formula like the following one to mix the views \( \hat{\mu} \) and \( \mu_{BL} \):

$$
\tilde{\mu} = (\Sigma^{-1} + (\tau \Sigma)^{-1})^{-1}(\Sigma^{-1}\mu_{BL} + (\tau \Sigma)^{-1}\hat{\mu})
$$

Black and Litterman would then advise the customer to hold the portfolio associated with these views implied by rule (2):

$$
\tilde{w} = (\delta \Sigma)^{-1}\tilde{\mu}
$$

This portfolio \( \tilde{w} \) will deviate from the portfolio \( w_{BL} \) in amounts that depend on the mixing parameter \( \tau \).

If \( \hat{\mu} \) is the maximum likelihood estimator and \( \tau \) is chosen heavily to weight this view, then the customer’s portfolio will involve big short-long positions.

```
In [4]: def black_litterman(λ, μ1, μ2, Σ1, Σ2):
    """
    This function calculates the Black-Litterman mixture mean excess return and covariance matrix
    """
    Σ1_inv = np.linalg.inv(Σ1)
    Σ2_inv = np.linalg.inv(Σ2)

    μ_tilde = np.linalg.solve(Σ1_inv + λ * Σ2_inv, Σ1_inv @ μ1 + λ * Σ2_inv @ μ2)
    return μ_tilde

τ = 1
μ_tilde = black_litterman(1, μ_m, μ_est, Σ_est, τ * Σ_est)

# The Black-Litterman recommendation for the portfolio weights
w_tilde = np.linalg.solve(δ * Σ_est, μ_tilde)

τ_slider = FloatSlider(min=0.05, max=10, step=0.5, value=τ)

@interact(τ=τ_slider)
def BL_plot(τ):
    μ_tilde = black_litterman(1, μ_m, μ_est, Σ_est, τ * Σ_est)
    w_tilde = np.linalg.solve(δ * Σ_est, μ_tilde)

    fig, ax = plt.subplots(1, 2, figsize=(16, 6))
    ax[0].plot(np.arange(N)+1, μ_est, 'o', c='k',
               label=r'\hat{\mu} (subj view)')
    ax[0].plot(np.arange(N)+1, μ_m, 'o', c='r',
               label=r'\mu_{BL} (market)')
2.8 Bayes Interpretation of the Black-Litterman Recommendation

Consider the following Bayesian interpretation of the Black-Litterman recommendation. The prior belief over the mean excess returns is consistent with the market portfolio and is given by

\[ \mu \sim \mathcal{N}(\mu_{BL}, \Sigma) \]

Given a particular realization of the mean excess returns \( \mu \) one observes the average excess returns \( \tilde{\mu} \) on the market according to the distribution
\[ \hat{\mu} | \mu, \Sigma \sim \mathcal{N}(\mu, \tau \Sigma) \]

where \( \tau \) is typically small capturing the idea that the variation in the mean is smaller than the variation of the individual random variable.

Given the realized excess returns one should then update the prior over the mean excess returns according to Bayes rule.

The corresponding posterior over mean excess returns is normally distributed with mean

\[
(\Sigma^{-1} + (\tau \Sigma)^{-1})^{-1}(\Sigma^{-1}\mu_{BL} + (\tau \Sigma)^{-1}\hat{\mu})
\]

The covariance matrix is

\[
(\Sigma^{-1} + (\tau \Sigma)^{-1})^{-1}
\]

Hence, the Black-Litterman recommendation is consistent with the Bayes update of the prior over the mean excess returns in light of the realized average excess returns on the market.

### 2.9 Curve Decolletage

Consider two independent “competing” views on the excess market returns

\[ \vec{r}_e \sim \mathcal{N}(\mu_{BL}, \Sigma) \]

and

\[ \vec{r}_e \sim \mathcal{N}(\hat{\mu}, \tau \Sigma) \]

A special feature of the multivariate normal random variable \( Z \) is that its density function depends only on the (Euclidian) length of its realization \( z \).

Formally, let the \( k \)-dimensional random vector be

\[ Z \sim \mathcal{N}(\mu, \Sigma) \]

then

\[ \tilde{Z} \equiv \Sigma(Z - \mu) \sim \mathcal{N}(0, I) \]

and so the points where the density takes the same value can be described by the ellipse

\[ \tilde{z} \cdot \tilde{z} = (z - \mu)'\Sigma^{-1}(z - \mu) = \tilde{d} \quad (5) \]

where \( \tilde{d} \in \mathbb{R}_+ \) denotes the (transformation) of a particular density value.

The curves defined by equation (5) can be labeled as iso-likelihood ellipses.
**Remark:** More generally there is a class of density functions that possesses this feature, i.e.

\[ ∃ g : \mathbb{R}_+ \mapsto \mathbb{R}_+ \text{ and } c ≥ 0, \text{ s.t. the density } f \text{ of } Z \text{ has the form } f(z) = cg(z \cdot z) \]

This property is called **spherical symmetry** (see p. 81 in Leamer (1978) [5]).

In our specific example, we can use the pair \((\tilde{d}_1, \tilde{d}_2)\) as being two “likelihood” values for which the corresponding iso-likelihood ellipses in the excess return space are given by

\[ (\tilde{r}_e - \mu_{BL})^{\prime} \Sigma^{-1} (\tilde{r}_e - \mu_{BL}) = \tilde{d}_1 \]
\[ (\tilde{r}_e - \hat{\mu})^{\prime} (\tau \Sigma)^{-1} (\tilde{r}_e - \hat{\mu}) = \tilde{d}_2 \]

Notice that for particular \(\tilde{d}_1\) and \(\tilde{d}_2\) values the two ellipses have a tangency point.

These tangency points, indexed by the pairs \((\tilde{d}_1, \tilde{d}_2)\), characterize points \(\tilde{r}_e\) from which there exists no deviation where one can increase the likelihood of one view without decreasing the likelihood of the other view.

The pairs \((\tilde{d}_1, \tilde{d}_2)\) for which there is such a point outlines a curve in the excess return space. This curve is reminiscent of the Pareto curve in an Edgeworth-box setting.

Dickey (1975) [2] calls it a **curve decolletage**.

Leamer (1978) [5] calls it an **information contract curve** and describes it by the following program: maximize the likelihood of one view, say the Black-Litterman recommendation while keeping the likelihood of the other view at least at a prespecified constant \(\tilde{d}_2\)

\[ \tilde{d}_1(\tilde{d}_2) \equiv \max_{\tilde{r}_e} (\tilde{r}_e - \mu_{BL})^{\prime} \Sigma^{-1} (\tilde{r}_e - \mu_{BL}) \]
\[ \text{subject to } (\tilde{r}_e - \hat{\mu})^{\prime} (\tau \Sigma)^{-1} (\tilde{r}_e - \hat{\mu}) ≥ \tilde{d}_2 \]

Denoting the multiplier on the constraint by \(\lambda\), the first-order condition is

\[ 2(\tilde{r}_e - \mu_{BL})^{\prime} \Sigma^{-1} + \lambda 2(\tilde{r}_e - \hat{\mu})^{\prime} (\tau \Sigma)^{-1} = 0 \]

which defines the **information contract curve** between \(\mu_{BL}\) and \(\hat{\mu}\)

\[ \tilde{r}_e = (\Sigma^{-1} + \lambda (\tau \Sigma)^{-1})^{-1}(\Sigma^{-1} \mu_{BL} + \lambda (\tau \Sigma)^{-1} \hat{\mu}) \tag{6} \]

Note that if \(\lambda = 1\), (6) is equivalent with (4) and it identifies one point on the information contract curve.

Furthermore, because \(\lambda\) is a function of the minimum likelihood \(\tilde{d}_2\) on the RHS of the constraint, by varying \(\tilde{d}_2\) (or \(\lambda\)), we can trace out the whole curve as the figure below illustrates.

In [5]: np.random.seed(1987102)

\[
\begin{align*}
\text{N} & = 2 & \# \text{Number of assets} \\
\text{T} & = 200 & \# \text{Sample size} \\
\tau & = 0.8
\end{align*}
\]
# Random market portfolio (sum is normalized to 1)

\[
\begin{align*}
w_m &= np.random.rand(N) \\
w_m &= w_m / (w_m \cdot \text{sum}())
\end{align*}
\]

\[
\begin{align*}
\mu &= (np.random.randn(N) + 5) / 100 \\
S &= np.random.randn(N, N) \\
V &= S \otimes S^T \\
\Sigma &= V * (w_m \otimes \mu)^*2 / (w_m \otimes V \otimes w_m)
\end{align*}
\]

\[
\begin{align*}
excess\_return &= \text{stat.mutivariate\_normal}(\mu, \Sigma) \\
sample &= excess\_return.rvs(T)
\end{align*}
\]

\[
\begin{align*}
\mu\_est &= sample.mean(0).reshape(N, 1) \\
\Sigma\_est &= np.cov(sample.T)
\end{align*}
\]

\[
\begin{align*}
\sigma_m &= w_m \otimes \Sigma\_est \otimes w_m \\
d\_m &= (w_m \otimes \mu\_est) / \sigma_m \\
\mu\_m &= (d\_m * \Sigma\_est \otimes w_m).reshape(N, 1)
\end{align*}
\]

\[
\begin{align*}
N\_r1, N\_r2 &= 100, 100 \\
r1 &= np.linspace(-.04, .1, N\_r1) \\
r2 &= np.linspace(-.02, .15, N\_r2)
\end{align*}
\]

\[
\begin{align*}
\lambda\_grid &= np.linspace(.001, 20, 100) \\
curve &= np.asarray([black\_litterman(\lambda, \mu\_m, \mu\_est, \Sigma\_est, \\
\tau * \Sigma\_est).flatten() for \lambda in \lambda\_grid])
\end{align*}
\]

\[
\begin{align*}
\lambda\_slider &= \text{FloatSlider(min=.1, max=7, step=.5, value=1)}
\end{align*}
\]

\[
\begin{align*}
def decolletage(\lambda): \\
\text{dist\_r\_BL} &= \text{stat.mutivariate\_normal}(\mu\_m\_squeeze(), \Sigma\_est) \\
\text{dist\_r\_hat} &= \text{stat.mutivariate\_normal}(\mu\_est\_squeeze(), \tau * \Sigma\_est)
\end{align*}
\]

\[
\begin{align*}
X, Y &= np.meshgrid(r1, r2) \\
Z\_BL &= np.zeros((N\_r1, N\_r2)) \\
Z\_hat &= np.zeros((N\_r1, N\_r2))
\end{align*}
\]

\[
\begin{align*}
\text{for i in range(N\_r1):} \\
\text{for j in range(N\_r2):} \\
\text{Z\_BL[i, j] = dist\_r\_BL.pdf(np.hstack([X[i, j], Y[i, j]]))} \\
\text{Z\_hat[i, j] = dist\_r\_hat.pdf(np.hstack([X[i, j], Y[i, j]]))}
\end{align*}
\]

\[
\begin{align*}
\mu\_tilde &= \text{black\_litterman}(\lambda, \mu\_m, \mu\_est, \Sigma\_est, \\
\tau * \Sigma\_est).flatten() \\
\text{fig, ax} &= \text{plt.subplots(figsize=(10, 6))} \\
\text{ax.contourf(X, Y, Z\_hat, cmap='viridis', alpha=.4)} \\
\text{ax.contourf(X, Y, Z\_BL, cmap='viridis', alpha=.4)} \\
\text{ax.contour(X, Y, Z\_BL, [dist\_r\_BL.pdf(\mu\_tilde)], cmap='viridis',} \\
\text{alpha=.9)} \\
\text{ax.contour(X, Y, Z\_hat, [dist\_r\_hat.pdf(\mu\_tilde)], cmap='viridis',} \\
\text{alpha=.9)} \\
\text{ax.scatter(\mu\_est[0], \mu\_est[1])} \\
\text{ax.scatter(\mu\_m[0], \mu\_m[1])} \\
\text{ax.scatter(\mu\_tilde[0], \mu\_tilde[1], c='k', s=20*3)} \\
\text{ax.plot(curve[:, 0], curve[:, 1], c='k',)}
\end{align*}
\]
Note that the line that connects the two points $\hat{\mu}$ and $\mu_{BL}$ is linear, which comes from the fact that the covariance matrices of the two competing distributions (views) are proportional to each other.

To illustrate the fact that this is not necessarily the case, consider another example using the same parameter values, except that the “second view” constituting the constraint has covariance matrix $\tau I$ instead of $\tau \Sigma$.

This leads to the following figure, on which the curve connecting $\hat{\mu}$ and $\mu_{BL}$ are bending
for j in range(N_r2):
    Z_BL[i, j] = dist_r_BL.pdf(np.hstack([X[i, j], Y[i, j]]))
    Z_hat[i, j] = dist_r_hat.pdf(np.hstack([X[i, j], Y[i, j]]))

μ_tilde = black_litterman(λ, μ_m, μ_est, Σ_est, τ * np.eye(N)).flatten()

fig, ax = plt.subplots(figsize=(10, 6))
ax.contourf(X, Y, Z_hat, cmap='viridis', alpha=.4)
ax.contourf(X, Y, Z_BL, cmap='viridis', alpha=.4)
ax.contour(X, Y, Z_BL, [dist_r_BL.pdf(μ_tilde)], cmap='viridis', alpha=.9)
ax.contour(X, Y, Z_hat, [dist_r_hat.pdf(μ_tilde)], cmap='viridis', alpha=.9)
ax.scatter(μ_est[0], μ_est[1])
ax.scatter(μ_m[0], μ_m[1])
ax.scatter(μ_tilde[0], μ_tilde[1], c='k', s=20*3)
ax.plot(curve[:, 0], curve[:, 1], c='k')
ax.axhline(0, c='k', alpha=.8)
ax.axvline(0, c='k', alpha=.8)
ax.set_xlabel(r'Excess return on the first asset, $r_{e,1}$')
ax.set_ylabel(r'Excess return on the second asset, $r_{e,2}$')
ax.text(μ_est[0] + 0.003, μ_est[1], r'$\hat{\mu}_{e,1}$')
ax.text(μ_m[0] + 0.003, μ_m[1] + 0.005, r'$\mu_{BL}$')
plt.show()

2.10 Black-Litterman Recommendation as Regularization

First, consider the OLS regression

$$\min_\beta \|X\beta - y\|^2$$
which yields the solution

\[ \hat{\beta}_{OLS} = (X'X)^{-1}X'y \]

A common performance measure of estimators is the \textit{mean squared error (MSE)}. An estimator is “good” if its MSE is relatively small. Suppose that \( \beta_0 \) is the “true” value of the coefficient, then the MSE of the OLS estimator is

\[
\text{mse}(\hat{\beta}_{OLS}, \beta_0) := \mathbb{E} \| \hat{\beta}_{OLS} - \beta_0 \|^2 = \mathbb{E} \| \hat{\beta}_{OLS} - \mathbb{E} \hat{\beta}_{OLS} \|^2 + \mathbb{E} \| \mathbb{E} \hat{\beta}_{OLS} - \beta_0 \|^2
\]

From this decomposition, one can see that in order for the MSE to be small, both the bias and the variance terms must be small.

For example, consider the case when \( X \) is a \( T \)-vector of ones (where \( T \) is the sample size), so \( \hat{\beta}_{OLS} \) is simply the sample average, while \( \beta_0 \in \mathbb{R} \) is defined by the true mean of \( y \).

In this example the MSE is

\[
\text{mse}(\hat{\beta}_{OLS}, \beta_0) = \frac{1}{T^2} \mathbb{E} \left( \sum_{t=1}^{T} (y_t - \beta_0) \right)^2 + 0
\]

However, because there is a trade-off between the estimator’s bias and variance, there are cases when by permitting a small bias we can substantially reduce the variance so overall the MSE gets smaller.

A typical scenario when this proves to be useful is when the number of coefficients to be estimated is large relative to the sample size.

In these cases, one approach to handle the bias-variance trade-off is the so called \textit{Tikhonov regularization}.

A general form with regularization matrix \( \Gamma \) can be written as

\[
\min_{\beta} \left\{ \| X\beta - y \|^2 + \| \Gamma(\beta - \tilde{\beta}) \|^2 \right\}
\]

which yields the solution

\[ \hat{\beta}_{Reg} = (X'X + \Gamma'\Gamma)^{-1}(X'y + \Gamma'\tilde{\beta}) \]

Substituting the value of \( \hat{\beta}_{OLS} \) yields

\[ \hat{\beta}_{Reg} = (X'X + \Gamma'\Gamma)^{-1}(X'X\hat{\beta}_{OLS} + \Gamma'T\tilde{\beta}) \]

Often, the regularization matrix takes the form \( \Gamma = \lambda I \) with \( \lambda > 0 \) and \( \tilde{\beta} = 0 \).

Then the Tikhonov regularization is equivalent to what is called \textit{ridge regression} in statistics.

To illustrate how this estimator addresses the bias-variance trade-off, we compute the MSE of the ridge estimator
\[
\text{mse} (\hat{\beta}_{\text{ridge}}, \beta_0) = \frac{1}{(T + \lambda)^2} \mathbb{E} \left( \sum_{t=1}^{T} (y_t - \beta_0)^2 \right) + \left( \frac{\lambda}{T + \lambda} \right)^2 \beta_0^2
\]

The ridge regression shrinks the coefficients of the estimated vector towards zero relative to the OLS estimates thus reducing the variance term at the cost of introducing a “small” bias.

However, there is nothing special about the zero vector.

When \( \tilde{\beta} \neq 0 \) shrinkage occurs in the direction of \( \tilde{\beta} \).

Now, we can give a regularization interpretation of the Black-Litterman portfolio recommendation.

To this end, simplify first the equation (4) characterizing the Black-Litterman recommendation

\[
\tilde{\mu} = (\Sigma^{-1} + (\tau \Sigma)^{-1})^{-1} (\Sigma^{-1} \mu_{BL} + (\tau \Sigma)^{-1} \hat{\mu})
\]

\[
= (1 + \tau^{-1})^{-1} \Sigma \Sigma^{-1} (\mu_{BL} + \tau^{-1} \hat{\mu})
\]

\[
= (1 + \tau^{-1})^{-1} (\mu_{BL} + \tau^{-1} \hat{\mu})
\]

In our case, \( \hat{\mu} \) is the estimated mean excess returns of securities. This could be written as a vector autoregression where

- \( y \) is the stacked vector of observed excess returns of size \((NT \times 1) - N \) securities and \( T \) observations.
- \( X = \sqrt{T^{-1}}(I_N \otimes \iota_T) \) where \( I_N \) is the identity matrix and \( \iota_T \) is a column vector of ones.

Correspondingly, the OLS regression of \( y \) on \( X \) would yield the mean excess returns as coefficients.

With \( \Gamma = \sqrt{T^{-1}}(I_N \otimes \iota_T) \) we can write the regularized version of the mean excess return estimation

\[
\hat{\beta}_{\text{Reg}} = (X'X + \Gamma \Gamma')^{-1} (X'X \hat{\beta}_{OLS} + \Gamma \Gamma \tilde{\beta})
\]

\[
= (1 + \tau)^{-1} X'X (X'X)^{-1} (\hat{\beta}_{OLS} + \tau \tilde{\beta})
\]

\[
= (1 + \tau)^{-1} \hat{\beta}_{OLS} + \tau \tilde{\beta}
\]

Given that \( \hat{\beta}_{OLS} = \tilde{\mu} \) and \( \tilde{\beta} = \mu_{BL} \) in the Black-Litterman model, we have the following interpretation of the model’s recommendation.

The estimated (personal) view of the mean excess returns, \( \tilde{\mu} \) that would lead to extreme short-long positions are “shrunk” towards the conservative market view, \( \mu_{BL} \), that leads to the more conservative market portfolio.

So the Black-Litterman procedure results in a recommendation that is a compromise between the conservative market portfolio and the more extreme portfolio that is implied by estimated “personal” views.
2.11 Digression on A Robust Control Operator

The Black-Litterman approach is partly inspired by the econometric insight that it is easier to estimate covariances of excess returns than the means.

That is what gave Black and Litterman license to adjust investors’ perception of mean excess returns while not tampering with the covariance matrix of excess returns.

The robust control theory is another approach that also hinges on adjusting mean excess returns but not covariances.

Associated with a robust control problem is what Hansen and Sargent \[4\], \[3\] call a \(T\) operator.

Let’s define the \(T\) operator as it applies to the problem at hand.

Let \(x\) be an \(n \times 1\) Gaussian random vector with mean vector \(\mu\) and covariance matrix \(\Sigma = CC'\). This means that \(x\) can be represented as

\[ x = \mu + C\epsilon \]

where \(\epsilon \sim N(0, I)\).

Let \(\phi(\epsilon)\) denote the associated standardized Gaussian density.

Let \(m(\epsilon, \mu)\) be a likelihood ratio, meaning that it satisfies

\[ m(\epsilon, \mu) > 0 \]
\[ \int m(\epsilon, \mu)\phi(\epsilon)d\epsilon = 1 \]

That is, \(m(\epsilon, \mu)\) is a non-negative random variable with mean 1.

Multiplying \(\phi(\epsilon)\) by the likelihood ratio \(m(\epsilon, \mu)\) produces a distorted distribution for \(\epsilon\), namely

\[ \tilde{\phi}(\epsilon) = m(\epsilon, \mu)\phi(\epsilon) \]

The next concept that we need is the entropy of the distorted distribution \(\tilde{\phi}\) with respect to \(\phi\).

**Entropy** is defined as

\[ \text{ent} = \int \log m(\epsilon, \mu)m(\epsilon, \mu)\phi(\epsilon)d\epsilon \]

or

\[ \text{ent} = \int \log m(\epsilon, \mu)\tilde{\phi}(\epsilon)d\epsilon \]

That is, relative entropy is the expected value of the likelihood ratio \(m\) where the expectation is taken with respect to the twisted density \(\tilde{\phi}\).

Relative entropy is non-negative. It is a measure of the discrepancy between two probability distributions.

As such, it plays an important role in governing the behavior of statistical tests designed to discriminate one probability distribution from another.
We are ready to define the $T$ operator.

Let $V(x)$ be a value function.

Define

$$T(V(x)) = \min_{m(\epsilon, \mu)} \int m(\epsilon, \mu) [V(\mu + C\epsilon) + \theta \log m(\epsilon, \mu)] \phi(\epsilon) d\epsilon$$

$$= -\log \theta \int \exp \left( \frac{-V(\mu + C\epsilon)}{\theta} \right) \phi(\epsilon) d\epsilon$$

This asserts that $T$ is an indirect utility function for a minimization problem in which an evil agent chooses a distorted probability distribution $\tilde{\phi}$ to lower expected utility, subject to a penalty term that gets bigger the larger is relative entropy.

Here the penalty parameter

$$\theta \in [\theta, +\infty]$$

is a robustness parameter when it is $+\infty$, there is no scope for the minimizing agent to distort the distribution, so no robustness to alternative distributions is acquired. As $\theta$ is lowered, more robustness is achieved.

**Note:** The $T$ operator is sometimes called a *risk-sensitivity* operator.

We shall apply $T$ to the special case of a linear value function $w'(\tilde{r} - rf1)$ where $\tilde{r} - rf1 \sim \mathcal{N}(\mu, \Sigma)$ or $\tilde{r} - rf1 = \mu + C\epsilon$ and $\epsilon \sim \mathcal{N}(0, I)$.

The associated worst-case distribution of $\epsilon$ is Gaussian with mean $v = -\theta^{-1}C'w$ and covariance matrix $I$ (When the value function is affine, the worst-case distribution distorts the mean vector of $\epsilon$ but not the covariance matrix of $\epsilon$).

For utility function argument $w'(\tilde{r} - rf1)$

$$T(\tilde{r} - rf1) = w'\mu + \zeta - \frac{1}{2\theta}w'\Sigma w$$

and entropy is

$$\frac{v' v}{2} = \frac{1}{2\theta^2} w'Cw$$

### 2.12 A Robust Mean-variance Portfolio Model

According to criterion (1), the mean-variance portfolio choice problem chooses $w$ to maximize

$$E[w(\tilde{r} - rf1)] - \text{var}[w(\tilde{r} - rf1)]$$

which equals

$$w'\mu - \frac{\delta}{2}w'\Sigma w$$
A robust decision maker can be modeled as replacing the mean return \( E[w(\tilde{r} - r_f 1)] \) with the risk-sensitive

\[
T[w(\tilde{r} - r_f 1)] = w' \mu - \frac{1}{2\theta} w' \Sigma w
\]

that comes from replacing the mean \( \mu \) of \( \tilde{r} - r_f 1 \) with the worst-case mean

\[
\mu - \theta^{-1} \Sigma w
\]

Notice how the worst-case mean vector depends on the portfolio \( w \).

The operator \( T \) is the indirect utility function that emerges from solving a problem in which an agent who chooses probabilities does so in order to minimize the expected utility of a maximizing agent (in our case, the maximizing agent chooses portfolio weights \( w \)).

The robust version of the mean-variance portfolio choice problem is then to choose a portfolio \( w \) that maximizes

\[
T[w(\tilde{r} - r_f 1)] - \frac{\delta}{2} w' \Sigma w
\]

or

\[
w' (\mu - \theta^{-1} \Sigma w) - \frac{\delta}{2} w' \Sigma w \tag{7}
\]

The minimizer of (7) is

\[
w_{rob} = \frac{1}{\delta + \gamma} \Sigma^{-1} \mu
\]

where \( \gamma \equiv \theta^{-1} \) is sometimes called the risk-sensitivity parameter.

An increase in the risk-sensitivity parameter \( \gamma \) shrinks the portfolio weights toward zero in the same way that an increase in risk aversion does.

### 3 Appendix

We want to illustrate the “folk theorem” that with high or moderate frequency data, it is more difficult to estimate means than variances.

In order to operationalize this statement, we take two analog estimators:

- sample average: \( \bar{X}_N = \frac{1}{N} \sum_{i=1}^{N} X_i \)
- sample variance: \( S_N = \frac{1}{N-1} \sum_{i=1}^{N} (X_i - \bar{X}_N)^2 \)

to estimate the unconditional mean and unconditional variance of the random variable \( X \), respectively.

To measure the “difficulty of estimation”, we use mean squared error (MSE), that is the average squared difference between the estimator and the true value.
Assuming that the process \( \{X_i\} \) is ergodic, both analog estimators are known to converge to their true values as the sample size \( N \) goes to infinity.

More precisely for all \( \varepsilon > 0 \)

\[
\lim_{N \to \infty} P \{|\bar{X}_N - \mathbb{E}X| > \varepsilon\} = 0
\]

and

\[
\lim_{N \to \infty} P \{|S_N - \mathbb{V}X| > \varepsilon\} = 0
\]

A necessary condition for these convergence results is that the associated MSEs vanish as \( N \) goes to infinity, or in other words,

\[
\text{MSE}(\bar{X}_N, \mathbb{E}X) = o(1) \quad \text{and} \quad \text{MSE}(S_N, \mathbb{V}X) = o(1)
\]

Even if the MSEs converge to zero, the associated rates might be different. Looking at the limit of the relative MSE (as the sample size grows to infinity)

\[
\frac{\text{MSE}(S_N, \mathbb{V}X)}{\text{MSE}(\bar{X}_N, \mathbb{E}X)} = \frac{o(1)}{o(1)} \xrightarrow{N \to \infty} B
\]

can inform us about the relative (asymptotic) rates.

We will show that in general, with dependent data, the limit \( B \) depends on the sampling frequency.

In particular, we find that the rate of convergence of the variance estimator is less sensitive to increased sampling frequency than the rate of convergence of the mean estimator.

Hence, we can expect the relative asymptotic rate, \( B \), to get smaller with higher frequency data, illustrating that “it is more difficult to estimate means than variances”.

That is, we need significantly more data to obtain a given precision of the mean estimate than for our variance estimate.

3.1 A Special Case – IID Sample

We start our analysis with the benchmark case of IID data. Consider a sample of size \( N \) generated by the following IID process,

\[X_i \sim \mathcal{N}(\mu, \sigma^2)\]

Taking \( \bar{X}_N \) to estimate the mean, the MSE is

\[
\text{MSE}(\bar{X}_N, \mu) = \frac{\sigma^2}{N}
\]

Taking \( S_N \) to estimate the variance, the MSE is
MSE($S_N, \sigma^2$) = \( \frac{2\sigma^4}{N-1} \)

Both estimators are unbiased and hence the MSEs reflect the corresponding variances of the estimators.

Furthermore, both MSEs are \( o(1) \) with a (multiplicative) factor of difference in their rates of convergence:

\[
\frac{\text{MSE}(S_N, \sigma^2)}{\text{MSE}(\hat{X}_N, \mu)} = \frac{N2\sigma^2}{N-1} \rightarrow_{N \to \infty} 2\sigma^2
\]

We are interested in how this (asymptotic) relative rate of convergence changes as increasing sampling frequency puts dependence into the data.

### 3.2 Dependence and Sampling Frequency

To investigate how sampling frequency affects relative rates of convergence, we assume that the data are generated by a mean-reverting continuous time process of the form

\[
dX_t = -\kappa(X_t - \mu)dt + \sigma dW_t
\]

where \( \mu \) is the unconditional mean, \( \kappa > 0 \) is a persistence parameter, and \( \{W_t\} \) is a standardized Brownian motion.

Observations arising from this system in particular discrete periods \( \mathcal{T}(h) \equiv \{nh : n \in \mathbb{Z}\} \) with \( h > 0 \) can be described by the following process

\[X_{t+1} = (1 - \exp(-\kappa h))\mu + \exp(-\kappa h)X_t + \epsilon_{t,h}\]

where

\[\epsilon_{t,h} \sim \mathcal{N}(0, \Sigma_h) \quad \text{with} \quad \Sigma_h = \frac{\sigma^2(1 - \exp(-2\kappa h))}{2\kappa}\]

We call \( h \) the frequency parameter, whereas \( n \) represents the number of lags between observations.

Hence, the effective distance between two observations \( X_t \) and \( X_{t+n} \) in the discrete time notation is equal to \( h \cdot n \) in terms of the underlying continuous time process.

Straightforward calculations show that the autocorrelation function for the stochastic process \( \{X_t\}_{t \in \mathcal{T}(h)} \) is

\[\Gamma_h(n) \equiv \text{corr}(X_{t+hn}, X_t) = \exp(-\kappa hn)\]

and the auto-covariance function is

\[\gamma_h(n) \equiv \text{cov}(X_{t+hn}, X_t) = \frac{\exp(-\kappa hn)\sigma^2}{2\kappa}\].
It follows that if $n = 0$, the unconditional variance is given by $\gamma_h(0) = \frac{\sigma^2}{2\kappa}$ irrespective of the sampling frequency.

The following figure illustrates how the dependence between the observations is related to the sampling frequency:

- For any given $h$, the autocorrelation converges to zero as we increase the distance -- $n$ -- between the observations. This represents the “weak dependence” of the $X$ process.
- Moreover, for a fixed lag length, $n$, the dependence vanishes as the sampling frequency goes to infinity. In fact, letting $h$ go to $\infty$ gives back the case of IID data.

In [7]:
\[
\begin{align*}
\mu &= .0 \\
\sigma &= .5 \\
\text{var}_\text{uncond} &= \sigma^2 / (2 * x)
\end{align*}
\]

\[
\begin{align*}
n_{\text{grid}} &= \text{np.linspace}(0, 40, 100) \\
\text{autocorr}_h1 &= \text{np.exp}(-x * n_{\text{grid}} * 1) \\
\text{autocorr}_h2 &= \text{np.exp}(-x * n_{\text{grid}} * 2) \\
\text{autocorr}_h5 &= \text{np.exp}(-x * n_{\text{grid}} * 5) \\
\text{autocorr}_h1000 &= \text{np.exp}(-x * n_{\text{grid}} * 1e8)
\end{align*}
\]

\[
\begin{align*}
\text{fig}, \text{ax} &= \text{plt.subplots(figsize=(8, 4))} \\
\text{ax.plot}(n_{\text{grid}}, \text{autocorr}_h1, \text{label}='h=1'$, \text{c}=\text{'darkblue'}, \text{lw}=2) \\
\text{ax.plot}(n_{\text{grid}}, \text{autocorr}_h2, \text{label}='h=2'$, \text{c}=\text{'darkred'}, \text{lw}=2) \\
\text{ax.plot}(n_{\text{grid}}, \text{autocorr}_h5, \text{label}='h=5'$, \text{c}=\text{'orange'}, \text{lw}=2) \\
\text{ax.plot}(n_{\text{grid}}, \text{autocorr}_h1000, \text{label}='h=\infty'$, \text{c}=\text{'darkgreen'}, \text{lw}=2) \\
\text{ax.legend}() \\
\text{ax.grid}() \\
\text{ax.set}(\text{title}='Autocorrelation functions, $\Gamma_h(n)$', \\
\text{xlabel}='Lags between observations, $n$')
\]

\[
\text{plt.show}()
\]
3.3 Frequency and the Mean Estimator

Consider again the AR(1) process generated by discrete sampling with frequency $h$. Assume that we have a sample of size $N$ and we would like to estimate the unconditional mean – in our case the true mean is $\mu$.

Again, the sample average is an unbiased estimator of the unconditional mean

$$
\mathbb{E}[\bar{X}_N] = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}[X_i] = \mathbb{E}[X_0] = \mu
$$

The variance of the sample mean is given by

$$
\mathbb{V}(\bar{X}_N) = \mathbb{V}\left(\frac{1}{N} \sum_{i=1}^{N} X_i\right) = \frac{1}{N^2} \left( \sum_{i=1}^{N} \mathbb{V}(X_i) + 2 \sum_{i=1}^{N-1} \sum_{s=i+1}^{N} \text{cov}(X_i,X_s) \right)
$$

$$
= \frac{1}{N^2} \left( N \gamma(0) + 2 \sum_{i=1}^{N-1} i \cdot \gamma(h \cdot (N - i)) \right)
$$

$$
= \frac{1}{N^2} \left( N \frac{\sigma^2}{2\kappa} + 2 \sum_{i=1}^{N-1} i \cdot \exp(-\kappa h (N - i)) \frac{\sigma^2}{2\kappa} \right)
$$

It is explicit in the above equation that time dependence in the data inflates the variance of the mean estimator through the covariance terms. Moreover, as we can see, a higher sampling frequency—smaller $h$—makes all the covariance terms larger, everything else being fixed. This implies a relatively slower rate of convergence of the sample average for high-frequency data.

Intuitively, the stronger dependence across observations for high-frequency data reduces the “information content” of each observation relative to the IID case.

We can upper bound the variance term in the following way

$$
\mathbb{V}(\bar{X}_N) = \frac{1}{N^2} \left( N \sigma^2 + 2 \sum_{i=1}^{N-1} i \cdot \exp(-\kappa h (N - i)) \sigma^2 \right)
$$

$$
\leq \frac{\sigma^2}{2\kappa N} \left( 1 + 2 \sum_{i=1}^{N-1} \exp(-\kappa h (i)) \right)
$$

$$
= \frac{\sigma^2}{2\kappa N} \left( 1 + 2 \frac{1 - \exp(-\kappa h)^{N-1}}{1 - \exp(-\kappa h)} \right)
$$

Asymptotically the $\exp(-\kappa h)^{N-1}$ vanishes and the dependence in the data inflates the benchmark IID variance by a factor of

$$
\left( 1 + 2 \frac{1}{1 - \exp(-\kappa h)} \right)
$$

This long run factor is larger the higher is the frequency (the smaller is $h$).
Therefore, we expect the asymptotic relative MSEs, \(B\), to change with time-dependent data. We just saw that the mean estimator’s rate is roughly changing by a factor of

\[
\frac{1}{1 - \exp(-\kappa h)} \left(1 + 2 \frac{1}{1 - \exp(-\kappa h)} \right)
\]

Unfortunately, the variance estimator’s MSE is harder to derive.

Nonetheless, we can approximate it by using (large sample) simulations, thus getting an idea about how the asymptotic relative MSEs changes in the sampling frequency \(h\) relative to the IID case that we compute in closed form.

In [8]:
```python
def sample_generator(h, N, M):
    \phi = (1 - \exp(-x * h)) \cdot \mu
    \rho = \exp(-x * h)
    s = \sigma**2 \cdot (1 - \exp(-2 * x * h)) / (2 * x)
    mean_uncond = \mu
    std_uncond = \sigma**2 / (2 * x)
    \varepsilon\_path = \text{stat.norm}(0, \sqrt{s}).\text{rvs}((M, N))
    y\_path = \text{np.zeros}((M, N + 1))
    y\_path[:, 0] = \text{stat.norm}(mean_uncond, std_uncond).\text{rvs}(M)
    for i in range(N):
        y\_path[:, i + 1] = \phi + \rho \cdot y\_path[:, i] + \varepsilon\_path[:, i]
    return y\_path
```

In [9]:
```python
# Generate large sample for different frequencies
N\_app, M\_app = 1000, 30000  # Sample size, number of simulations
h\_grid = \text{np.linspace}(.1, 80, 30)
var\_est\_store = []
mean\_est\_store = []
labels = []

for h in h\_grid:
    sample = sample_generator(h, N\_app, M\_app)
    mean\_est\_store.append(np.mean(sample, 1))
    var\_est\_store.append(np.var(sample, 1))

del labels
var\_est\_store = \text{np.array}(var\_est\_store)
mean\_est\_store = \text{np.array}(mean\_est\_store)

# Save mse of estimators
mse\_mean = np.var(mean\_est\_store, 1) + (np.mean(mean\_est\_store, 1) - \mu)**2
mse\_var = np.var(var\_est\_store, 1) \\ 
\* (np.mean(var\_est\_store, 1) - var\_uncond)**2
benchmark\_rate = 2 * var\_uncond  # IID case

# Relative MSE for large samples
rate\_h = mse\_var / mse\_mean
```

23
The above figure illustrates the relationship between the asymptotic relative MSEs and the sampling frequency

- We can see that with low-frequency data – large values of $h$ – the ratio of asymptotic rates approaches the IID case.
- As $h$ gets smaller – the higher the frequency – the relative performance of the variance estimator is better in the sense that the ratio of asymptotic rates gets smaller. That is, as the time dependence gets more pronounced, the rate of convergence of the mean estimator’s MSE deteriorates more than that of the variance estimator.

References


