

# Minimum Variance

## 1 Overview

The minimum variance portfolio estimates the variance-covariance matrix using a factor model approach similar to MSCI Barra and constructs the portfolio that minimizes variance. Returns are assumed to follow a linear factor model, where security characteristics serve as observable factor loadings and latent factor returns are inferred from cross-sectional regressions.

## 2 Data Preparation

Before entering the factor model, stock characteristics are preprocessed as follows. First, each characteristic is percentile-ranked cross-sectionally (within country and month) and centered by subtracting 0.5, so that values lie in  $[-0.5, 0.5]$ . Missing values are imputed with zeros. The characteristics are then standardized to have mean zero and unit standard deviation within each month.

## 3 Variance-Covariance Estimation

### 3.1 Factor Model

To estimate the variance-covariance matrix, we use an approach similar to the one used by MSCI Barra, which is based on the assumption that returns follow a linear factor model. The idea is to treat security characteristics as observable factor loadings and infer the latent factor returns from cross-sectional regressions of excess returns on security characteristics. Specifically, each day we estimate the cross-sectional ridge regression:

$$r_{i,t+1} = x'_{i,t} \hat{f}_{t+1} + \hat{\epsilon}_{i,t+1}, \quad (1)$$

where  $r_{i,t+1}$  is the stock's realized excess return,  $x$  is a vector of the stock's characteristics,  $\hat{\epsilon}_{i,t+1}$  is the regression residual, and  $\hat{f}_{t+1}$  is the estimated ridge regression parameters, which we treat as estimated factor returns. The regression has no intercept; instead, all 12 industry dummies are included so that they absorb what would otherwise be the intercept. The ridge penalty is set to  $10^{-4}$ . The stock characteristics are the 402 **features** and 12 industry dummies based on SIC codes and the industry definition from Kenneth French's webpage.<sup>1</sup>

---

<sup>1</sup>Our approach differs from Jensen et al. (2025), who use a compressed version of 13 factor themes (made from the original 153 JKP characteristics) plus the 12 industry dummies. We use the raw characteristics to enable the variance-covariance estimate to change as we change the set of characteristics. The Sharpe ratio is higher if we use the 13 factor themes instead of the 402 raw characteristics. We also differ by using ridge instead of OLS regression. We use ridge because some of our characteristics are highly correlated, and using a small penalty makes the estimates robust to near multicollinearity.

The structure in Equation 1 implies that the variance-covariance matrix is:

$$\hat{\Sigma}_t = X_t \text{Var}_t(\hat{f}_{t+1}) X_t' + \text{diag}(\text{Var}_t(\hat{\epsilon}_{t+1})),$$

where  $X_t$  is the matrix of stock characteristics,  $\text{Var}_t(\hat{f}_{t+1})$  is the variance-covariance matrix of factor returns, and  $\text{diag}(\text{Var}_t(\hat{\epsilon}_{t+1}))$  is a matrix with the idiosyncratic variances in the diagonal and zero elsewhere.

### 3.2 Factor Return Covariance

We estimate  $\text{Var}_t(\hat{f}_{t+1})$  as the exponentially weighted sample variance-covariance matrix of the last ten years of daily returns. The exponential weighting scheme gives observations  $j$  days from  $t$  a weight of  $w_{t-j} = c \cdot 0.5^{j/\text{half-life}}$ , where  $c$  is a constant ensuring that the weights sum to one, and it ensures that recent observations affect the estimate more than distant ones. Following the MSCI Barra USE4S model, we use a half-life of 504 days for correlations and 84 days for variances (Menchero, Orr, and Wang 2011, Table 4.1).

### 3.3 Idiosyncratic Variance

Similarly, we estimate each stock’s idiosyncratic variance,  $\text{Var}_t(\hat{\epsilon}_{i,t+1})$ , as the exponentially weighted moving average of squared residuals,  $\hat{\epsilon}_{i,t+1}$  from Equation 1, with a half-life of 84 days. The EWMA variance is initialized (seeded) from the first 63 observations (~3 months). The half-life is again chosen as the one from the MSCI Barra USE4S model (Menchero, Orr, and Wang 2011, Table 5.1). To estimate the idiosyncratic variance we require at least 200 non-missing observations within the last 252 trading days. For stocks without an idiosyncratic variance estimate, we estimate the function,  $\ln\left(\sqrt{\hat{\text{Var}}_t(\hat{\epsilon}_{i,t+1})}\right) = \hat{f}_t(x_{i,t})$ , using a ridge regression model with a small ridge penalty of  $10^{-4}$ , and use it to estimate the missing variances.<sup>2</sup>

## 4 Portfolio Construction

Given the estimated variance-covariance matrix  $\hat{\Sigma}_t$ , the minimum-variance portfolio weights are:

$$w_t = \frac{\hat{\Sigma}_t^{-1} \mathbf{1}}{\mathbf{1}' \hat{\Sigma}_t^{-1} \mathbf{1}},$$

where  $\mathbf{1}$  is a vector of ones. The weights sum to one by construction. To avoid inverting the large  $N \times N$  covariance matrix directly, we exploit the factor structure and use the Woodbury matrix identity, which reduces the problem to inverting a  $K \times K$  matrix (where  $K$  is the number of factors).

## References

- Jensen, Theis Ingerslev, Bryan T Kelly, Semyon Malamud, and Lasse Heje Pedersen. 2025. “Machine Learning and the Implementable Efficient Frontier.” *Review of Financial Studies (Forthcoming)*.  
Menchero, Jose, D. J. Orr, and Jun Wang. 2011. “Barra US Equity Model (USE4).” Methodology Notes. MSCI Barra.

---

<sup>2</sup>Our approach follows MSCI Barra (Menchero, Orr, and Wang 2011, Eq. 5.3), except that they use an OLS regression. We use a ridge regression because some of our characteristics are highly correlated, and using a small penalty helps make the estimates robust to near multicollinearity.