October 17, 2023

Regularization of Covariance Matrices and Returns for the Purposes of Mean-Variance Optimization

The first part of this document provides a potential approach for regularizing the potentially low quality covariance matrices that could be fed into the optimizer. The approach generally follows three steps, the latter two of which can be skipped in some circumstances:

1. (**Raw Covariance**) Load or estimate a raw covariance matrix, which may not be positive definite or even positive semi-definite. This is discussed in Section
2. (**Sample Covariance**) If necessary, use the algorithm from Higham (2002) to identify the nearest positive semi-definite matrix.
3. (**Regularized Covarianc**e) If the matrix is not positive definite or additional robustness is desired, shrink the matrix towards a positive definite target. The shrinkage can either be arbitrary or calibrated to an optimal amount. The calibration can take place either via bootstrapping or the standard Ledoit-Wolf (2004) optimal shrinkage calculation.

The second step is strictly unnecessary if the matrix is at least positive semi-definite. However, the notoriously unstable estimation of historical covariance matrixes generally indicate the regularization of the third step in many circumstances.

Section 1 discusses loading the raw covariance matrix. Section 2 describes the Modified Repeating Projection technique to force the covariance to be positive semi-definite. Section 3 considers the Bootstrapping and Ledoit-Wolf shrinkage techniques.

Several other topics are covered related to optimization outside of regularizing the covariance matrix. Section 4 describes return shrinkage using a rudimentary empirical Bayesian model based on the James-Stein (1956) estimator. The James-Stein estimator itself is implemented in Section 6 in the appendix. Finally Section 5 provides a simple algorithm for handling structured investments.

Code prototyping the techniques described in Sections 1, 2, 4, and 6, and the Ledoit-Wolf optimal shrinkage technique of Section 3 is availble on GitHub at darkbluefactore/emf_regularization. At the same time, the production prototype contains both implementations of the Bootstrapping approach from Section 3 along with code that involes the actual optimization.

1 Reading the covariance matrix

The first step is to read in the raw covariance matrix. For a CMA-based approach, the raw covariance matrix is generally provided directly. The below discusses three reasonable alternatives for historical returns.

1.1 Completely shared overlap

If the asset with the shortest track record (T_{min}) is significantly larger than the number of assets, than this method is often preferred as it generally produces an asymptotically correct positive definite covariance matrix. However outside of ideal circumstances, the method squanders substantial quantities of information. For instance, a portfolio consisting of the S&P 500, the Barclays Aggregate Bond Index, and a secondaries private equity fund that launched two years prior would only provide 8 quarterly data points. While this is sufficient for the purposes of creating a positive definite matrix, the shear quantity of information lost from subtotaling returns to a quarterly frequency and cutting off the track record at two years suggests that this approach is sub-optimal in finite samples. In cases where $N > T$, the covariance matrix is singular and still must be regularized prior to unrestricted optimization.

1.2 Pairwise overlap

This approach seems to provide a natural solution to the problem of track records with differing length. Specifically, for any two assets, compute the pairwise covariance by first aggregating the track records to the lowest frequency:

$$\hat{\sigma}_{ij} = \frac{s_{ij}}{T_{ij}} \sum_{t \in 1:T_{ij}} \left(x_{it} - \overline{x}_i\right)\left(x_{jt} - \overline{x}_j\right)$$

where x_{it} is the return of asset i at time t, T_{ij} is the number of data points available during the overlapping period with data measured at a common frequency, and s_{ij} is a scaling factor that adjusts all covariances to a common frequency (typically monthly, quarterly, or annual).

The property that the resulting sample covariance matrix does not need to be, and is often not, positive semi-definite presents a significant disadvantage. Furthermore, despite the improved amount of information used, each pairwise calculation still loses significant information when the track records are of differing lengths and frequencies.

The pairwise approach can be modified to improve informational efficiency. One option is to compute the pair-wise correlation matrix but compute the variances using the entire track record. If using this approach, the covariance matrix is calculated as follows:

$$\begin{aligned}
\rho_{ij} &= \frac{\frac{1}{T_{ij}} \sum_{t \in 1:T_{ij}} \left(x_{it} - \overline{x}_i\right)\left(x_{jt} - \overline{x}_j\right)}{\left(\frac{1}{T_{ij}} \sum \left(x_{it} - \overline{x}_i\right)^2\right)^{1/2} \left(\frac{1}{T_{ij}} \sum \left(x_{jt} - \overline{x}_j\right)^2\right)^{1/2}} \\
\sigma_{ij} &= \rho_{ij} \sigma_{ii} \sigma_{jj}
\end{aligned}$$

where σ_{ii} and σ_{jj} are calculated as previously described. This approach is an improves the previously described pair-wise methedology by not entirely ignoring portions of the track record for assets with longer history. However, greater informational efficiency is still possible, as described in the subsequent section.

1.3 Informationally efficient pairwise overlap

In population moments, the pairwise covariance matrix can be written as:

$$\sigma_{ij} = s_{ij} \times \left(E\left[x_i x_j\right] - E\left[x_i\right] E\left[x_j\right] \right)$$

Importantly, the second term does not require a common overlapping period in a finite sample. Therefore the following plug-in estimator can be used:

$$\hat{\sigma}_{ij} = s_{ij} \times \left(\left(\frac{1}{T_{ij}} \sum x_i x_j \right) - \left(\frac{1}{T_i} \sum x_i \right) \left(\frac{1}{T_j} \sum x_j \right) \right)$$

As with the standard pairwise approach, this variant does not generally result in a positive semi-definite covariance matrix.

1.4 Additional potential enhancements

- A Bartlett or Cochrane kernel could account for autocorrelation.
- Imposing a factor structure, in the manner of G3 or otherwise, could improve pairwise correlation estimates by reducing the proportion of the covariance that is calculated in a pairwise manner.

2 Modified Repeated Projection Method

The methodology for creating the positive-semidefinite matrix, follows Higham 2002 with some modifications. It follows the usual methodology of first ensuring that the correlation matrix is positive semi-definite, then applying an additional regularization to impose a unique solution.

Start by computing the correlation matrix ρ. Higham proves a fixed point theorem for an algorithm that identifies the solution to the problem

$$\begin{aligned} &\min_{\tilde{\rho}} ||\rho - \overline{\rho}|| \\ &s.t. \\ &\overline{\rho} \geq 0 \text{ (is postive-semi-definite)} \\ &1 = diag\left(\overline{\rho}\right) \end{aligned}$$

s.t. both ρ and $\overline{\rho}$ are correlation matrices. The constraints imply that $\tilde{\rho}$is a correlation matrix. For a $K \times K$ correlation matrix, Higham's algorithm is as follows:

1. Initialize $dS = 0_{K \times K}$ and $Y = \rho$
2. Let $R = Y - dS$. This applies the Dykstra correction which is necessary for the convergence proofs in the Higham paper.
3. Compute $X = \left(V^R\right) \left\{ \max\left(\lambda^R, 0\right) \right\} \left(V^R\right)'$ where λ^R and V^R are the eigenvalue and eigenvector decompositions of R.

4. Compute $dS = X - R$

5. Compute $Y_{ij} = \begin{cases} X_{i,j} & i \neq j \\ 1.0 & i = j \end{cases}$ for all $i, j \in 1 : K$

6. Repeat steps $2 - 5$ many times (from a practical standpoint, this is until$||X - Y|| < tol$)

7. The matrix $\overline{\rho} = Y$ now approximately solves the problem above.

This does not guarantee a unique solution to a mean-variance portfolio optimization, as $\overline{\rho}$ may not be positive definite. If the singular optimization approach is employed (see the appendix), a unique solution may not be needed. Otherwise, shrinkage represents a reasonable way to regularize the correlation matrix.

3 Robustness and uniqueness

A unique solution to the optimization problem generally entails a positive definite matrix.

3.1 A minimally unique solution via shrinkage

To apply a minimal degree of regularization to the problem, modify the standard algorithm through the following transformation

$$\tilde{\rho} = (1 - \alpha)\,\overline{\rho} + \alpha \rho^S$$

where ρ^S is a positive definite shrinkage target with a unit diagonal.Two reasonable options for selecting α include picking a minimum eigenvalue threshold and cross-validation/bootstrapping.

In the minimum eigenvalue approach, α is selected such that

$$\begin{aligned} c \leq & \lambda \\ V \Lambda V' \equiv & \tilde{\rho} \end{aligned}$$

where λ is the vector of eigenvalues, Λ is the corresponding diagonal matrix, V is the matrix of eigenvectors, and c is the minimum allowable eigenvalue.

In the special case where $\rho^S = I_K$, the minimum shrinkage α is given by:

$$\alpha_{min} = \begin{cases} \frac{c - \min(\lambda)}{1 - \min(\lambda)} & 0 \leq \min(\lambda) < c \\ 0 & \min(\lambda) \geq c \end{cases}$$

for all $c \in [0, 1]$. This follows from well known properties of eigenvalues in the context of symmetric

and diagonal matrices:

$$\begin{aligned} eig\left((1-\alpha)A\right) &= (1-\alpha)\lambda \text{ (follows from scaling)} \\ \implies eig\left((1-\alpha)A+\alpha I\right) &= (1-\alpha)\lambda+\alpha \text{ (follows from } eig\left(M+\kappa I\right)=eig\left(M\right)+\kappa) \\ \implies c &= (1-\alpha)\min\left(\lambda\right)+\alpha \\ \alpha &= \frac{c-\min\left(\lambda\right)}{1-\min\left(\lambda\right)} \end{aligned}$$

The previous result can be partially generalized by Weyl's inequality:

$$\min\left(\tilde{\lambda}\right) \geq (1-\alpha)\min\left(\lambda\right)+\alpha\min\left(\lambda^S\right)$$

where λ^S is the eigenvalue vector for the shrinkage target. Thus for minimum bound c:

$$\begin{aligned} c &= (1-\gamma)\min\left(\lambda\right)+\alpha\min\left(\lambda^S\right) \leq \min\left(\tilde{\lambda}\right) \\ \gamma^S &= \frac{c-\min\left(\lambda\right)}{\min\left(\lambda^S\right)-\min\left(\lambda\right)} \end{aligned}$$

using the above as a mixing target will guarantee a regularized matrix where $\min\left(\lambda\right) \geq c$. This does not guarantee that γ^S is the smallest acceptable value of γ, just that γ is acceptable with respect to a constraint on the minimum eigenvalue.

The identity matrix is an aggressive shrinkage target in that it assumes greater benefits from diversification than are likely. If γ^S is selected as above, this effect is unlikely to be material for a suitably small eigenvalue minimum c. However, consider using a slightly more sophisticated target if γ^S is large enough for the shrinkage target to materially impact the results. This could include matrices derived from single factor models (recommended in Ledoit-Wolf 2003) or a constant correlation restriction (recommended in Ledoit-Wolf 2004).

With a single factor model, all assets would be modeled as $r_t = \beta f_t + \varepsilon_t$. If f_t is represented by one of the considered assets (say equity markets), $\beta_i = \frac{\rho_{if}\sigma_i}{\sigma_f}$, which further implies $\sigma_\varepsilon^2 = \sigma_i^2 - \beta^2\sigma_f^2$. Then $\rho_{ij} = \frac{\beta_i\beta_j\sigma_f^2}{\sigma_i\sigma_j}$ for $i \neq j$.

3.2 Bootstrapping

As with any mean-variance approach, out of sample efficacy is a concern. Regularization can help, but determining the optimal amount of regularization requires specialized techniques. Two reasonable approaches are bootstrapping (either parametric or non-parametric) and cross-validation. If the covariance matrix and expected return moments are supplied directly, cross-validation and standard non-parametric bootstrapping are not applicable. Parametric bootstrapping represents a reasonable alternative.

The parametric bootstrap is widely employed in applied statistics and finance (See for instance Wasserman 2003 for a textbook implementation or Tepper 2020 for an example in a finance context).

Suppose the expected returns and covariance matrices are sampled from T_{IS} data points. T_{IS} can be either an assumption, or if historical data is supplied, the number of data points utilized. Let B

be the number of bootstrap iterations, and θ be the regularization parameter, either the number of principle components targeted for retention. (For research purposes, examining T_{IS} in the range of 24-60 months seems to be a reasonable starting point). Then the parametric bootstrap operates as follows:

1. Calculate or estimate valid expected returns $\hat{\mu}$, variances $\hat{\sigma}$, and correlations $\hat{\rho}$ from the provided inputs. If the provided covariance matrix is not positive definite, use the previously mentioned techniques to make it valid.

2. Simulate B track-records of length T_{IS}, using the moments previously calculated. These track records need not be normal, though the normal variant is more tractable for analytical analysis and allows for fast approximations. Several reasonable options:

   (a) Multivariate normal (see remarks for some approximations that could be applied).

   (b) A multivariate t distribution with scale parameter $\Sigma^{MVT} = \frac{\nu-2}{\nu}\Sigma$ and dof selected where $\frac{6}{\nu-4} = kurt$, where $kurt$ is some reasonable number (say 2, roughly the excess kurtosis of the Wilshire 5000, implying $\nu = 7$).

   (c) A VAR. Assumptions would need to be made on the autocorrelation.

3. Compute the sample moments $\hat{\mu}_b$, $\hat{\sigma}_b$ and $\hat{\Sigma}_b$ over the T_{IS} period, implying B estimates of each input.

4. Optimize the portfolio for each set of inputs given regularization parameter θ, and the sample moments from the simulated track record, and score the objective function for the portfolio according to the original sample moments $\hat{\mu}$ and $\hat{\Sigma}$. Average the B resulting scores.

5. Repeat steps 3 and 4 until optimal θ is identified.

Note that the scoring function is not necessarily the objective function when the constraints are measured relative to the population moments, which are unobservable. While constraints on the weights will never be violated as the weights are perfectly measured, imprecisely measured moments may lead to violation of the constraints. This becomes an issue when maximizing return given a volatility cap or minimizing volatility given a return target.

To account for this, impose a second set of constraints for each path. Specifically, include the constraints implied by the originally estimated moments $g(w) \leq 0$, in addition to the estimated moments from each path $g_b(w) \leq 0$. For a minimum return, this would mean adding a constraint $w'\hat{\mu} \geq r_{min}$ in addition to $w'\mu_b \geq r_{min}$. For a volatility cap, the additional constraint would be $w'\hat{\Sigma}w \leq \sigma^2_{max}$ in addition to $w'\hat{\Sigma}_b w \leq \sigma^2_{max}$. This approach scores the portfoliosin accordance to the objective function while ensuring that non-feasible solutions are avoided.

With respect to the regularization parameter, two choices seem reasonable. First, θ could correspond to the number of principle components kept. . To compute the principle component regularized matrix, calculate the truncated singular value decomposition of the correlation matrix:

$$\overline{\rho} = \overline{USV}'$$

where $\overline{S}$ is a matrix with the first θ principle components and zero elsewhere and $\overline{U}$ and $\overline{V}'$ are the left and right singular value matrices. Note that $\overline{\rho}$ is NOT a valid correlation matrix, and as a result the MRPM method needs to be run on the truncated matrix again. Despite this shortcoming, this methodology may be fast given that only K scenarios need be tested.

Second, θ could correspond to the shrinkage parameter $\gamma \in [0, 1]$. Shrinkage in this manner is a well-accepted approach to regularizing covariance matrices (See Ledoit-Wolfe 2003 and 2004). It also has an advantage in always corresponding to a unique positive definite correlation matrix. If this approach is used, the constant correlation shrinkage target is more conservative and better grounded than the identity matrix, with the choice having greater meaning given that the shrinkage amount is no longer de minimis. The primary downside of the approach is θ is now a continuous variable.

3.2.1 Direct modeling of the covariance matrix

In some circumstances, a combination of assumptions and approximations allows for direct sampling from the distribution of sample covariance matrices. This would allow for application of the bootrapping procedure without simulating the track records by drawing the covariance matrix from a specified probability distribution. In the case of a multi-variate normal, the distribution of a sample covariance matrix given normally distributed variables given the true mean μ_r is given by:

$$\Sigma_W \sim W_K(\Sigma, T)$$

Note that the above is an approximation as the true mean is not known. However, the sample mean is distributed as:

$$\overline{r} \sim N\left(\mu_r, \frac{1}{T}\Sigma\right)$$

Then the distribution of the sample covariance can be derived as follows:

$$\begin{aligned}
\Sigma_W =& \frac{1}{T}\sum_t r_t r_t' - \mu_r \mu_r' \\
\hat{\Sigma} =& \frac{1}{T}\sum_t r_t r_t' - \overline{r}\overline{r}' \\
\implies \hat{\Sigma} =& \Sigma_W - \overline{r}\overline{r}' + \mu_r \mu_r'
\end{aligned}$$

The above is still an approximation as it relies on the independence of Σ_W and $\overline{r}$.

For a more general approximation with respect to the sample covariance matrix, apply the central limit theorem:

$$\begin{aligned}
\sqrt{T}\left(\hat{\sigma}_{ij} - \sigma_{ij}\right) \rightsquigarrow& N\left(0, var\left[\sigma_{ij}\right]\right) \\
var\left[\sigma_{ij}\right] \equiv& var\left[\frac{1}{T}\sum_t \left(r_{it} - \mu_i\right)\left(r_{jt} - \mu_j\right)\right] \\
=& E\left[\left(r_i - \mu_i\right)^2\left(r_j - \mu_j\right)^2\right] - \sigma_{ij}^2
\end{aligned}$$

Application of the above relies on knowing the co-kurtosis of the two series, and relies on large T asymptotics.

If r is multivariate normal, then the distribution follows a bi-variate normal and the co-kurtosis is well-known:

$$E\left[(r_i - \mu_i)^2 (r_j - \mu_j)^2\right] = \sigma_i^2 \sigma_j^2 + 2\sigma_{ij}^2$$
$$\sqrt{T}\left(\hat{\sigma}_{ij} - \sigma_{ij}\right) \rightsquigarrow N\left(0, \sigma_i^2 \sigma_j^2 + \sigma_{ij}^2\right)$$

However, the Wishart approximation is likely more accurate in this scenario.

Returning to the multi-variate normal scenario, if an approximation is applied for the in-sample portion and out-of-sample portion separately, then actual track records may not need to be generated. However, the track record approach avoids any of the the above approximations.

3.3 Ledoit-Wolf Optimal Shrinkage

The Ledoit-Wolf methodology uses the realized track records to estimate the shrinkage coefficient which minimizes the expected difference between the population covariance matrix and the sample covariance matrix. The method is not applicable when using CMAs. Furthermore, the input matrix must be at least positive semi-definite, meaning that the MPR method should be used if necessary before computing the optimal shrinkage. Several modifications account for the pairwise nature of the covariance estimates. Following the 2003 and 2004 papers, the below discussion applies the shrinkage to the covariance rather than the correlation matrixes:

$$\Sigma = (1 - \alpha) S + \gamma F$$

where S is a positive semi-definite "sample" covariance matrix and F is a positive definite shrinkage matrix. The positive definite result follows from the property that any convex combination of S and F (excluding S) yields a positive definite matrix.

3.3.1 Derivation of the optimal shrinkage coefficient

Since S is the plug-in estimator for $E[S_t]$, the goal is to pick α such that following problem is solved:

$$arg \min_{\alpha \in (0,1]} E\| (1 - \alpha) S + \alpha F - \Sigma \|$$

Unfortunately, the plug-in estimator for S_t may not be positive semi-definite if it is estimated pairwise, and instead reflects covariance matrix implied by the nearest positive semi-definite correlation matrix.

The risk function is given by:

$$\begin{aligned}
&E\| (1-\alpha) S + \alpha F - \Sigma \| \\
=&\sum_{i,j} V\left[(1-\alpha) S + \alpha F - \Sigma\right] + E\left[(1-\alpha) S + \alpha F - \Sigma\right]^2 \\
=&\sum_{i,j} V\left[(1-\alpha) S + \alpha F\right] + E\left[(1-\alpha) S + \alpha F - \Sigma\right]^2 \\
=&\sum_{i,j} \Bigg[(1-\alpha)^2 V[S] + \alpha^2 V(F) \\
&+ 2(1-\alpha)\alpha Cov(S,F) + \alpha^2 \left(E[F] - \Sigma\right)^2 \Bigg]
\end{aligned}$$

The above assumes that S_t is a consistent estimator of Σ (very confident this is true for the pairwise matrix but still should be checked). All moments are computed at the element level. Assuming convexity and taking FOC to solve for optimal α:

$$\begin{aligned}
0 =&\sum_{i,j} \Bigg[-(1-\alpha) V[S] + \alpha V(F) \\
&+ (1-2\alpha) Cov(S,V) + \alpha \left(E[F] - \Sigma\right)^2 \Bigg] \\
\alpha =&\frac{\sum_{i,j} \left[V[S] - Cov(S,F)\right]}{\sum_{i,j} \left[V[S] + V(F) - 2Cov(S,F) + \left(E[F] - \Sigma\right)^2\right]} \\
\implies \alpha =&\frac{\sum_{i,j} \left[V[S] - Cov(S,V)\right]}{\sum_{ij} \left[V[S-F] + \left(E[F] - \Sigma\right)^2\right]}
\end{aligned}$$

The above makes no assumptions about the track-record lengths. However to plug in finite sample analogues, let T_{ij} be the length of the overlapping portion of asset i and j, with $T_i = T_{ii}$the length of the track record of asset i. Then:

$$\hat{\alpha} = \frac{\sum_{i,j} \frac{1}{T_{ij}} V\left[\sqrt{T} s_{ij}\right] - \sum_{i,j} \frac{1}{T_{ij}} Cov\left(\sqrt{T_{ij}} s_{ij}, \sqrt{T_{ij}} f_{ij}\right)}{\sum_{ij} \left[V[S-F] + \left(E[F] - \Sigma\right)^2\right]}$$

Assume that $s_{ij} = \frac{1}{T_{ij}} \sum s_{ijt}$ and $f_{ij} = \frac{1}{T_{ij}} \sum f_{ijt}$. Then by the CLT:

$$\sqrt{T_{ij}} \left(\begin{bmatrix} s_{ij} \\ f_{ij} \end{bmatrix} - \begin{bmatrix} \sigma_{ij} \\ E[f_{ij}] \end{bmatrix} \right) \rightsquigarrow N\left(0, \begin{bmatrix} V[s_{ij}] & Cov(s_{ij}, f_{ij}) \\ Cov(s_{ij}, f_{ij}) & V[f_{ij}] \end{bmatrix}\right)$$

$$\frac{\sum_{i,j} \frac{1}{T_{ij}} V[s_{ijt}] - \sum_{i,j} \frac{1}{T_{ij}} Cov(s_{ijt}, f_{ijt})}{\sum_{ij} \left[V[S-F] + \left(E[F] - \Sigma\right)^2\right]} \rightsquigarrow \alpha$$

Similarly, by the delta method, $\nabla\left(\sigma_{ij}-E\left[f_{ij}\right]\right)=\begin{bmatrix}-E\left[f_{ij}\right]\\ \sigma_{ij}\end{bmatrix}$. Therefore:

$$\sqrt{T_{ij}}\left(\left(s_{ij}-f_{ij}\right)-\left(\sigma_{ij}-E\left[f_{ij}\right]\right)\right)\rightsquigarrow N\left(0,\delta_{ij}\right)$$

From the above, $V\left(s_{ij}-f_{ij}\right)$tends to zero at $O\left(\frac{1}{T_{ij}}\right)$ while while $\left(E\left[F\right]-\Sigma\right)^2$ remains fixed. Therefore:

$$\begin{aligned}\alpha \approx & \frac{\sum_{ij}\frac{1}{T_{ij}}\left(\xi_{ij}-\theta_{ij}\right)}{\gamma}\\ s.t.&\\ \xi_{ij} \equiv & V\left[\sqrt{T_{ij}}s_{ij}\right]\\ \theta_{ij} \equiv & Cov\left(\sqrt{T_{ij}}s_{ij},\sqrt{T_{ij}}f_{ij}\right)\\ \gamma \equiv & \sum_{ij}\left(E\left[F\right]_{ij}-\Sigma_{ij}\right)^2\end{aligned}$$

To derive estimators, first note that $\hat{q}_{ij}\equiv\frac{1}{T_{ij}}\sum_t\left(x_{it}-\overline{x}_t\right)\left(x_{jt}-\overline{x}_j\right)\xrightarrow{P}\sigma_{ij}$. If the covariance matrix is computed pairwise, $s_{it}=q_{it}$, otherwise it is the covariance calculated via the MRP method. The MRP covariance matrix can be viewed as the covariance matrix implied by the closest positive-semidefinite correlation matrix. The estimator for s_{ij} follows:

$$\sqrt{T}_{ij}\left(\hat{s}_{ij}-\sigma_{ij}\right)\rightsquigarrow N\left(0,V\left[T_{ij}s_{ij}\right]\right)$$

Estimating the variance of s_{ij} is therefore a straight-forward estimation of plug-in estimators

$$\hat{\xi}_{ij}\equiv\frac{1}{T_{ij}}\sum_{t\in1:T_{ij}}\left[\left(\hat{q}_{ijt}-\hat{s}_{ij}\right)^2\right]$$

If the MRP method is used, the accuracy as measured by the similarity between Q and S will likely increase for larger values of T. Any drop in the shrinkage coefficient form the use of the MRP method is potentially partially mitigated by using s_{ij}as a plug-in estimator.

The estimator for γ is a straight forward plug in estimation:

$$\hat{\gamma}=\sum_{ij}\left(f_{ij}-\hat{s}_{ij}\right)^2$$

The estimator for θ is less straight forward and is specific to the shrinkage target employed. The

below ONLY considers the constant correlation target. To be precise:

$$\hat{f}_{ij} = \begin{cases} \hat{s}_{ij} & i = j \\ \overline{r}\sqrt{\hat{s}_{ii}\hat{s}_{jj}} & otherwise \end{cases}$$

$$s.t.$$

$$\overline{r} = \frac{2}{(N-1)(N-2)} \sum_{i \in 2:N} \sum_{j \in 1:i-1} \frac{\hat{s}_{ij}}{\sqrt{\hat{s}_{ii}\hat{s}_{jj}}}$$

the definition of $\overline{r}$ being a simple averaging of the off-diagonal terms of the correlation matrix. Apply the delta method to the standard deviations:

$$\theta_{ij,i \neq j} = Cov\left(\sqrt{T_{ij}}\overline{r}\sqrt{s_{ii}s_{jj}}, \sqrt{T_{ij}}s_{ij}\right)$$

The following result will be useful. Suppose we want to apply the delta method to $cov\left(g\left(z\right), h\left(z\right)\right)$ where

$$\sqrt{T}\left(\hat{z} - \mu_z\right) \rightsquigarrow N\left(0, \Sigma^z_{jk}\right)$$

denotes the asymptotic covariance matrix for z. Then a first order expansion is given by:

$$\begin{aligned} E\left[g\left(z\right)h\left(z\right)\right] \approx & \left(g\left(\mu_z\right) + \sum_{k \in 1:K} \nabla g\left(\mu_z\right)\left(z - \mu_z\right)\right)\left(h\left(\mu_z\right) + \sum_{k \in 1:K} \nabla h\left(\mu_z\right)\left(z - \mu_z\right)\right) \\ = & \sum_{j,k}\left[\Sigma^z_{jk} \odot \left[\left(\nabla g\left(\mu_z\right)\nabla h\left(\mu_z\right)'\right)\right]_{jk}\right] + g\left(\mu_z\right)h\left(\mu_z\right) \\ = & tr\left(\Sigma_z\left[\nabla g\left(\mu_z\right)\nabla h\left(\mu_z\right)'\right]'\right) + g\left(\mu_z\right)h\left(\mu_z\right) \end{aligned}$$

Now apply this to θ. Since $\{T_i, T_j\} \geq T_{ij}$

$$\sqrt{T_{ij}}\left(\begin{bmatrix}\hat{s}_{ii} \\ \hat{s}_{jj} \\ \hat{s}_{ij}\end{bmatrix} - \begin{bmatrix}s_{ii} \\ s_{jj} \\ s_{ij}\end{bmatrix}\right) \rightsquigarrow N\left(0, \Sigma^s_{ii,ij}\right)$$

$$\begin{aligned} \implies Cov\left(\sqrt{T_{ij}}\overline{r}\sqrt{s_{ii}s_{jj}}, \sqrt{T_{ij}}s_{ij}\right) \approx & \overline{r}1'\left[\Sigma^s_{ii,ij} \odot \left(\begin{bmatrix}\frac{1}{2}\sqrt{\frac{\hat{s}_{jj}}{\hat{s}_{ii}}} \\ \frac{1}{2}\sqrt{\frac{\hat{s}_{ii}}{\hat{s}_{jj}}} \\ 0\end{bmatrix}\begin{bmatrix}0 & 0 & 1\end{bmatrix}\right)\right]1 \\ = & \frac{\overline{r}}{2}\left(\sqrt{\frac{\hat{s}_{jj}}{\hat{s}_{ii}}}cov\left(s_{ii}, s_{ij}\right) + \sqrt{\frac{\hat{s}_{ii}}{\hat{s}_{jj}}}cov\left(s_{jj}, s_{ij}\right)\right) \end{aligned}$$

Now replace the covariance terms with plug-in estimators and total the results:

$$\begin{aligned}
\hat{\theta}_{ij,i\neq j} =& \frac{\overline{r}}{2}\frac{1}{T_{ij}}\sum_{t\in 1:T_{ij}}\left[\sqrt{\frac{\hat{s}_{jj}}{\hat{s}_{ii}}}\left(\left(x_{it}-\overline{x}_i\right)^2-\hat{s}_{ii}\right)\left(\left(x_{it}-\overline{x}_i\right)\left(x_{jt}-\overline{x}_j\right)-\hat{s}_{ij}\right)\right. \\
&\left.+\sqrt{\frac{\hat{s}_{ii}}{\hat{s}_{jj}}}\left(\left(x_{jt}-\overline{x}_j\right)^2-\hat{s}_{jj}\right)\left(\left(x_{it}-\overline{x}_i\right)\left(x_{jt}-\overline{x}_j\right)-\hat{s}_{ij}\right)\right] \\
=& \frac{\overline{r}}{2}\frac{1}{T_{ij}}\sum_{t\in 1:T_{ij}}\left(\sqrt{\frac{\hat{s}_{jj}}{\hat{s}_{ii}}}\left(q_{iit}-\hat{s}_{ii}\right)\left(q_{ijt}-\hat{s}_{ij}\right)+\sqrt{\frac{\hat{s}_{ii}}{\hat{s}_{jj}}}\left(q_{jjt}-\hat{s}_{jj}\right)\left(q_{ijt}-\hat{s}_{ij}\right)\right) \\
\hat{\theta}_{ii} =& \hat{\xi}_{ii}
\end{aligned}$$

Finally, to complete the estimator, force $\hat{\alpha}$such that $\hat{\alpha} \in [0,1]$. Then the asymptotically optimal shrinkage is then given by:

$$\hat{\alpha} = \min\left(1, \max\left(0, \frac{\sum_{ij}\frac{1}{T_{ij}}\left(\hat{\xi}_{ij}-\hat{\theta}_{ij}\right)}{\hat{\gamma}}\right)\right)$$

Note that the above shrinkage estimator represents an average of optimal shrinkage estimators for each entry in the matrix. As a result, it may overshrink long history estimates and undershrink shorter history estimates. A more conservative approach is to inflate the shrinkage estimate to reflect the shortest history track record. This gives:

$$\hat{\alpha}_{min} = \min\left(1, \max\left(0, \frac{\frac{1}{T_{min}}\sum_{ij}\left(\hat{\xi}_{ij}-\hat{\theta}_{ij}\right)}{\hat{\gamma}}\right)\right)$$

where T_{min} is the shortest history rack record.

3.3.2 Step-by-step procedure for deriving the optimal shrinkage parameter using a constant correlation shrinkage target

1. Compute the raw pairwise covariance matrix $\hat{Q}$.
2. Estimate the sample covariance matrix $\hat{S}$:
   (a) If $\hat{Q}$ is at least positive semi-definite, $\hat{S} = \hat{Q}$
   (b) Otherwise, use the MRP method on $\hat{Q}$ to derive $\hat{S}$.

3. Compute the shrinkage target:

$$\hat{f}_{ij} = \begin{cases} \hat{s}_{ij} & i = j \\ \overline{r}\sqrt{\hat{s}_{ii}\hat{s}_{jj}} & otherwise \end{cases}$$

$$s.t.$$

$$\overline{r} = \frac{2}{(N-1)(N-2)} \sum_{i \in 2:N} \sum_{j \in 1:i-1} \frac{\hat{s}_{ij}}{\sqrt{\hat{s}_{ii}\hat{s}_{jj}}}$$

Note that it would be reasonable to replace $\hat{s}$ here with $\hat{q}$.

4. Compute the following quantities:

$$\begin{aligned} \hat{\xi}_{ij} \equiv & \frac{1}{T_{ij}} \sum_{t \in 1:T_{ij}} \left[\left(\hat{q}_{ijt} - \hat{s}_{ij} \right)^2 \right] \\ \hat{\gamma} \equiv & \sum_{ij} \left(f_{ij} - \hat{s}_{ij} \right)^2 \\ \hat{\theta}_{ij, i \neq j} = & \frac{\overline{r}}{2} \frac{1}{T_{ij}} \sum_{t \in 1:T_{ij}} \left(\sqrt{\frac{\hat{s}_{jj}}{\hat{s}_{ii}}} \left(q_{iit} - \hat{s}_{ii} \right) \left(q_{ijt} - \hat{s}_{ij} \right) + \sqrt{\frac{\hat{s}_{ii}}{\hat{s}_{jj}}} \left(q_{jjt} - \hat{s}_{jj} \right) \left(q_{ijt} - \hat{s}_{ij} \right) \right) \\ \hat{\theta}_{ii} = & \hat{\xi}_{ii} \end{aligned}$$

5. The optimal shrinkage is then given by either:

$$\hat{\alpha} = \min \left(1, \max \left(0, \frac{\sum_{ij} \frac{1}{T_{ij}} \left(\hat{\xi}_{ij} - \hat{\theta}_{ij} \right)}{\hat{\gamma}} \right) \right)$$

or

$$\hat{\alpha}_{min} = \min \left(1, \max \left(0, \frac{\frac{1}{T_{min}} \sum_{ij} \left(\hat{\xi}_{ij} - \hat{\theta}_{ij} \right)}{\hat{\gamma}} \right) \right)$$

with the top quantity potentially under-shrinking the results and the bottom quantity more likely to over-shrink the results (less bias vs less variance).

4 Simple empirical Bayes in the Stein framework

Let $\hat{\Sigma}$ be an estimator for the **means** of K assets.

$$\overline{x} \sim N(\theta, \Sigma)$$

Assume

$$\theta \sim N(\mu, a)$$

Then:

$$
\begin{aligned}
p(\theta|x) \propto & p(x|\theta)\, p(\theta) \\
= & \frac{1}{(2\pi)^K a^{K/2} \det(\Sigma)^{1/2}} \exp\left[-\frac{1}{2}(x-\theta)' \hat{\Sigma}^{-1}(x-\theta) - \frac{(\theta-\mu)'(\theta-\mu)}{2a}\right] \\
= & \frac{1}{(2\pi)^K a^{K/2} \det(\Sigma)^{1/2}} \exp\left[-\frac{\theta'\left(a\Sigma^{-1}+I\right)\theta - 2\theta'\left(a\Sigma^{-1}x+\mu\right)}{2a}\right] \exp\left(-\frac{ax'\Sigma^{-1}x+\mu'\mu}{2a}\right) \\
= & \frac{1}{(2\pi)^K a^{K/2} \det(\Sigma)^{1/2}} \exp\left[-\frac{1}{2}(\theta-\delta)'\Xi^{-1}(\theta-\delta)\right] \exp\left(-\frac{ax'\Sigma^{-1}x+\mu'\mu-a\delta'\Xi^{-1}\delta}{2a}\right) \\
= & N(\delta,\Xi;\theta) \times \frac{\det(\Xi)^{1/2}}{(2\pi)^{K/2} a^{K/2} \det(\Sigma)^{1/2}} \times \exp\left(-\frac{ax'\Sigma^{-1}x+\mu'\mu-a\delta'\Xi^{-1}\delta}{2a}\right)
\end{aligned}
$$

s.t.

$$
\begin{aligned}
\Xi = & \left[\Sigma^{-1} + \frac{1}{a} I\right]^{-1} \\
\delta = & \Xi\left[\Sigma^{-1}x + \frac{1}{a}\mu\right]
\end{aligned}
$$

Plugging in the grand mean for μ and the grand variance for a yields an empirical Bayesian estimate for the means. Taking the expectation gives the point estimate (δ). A higher grand variance implies less shrinkage. Higher variance for x yields increased shrinkage. The estimate has several advantages over the standard Stein estimator. It avoids a coordinate transform, and accounts for the relative variance of the assets. A negative of this approach is the heuristic nature of the empirical Bayesian plug-in estimators.

4.1 Estimating the grand mean and variance

The grand variance is needed in order to apply the above methodology. Consider that the sample mean can be written in the following form.

$$
\begin{aligned}
\overline{x}_i = & \theta_i + \varepsilon_i \\
\varepsilon \sim & N(0, \Sigma) \\
\theta_i \sim & N(\mu, a)
\end{aligned}
$$

This implies two sensible approaches: traditional asymptotics and maximum likelihood.

4.1.1 Asymptotics

The standard asymptotics approach appeals to the central limit theorem for the grand mean

$$\begin{aligned}\sqrt{K}\left(\overline{\overline{x}}-\mu\right) \rightsquigarrow & N\left(0,a\right)\\ & s.t.\\ \overline{\overline{x}} = & \frac{1}{K}\sum_{k}\overline{x}_i\end{aligned}$$

First, a simple approach appeals to the central limit theorem:

$$\begin{aligned}gvar\left(\overline{x}_i\right) = & \hat{a}_i + \Sigma_{ii}\\ \sqrt{K}\left(\hat{a}-a\right) \rightsquigarrow & N\left(0,var\left(a_i\right)\right)\\ & s.t.\\ \hat{a} = & \end{aligned}$$

Utilizing the plug-in estimators gives:

$$\begin{aligned}\hat{a} = & \frac{1}{K}\sum \hat{a}_i\\ = & gvar\left(\overline{x}_i\right) + tr\left(\Sigma\right)\\ gvar\left(\overline{x}_i\right) = & \frac{1}{K}E\left[\sum\left(\overline{x}_i - \frac{1}{K}\sum_{j}x_j\right)^2\right]\end{aligned}$$

Note that the asymptotics could be suspect here for small portfolios. We can correct the above by explicitly accounting for the uncertainty of $\overline{\overline{x}}$. The correlation between the error terms of x_i means that this quantity could be substantial. Star by decomposing the grand variance estimator:

$$\begin{aligned}gvar\left(x_i\right) = & \frac{1}{K}E\left[\sum_{i}\left(\left(\overline{x}_i - \frac{1}{K}\sum_{j}x_j\right)^2\right)\right]\\ = & \frac{1}{K}E\left[\sum_{i}\left(\left(\overline{x}_i - \mu + \mu - \frac{1}{K}\sum_{j}\overline{x}_j\right)^2\right)\right]\\ = & \frac{1}{K}E\left[\sum_{i}\left(\left(\overline{x}_i-\mu\right)^2 + \left(\mu - \frac{1}{K}\sum_{j}\overline{x}_j\right)^2 + 2\left(\overline{x}_i-\mu\right)\left(\mu - \frac{1}{K}\sum_{j}\overline{x}_j\right)\right)\right]\end{aligned}$$

Analyze each of these three terms:

$$\begin{aligned}
\frac{1}{K}\left(\sum_i E\left[(\overline{x}_i-\mu)^2\right]\right) =& \frac{1}{K}\sum_i E\left[(\theta_i+\varepsilon_i-\mu)^2\right] \\
=& a+\frac{1}{K}Tr\left(\Sigma\right) \\
E\left[\left(\mu-\frac{1}{K}\sum_j \overline{x}_j\right)^2\right] =& E\left[\left(\mu-\frac{1}{K}\sum_j\left(\theta_i+\varepsilon_i\right)\right)^2\right] \\
=& \mu^2-2E\left[\mu\frac{1}{K}\sum_j\left(\theta_i+\varepsilon_i\right)\right]+E\left[\frac{1}{K^2}\left(\sum_j\left(\theta_i+\varepsilon_i\right)\right)^2\right] \\
=& \frac{a}{K}+\frac{1'\Sigma 1}{K^2} \\
\frac{1}{K}E\left[\sum_i 2\left(\theta_i+\varepsilon_i-\mu\right)\left(\mu-\frac{1}{K}\sum_j\left(\theta_j+\varepsilon_j\right)\right)\right] =& \frac{1}{K}E\left[\sum_i 2\left(\theta_i-\mu\right)\left(\mu-\frac{1}{K}\sum_j\theta_j\right)-\frac{2}{K}\sum_i\sum_k\varepsilon_i\varepsilon_j\right] \\
=& -\frac{1}{K^2}E\left[2\sum_i\sum_j\theta_i\theta_j\right]+2\mu^2-\frac{2}{K^2}1'\Sigma 1 \\
=& -\frac{2a}{K}-\frac{2}{K^2}1'\Sigma 1
\end{aligned}$$

Thus:

$$\begin{aligned}
\frac{1}{K}E\left[\sum\left(\overline{x}_i-\frac{1}{K}\sum_j x_j\right)^2\right] =& \frac{K-1}{K}a+\frac{1}{K}Tr\left(\Sigma\right)-\frac{1'\Sigma 1}{K^2} \\
\implies a =& \frac{K}{K-1}gvar\left(\overline{x}\right)-\frac{1}{K-1}Tr\left(\Sigma\right)+\frac{1'\Sigma 1}{K\left(K-1\right)}
\end{aligned}$$

This equation is intuitive: the variance increases with dispersion. The adjustment for the noise in the measurement reduces the estimated variance of the true mean, while correlations between estimates of x_i accounts for the additional information provided in the comeasurements of x_i.

4.1.2 Maximum likelihood

Maximum likelihood estimation presents an alternative:

$$\overline{x}\sim N\left(\mu, Ia+\Sigma\right)$$

Or written as a likelihood:

$$\begin{aligned}
logL\left(\overline{x}|\cdot\right) =& -\frac{K}{2}\log 2\pi-\frac{1}{2}\log Det\left(Ia+\Sigma\right) \\
& -\frac{1}{2}\left(\overline{x}-1\mu\right)'\left[Ia+\Sigma\right]^{-1}\left(\overline{x}-1\mu\right)
\end{aligned}$$

This can be then maximized with respect to a and μ. Note that an analytical solution to μ is straight forward from first order conditions:

$$\begin{aligned} D_\mu : 0 =& D_\mu \left[-2 * \mu 1' \left[Ia + \Sigma \right]^{-1} \overline{x} + \mu^2 1' \left[Ia + \Sigma \right]^{-1} 1 \right] \\ =& -\mu 1' \left[Ia + \Sigma \right]^{-1} \overline{x} + \mu 1' \left[Ia + \Sigma \right]^{-1} 1 \\ \implies \hat{\mu} =& \frac{1' \left[Ia + \Sigma \right]^{-1} \overline{x}}{1' \left[Ia + \Sigma \right]^{-1} 1} \end{aligned}$$

The above estimate of the grand mean effectively assigns higher weights to values of $\overline{x}_i$ with greater precision. Specifically, the grand mean will be tilted towards assets of lower volatility, generally leading to a lower estimate and greater shrinkage. A downside of this approach is

4.2 Practical Implementation

1. Compute the Ledoit-Wolf estimator for the covariance matrix.
2. Compute $\hat{\Sigma}$, where

$$\hat{\Sigma}_{ij} = \frac{1}{T_{ij}} \hat{\Sigma}_{ij}^{LW}$$

3. Compute the grand mean

$$\overline{\overline{x}} = \frac{1}{K} \sum_i \overline{x}_i$$

4. Compute the grand mean $\overline{\overline{x}}$ and grand variance a
   (a) If appealing to asymptotics:
      i. Estimate the grand mean as

$$\begin{aligned} \hat{\mu} =& \overline{\overline{x}} \\ \overline{\overline{x}} \equiv& \frac{1}{K} \sum_i \overline{x}_i \end{aligned}$$

      ii. Then estimate the grand variance either as

$$\begin{aligned} \hat{a} =& gvar\left(\overline{x}\right) - \frac{1}{K} tr\left(\Sigma\right) \\ & s.t. \\ gvar\left(\overline{x}\right) =& \frac{1}{K} E\left[\sum_i \left(\overline{x}_i - \overline{\overline{x}} \right)^2 \right] \end{aligned}$$

or the bias-corrected form:

$$\hat{a} = \frac{K}{K-1} gvar\left(\overline{x}\right) - \frac{1}{K-1} Tr\left(\Sigma\right) + \frac{1'\Sigma 1}{K\left(K-1\right)}$$

(b) If employing MLE:

i. First compute the grand mean:

$$\hat{\mu} = \frac{1'\left[Ia+\Sigma\right]^{-1}\overline{x}}{1'\left[Ia+\Sigma\right]^{-1}1}$$

Then solve for the grand variance:

$$\begin{aligned}\hat{a} =& arg\min_{a} logL\left(\overline{x}|\cdot\right)\\ s.t.&\\ logL\left(\overline{x}|\cdot\right) \equiv& -\frac{K}{2}\log 2\pi - \frac{1}{2}\log Det\left(Ia+\Sigma\right) - \frac{1}{2}\left(\overline{x}-\hat{\mu}\right)'\left[Ia+\Sigma\right]^{-1}\left(\overline{x}-\hat{\mu}\right)\end{aligned}$$

5. Use the empirical estimates to shrink the mean:

$$\delta = \left[\hat{\Sigma}^{-1} + \frac{1}{\hat{a}}I\right]^{-1}\left[\Sigma^{-1}\overline{x} + \frac{1}{\hat{a}}\hat{\mu}\right]$$

5 Handling structured investments

The below procedure is loosely based on Faias and Santa-Clara 2017. The idea here is to treat structured investments (SI) in a manner similar to the underlying asset. The following procedure should be relatively straight forward:

1. Compute the returns and covariance matrix for all non-structured investment asset classes, as described elsewhere in this document. Apply any desired shrinkage. Make sure that the underlying asset class is included in the covariance matrix.

2. Let $C\left(V_T\right)$ be the terminal payoff of the structured investment as a function of K portfolio payoffs V_T. Then the mean and covariance of the SI with the assets in the portfolio are:

$$\begin{aligned}\mu_C =& E\left[C\left(V_T\right)\right]\\ \sigma_C =& E\left[\left(C\left(V_T\right)-\mu_C\right)\left(V_T-\mu_V\right)\right]\end{aligned}$$

3. To make estimating the above feasible:

(a) Draw B times such that draw $b \in 1:B$ is equal to:

$$V_T^b \sim N\left(\hat{E}\left(V\right);\hat{\Sigma}_V\right)$$

and $\hat{E}(V)$ and $\hat{\Sigma}_V$ are the respective (possibly shrunken) expected returns and covariance matrix of all assets other than the focal SI(s).

(b) Use the draws for plug-in estimators of the population moments:

$$\hat{\mu}_C = \frac{1}{B} \sum_{b \in 1:B} C\left(V_T^b\right)$$
$$\hat{\sigma}_C = E\left[\left(C\left(V_T^b\right) - \hat{\mu}_C\right)\left(V_T^b - \hat{\mu}_V\right)\right]$$

Then the covariance matrix of $\hat{\Sigma}_V$ can be augmented with the values provided in $\hat{\sigma}_C$. There is a very small chance that the resulting matrix is not positive definite. In this unlikely scenario, the full covariance matrix could be estimated using the empirical returns of all of the simulated assets.

5.1 Remarks

In the case of historical returns, a non-parametric block-bootstrapping approach could potentially better capture the underlying distribution. The basic approach would be the same as above, with a few notable differences. First, the draws for the bootstrap would be from the historical time series, or overlapping time series in the case of the covariances. Instead of drawing individual slices of returns (e.g. randomly drawing with replacement different time indices to create the simulated track records), a block-bootstrapping approach would draw blocks of consecutive periods. This would account for autocorrelation.

Second, the calculated covariances would be unshrunk, and therefore would augment the raw covariance matrix as opposed to the processed covariance matrix. The MRP method and any shrinkage would be applied to the raw matrix, as discussed elsewhere in the document. If the Ledoit-Wolf method is used for compute optimal shrinkage, a reasonable heuristic for shrinkage would be to not recalculate the optimal shrinkage amount $\hat{\alpha}$ as the information set remains unchanged. Then the shrinkage target would use the same off-diagonal correlation term as before. Other heuristics might modestly improve efficacy.

6 Appendix: Stein Estimation of Expected Returns

As a practical matter, the long-run mean return of GMAM 3.0 is likely a better estimate than the Stein estimator due to the greater amount of information imparted via the GMAM 3.0 estimator. However this is not guaranteed- the Stein estimator always uses information in the cross-section, while the G3 estimator does not.

Suppose, for now, that all track records have the same length. Then

$$\overline{x} \sim N\left(\mu_x, \frac{1}{T}\Sigma\right)$$

This implies

$$\sqrt{T}\Sigma^{-1/2}\overline{x} = N\left(\sqrt{T}\Sigma^{-1/2}\mu_x, I\right)$$

Therefore,

$$\begin{aligned} z \sim & N(\theta, I) \\ & s.t. \\ z = & \sqrt{T}\Sigma^{-1/2}\overline{x} \end{aligned}$$

Following Efron and Morris (1973), Stein (1961) defines loss as

$$L(\theta, \delta) = \frac{1}{K}||\delta - \theta||$$

where δ is the estimator of θ. Note that if $\delta = z$, then

$$L(\theta, \delta) = \frac{1}{K}E(x - \theta)^2 = 1$$

Assume

$$\theta_i \sim N(\mu_i, a)$$

Then

$$\begin{aligned} p(\theta_i|z) \propto & p(z|\theta)\, z(\theta) \\ = & \frac{1}{2\pi a^{1/2}} \exp\left[-\left(\frac{z_i - \theta_i}{2}\right)^2 - \frac{(\theta_i - \mu_i)^2}{2a}\right] \\ = & \frac{1}{2\pi a^{1/2}} \exp\left[-\left(\frac{\mu_i^2 + az_i^2 - 2(az_i + \mu_i)\theta_i + (1+a)\theta_i^2}{2a}\right)^2\right] \\ = & \frac{1}{2\pi a^{1/2}} \exp\left[-\left(\frac{2((1-b)z_i + b\mu_i)\theta_i + \theta_i^2}{2(1-b)}\right)\right] \exp\left(-\frac{az_i^2 + \mu_i^2}{2}\right) \\ = & \frac{1}{2\pi a^{1/2}} \exp\left[-\left(\frac{(\theta_i - (1-b)z_i)^2}{2(1-b)}\right)\right] \exp\left(-\frac{az_i^2 + \mu_i^2}{2} - \frac{((1-b)z_i + b\mu_i)^2 z_i^2}{2(1-b)}\right) \end{aligned}$$

$$\begin{aligned} & s.t. \\ b \equiv & \frac{1}{a+1} \end{aligned}$$

Hence

$$\begin{aligned} p\left(\theta_i|z\right) =& \frac{1}{\left(2\pi\right)^{1/2}\left(1-b\right)^{1/2}}\exp\left[-\frac{\left(\theta_i-\delta_i\right)^2}{2\left(1-b\right)}\right] \\ & s.t. \\ \delta_i \equiv& \left(1-b\right)z_i + b\mu_i \end{aligned}$$

This implies

$$\begin{aligned} & \theta_i \sim N\left(\left(1-b\right)z_i + b\mu_i, 1-b\right) \\ \implies & L\left(\theta,\delta\right) = 1-b \end{aligned}$$

Hence loss saved is given by b, with significant benefits when a is small. However, the above must be adjusted for the fact that b is estimated with error. The James-Stein estimator uses the following estimate for b:

$$\begin{aligned} \tilde{b} =& \frac{K-2}{\hat{s}} \\ s \equiv& \sum_i \left(z_i-\mu_i\right)^2 \end{aligned}$$

which implies the following:

$$\begin{aligned} \delta &= \left(1-\frac{K-2}{s}\right)z_i + \left(\frac{K-2}{s}\right)\mu_i \\ &= \left[1-\left(\frac{K-2}{s}\right)\right]\left(z_i-\mu_i\right)+\mu_i \end{aligned}$$

Efron and Morris (1973) generally advise bounding 0 and 1 to avoid extreme scenarios. Furthermore,, the above depends on unobservable parameter μ. The following bounded estimator uses the results of Efron and Morris, bounded with a plug-in estimator for the mean:

$$\begin{aligned} \hat{\delta} =& \left[1-\hat{b}\right]\left(z_i-\overline{z}\right)+\overline{z} \\ & s.t. \\ \hat{b} =& \max\left(0, \min\left(1, \frac{K-3}{\hat{s}}\right)\right) \\ \hat{s} \equiv& \sum_i \left(z_i-\mu\right)^2 \end{aligned}$$

The above implies the following steps for simple shrinkage:

1. For $K \leq 3$, use a simple average of each column. For $K > 3$, use the following approach.
2. Compute the Ledoit-Wolf estimator for the covariance matrix.

3. Compute $\hat{\Sigma}$, where

$$\hat{\Sigma}_{ij} = \frac{1}{T_{ij}} \hat{\Sigma}_{ij}^{LW}$$

(alternatively, T_{ij} could be the minimum for a more conservative shrinkage estimate).

4. Compute

$$z \equiv \hat{\Sigma}^{-1/2} \overline{x}$$
$$\hat{\zeta} \equiv \hat{\Sigma}^{-1/2} 1 \overline{\overline{x}}$$

5. Estimate b from

$$\hat{s} = \sum_i \left(z_i - \hat{\zeta}_i \right)^2$$
$$\hat{b} = \max \left(0, \min \left(1, \frac{K-3}{\hat{s}} \right) \right)$$

6. Estimate δ from

$$\hat{\delta} = [1 - b] \left(z - \hat{\zeta} \right) + \zeta$$

7. Compute $\hat{\mu}_x$ from:

$$\hat{\mu}_x = \hat{\Sigma}^{1/2} \hat{\delta}$$

Note that this implies

$$\hat{\mu}_x = \left[1 - \max \left(0, \min \left(1, \frac{K-3}{\hat{s}} \right) \right) \right] \left(\overline{x} - \overline{\overline{x}} \right) + \overline{\overline{x}}$$

Note the above coordinate transformation leaves some efficiency behind by throwing out differences in variance between the estimates in assigning the shrinkage, as discussed by Efron and Morris in their remarks on equation 8.3. Anecdotally, the above seems to undershrink relative to other techniques.

7 Appendix: Alternative approaches to constraint incompatibility

Two alternative approaches to this issue were rejected to handle the differences in the constraints between the estimated moments from paths versus the estimation from the original sample:

1. An alternative is to adjust the objective function. Let w^* represent the original optima from the sample data. Let $U(w)$ be the objective function, and $g(w) \leq 0$ be the constraints. Then one such scoring function could be $\left(U(w_b) - U(w_b^*) \right)^2 + g(w)^2 \iota_{g(w)>0}$. To see why even higher

objective functions should be penalized when they exceed the objective function result from the original sample, first note that the parametric bootstrap treats the original sample moments as population moments. The original weights are treated as optimal, and therefore given optimal w^*, $U(w^*) \geq U(w)$ for all w where $g(w) \leq 0$, hence $U(w_b) > U(w^*) \implies g(w) > 0$. Finally, because the reverse is not necessarily true, an additional penalty on the constraint itself further ensures that $U(w^*)$ is the optimal portfolio. This approach was rejected due to the constraint being "soft," along with the approach's relative complexity

2. Ignore the issue. This effectively treats the constraints as based on the sample moments, which are perfectly measured. If the regularization leads to significant differences between the population and sample moments, the constraints could differ significantly from those implied by the population moments. When the objective function and scoring function are equal, the optimizer could favor solutions where such violations occurred. Still this approach is simple and correct under the sample moments interpretation.