Evaluation of Asset Pricing Models: Optimal Risk Premia and Goodness-of-Fit Measures

Robert L. Kimmel
罗伯特. 金默尔

Nanyang Business School
Nanyang Technological University

24 July 2020

We are interested in linear factor models of expected returns,

$$R_{i,t} = \mathrm{E}\left[R_{i,t}\right] + \sum_{j=1}^{N} \beta_{i,j}\left(F_{j,t} - \mathrm{E}\left[F_{j,t}\right]\right) + \epsilon_{i,t},$$

where

$$\mathrm{E}\left[\epsilon_{i,t}\right] = 0, \qquad \forall i, \qquad \text{and} \qquad \mathrm{Cov}\left[\epsilon_{i,t}, F_{j,t}\right] = 0, \qquad \forall i,j.$$

The expected level of returns is specified as

$$\underbrace{\mathrm{E}\left[R_{i,t}\right] - R_{f,t}}_{\text{Expected Excess Return}} = \sum_{j=1}^{N} \underbrace{\beta_{i,j}}_{\text{Risk Exposure}} \times \underbrace{\Lambda_j}_{\text{Risk Premium}} .$$

Well-grounded in economic theory—widely, but often poorly, understood.

Questions in which we might be interested:

1. Does the model explain the expected returns of all the assets?
2. Does the model explain the expected returns of the assets better than some other model?
3. Does the performance of the model degrade if a particular factor (or set of factors) is removed?
4. What are the risk premia of the factors?

When a perfect linear relation between the expected returns and the beta coefficients exists, the definition of the risk premia is clear—they are the slope coefficients of the perfect linear relation.

What if there is not a perfect linear relation (i.e., the model does not explain the expected returns perfectly)?

Can we still assign risk premia to the factors in a reasonable, and uniquely defined way?

Is there a reasonable, well-defined way to rank models which are misspecified?

If one choice of risk premia strictly dominates another (i.e., causes the prediction error of some assets to decrease, but does not cause the prediction error of any asset to increase), then it is clear that the first choice is better.

1. However, there is precisely one situation in which this is the case—if the first choice results in the perfect model, i.e., one that has no prediction error for any asset.
2. If the first choice of risk premia fails to predict all expected returns perfectly, it never strictly dominates any other choice of risk premia.

So unless we have a perfect model, choice of risk premia and model evaluation is something of a compromise.

We can choose the risk premia to do a better job predicting the expected returns of some assets, by allowing its performance to degrade on other assets.

Similarly, every misspecified model still predicts the expected returns of some assets well. So which assets are the most important to match? How much weight should be assigned to each asset?

The same principle holds for evaluation of fit—if two models are misspecified, each will always do a better job predicting expected returns of some particular assets than the other model.

Is there a method for choosing risk premia, and for evaluating the goodness-of-fit of models for expected returns, that is in some sense "optimal"? Or are the choices essentially arbitrary?

Methods of assignment of risk premia that have been used:

1. Two-pass OLS regression—seriously problematic (see Kandel and Stambaugh (1995), Hou and Kimmel (2020), others)
2. Two-pass GLS regression—much less problematic, but results can be found in much simpler ways (see same authors)
3. Hou and Kimmel (2020)—simple, straightforward, robust, rarely used.

Model evaluation criteria:

1. Gibbons/Ross/Shanken test, GMM tests—rigorous and well-defined, but give a binary reject/do-not-reject decision for a single model.
2. Sharpe ratio tests—compare misspecified models.
3. Analysis of second-pass coefficients in regression methods—commonly used, conceptually flawed, provides no useful information.
4. Hansen-Jagannathan distance.
5. Other methods?

Model evaluation techniques are often very ad-hoc and arbitrary.

Two fundamental guiding principles:

1. A model evaluation criterion should be based on the models' predictions.
    1. The purpose of the model is to describe the cross-sectional expected returns. So the performance of the model should be based on the expected returns it predicts, not on other criteria.
    2. If one model describes the expected return of some asset better than another model, but the two models agree on all assets returns that are uncorrelated with this particular asset, then the first model should be evaluated as "better".

The only model that ever strictly dominates (that is, that predicts the expected returns at least as well as, and sometimes better than) any other model is the perfect model that predicts all expected returns correctly.

If two models each have some degree of misspecification, each one will predict the expected returns of *some particular* assets better than the other, so there will not be strict dominance.

Factor rotation invariance—although a model evaluation criterion depends on the explanatory factors, it is factor rotation invariant if taking linear combinations of the same factors does not change the result.

Suppose factors F are replaced by other factors,

$$F^{\star} = \Omega F,$$

where Ω is a full-rank matrix.

A model with the evaluation criterion is invariant to factor rotation if the criterion is the same using F or $F^{\star}$.

In words—a model evaluation method is invariant to factor rotation if, when the factors are replaced by linear combinations of the same factors, the method returns the same score.

Simple example of factor rotation—consider Fama/French model, with factors

$$F = \begin{bmatrix} RMRF \\ \\ SMB \\ \\ HML \end{bmatrix}.$$

Replace with alternate factors

$$F^{\star} = \begin{bmatrix} RMRF \\ \\ SMB + HML \\ \\ SMB - HML \end{bmatrix}.$$

Would a reasonable model evaluation method treat the two models differently?

Suppose the beta coefficients for some asset on a set of factors F are given by β_i, and the factors have risk premia γ.

Then the beta coefficients for some asset on a set of factors $F^\star = \Omega F$ are

$$\beta_i^\star = \beta_i \Omega^{-1}.$$

If the alternate factors have the risk premia

$$\gamma^\star = \Omega\gamma,$$

then the model based on the alternate factors predicts exactly the same expected returns as the original model.

Could some other vector of risk premia, $\gamma_0^\star$ provide a better fit for the model with the alternate factors?

If so, then $\gamma_0 = \Omega^{-1}\gamma_0^\star$ provides the exact same "better" fit for the model with the original factors.

So if the expected returns predicted using $F^\star$ and $\gamma_0^\star$ are better than those using $F^\star$ and $\gamma^\star$, then should not the expected returns predicted using F and γ_0 be better than those found using F and γ?

If you respect the first guiding principle, and if γ are the "best" risk premia for the factors F, then

$$\gamma^\star = \Omega\gamma$$

must be the "best" risk premia for the factors $F^\star$.

Furthermore, also by the first principle, the fit of the model based on $F^\star$ must be the same as the fit of the model based on F.

Asset rotation invariance—although a model evaluation criterion depends on the universe of investment opportunities available, it is asset rotation invariant if it does not depend on how these investment opportunities are packaged into individual assets.

Suppose the assets R are replaced by another set of assets,

$$R^{\star} = \Omega R,$$

where Ω is a full-rank matrix, with the sum of elements in each column equal to one.

A model evaluation criterion is invariant to asset rotation if the criterion is the same using R or $R^{\star}$.

In words—a model evaluation method is invariant to asset rotation if, when the assets are replaced by portfolios that offer exactly the same investment opportunities as the original assets, the method returns the same score.

Simple example of asset rotation—two assets, R_1 and R_2.

Let

$$R_1^\star = \frac{1}{2}R_1 + \frac{1}{2}R_2 \qquad \text{and} \qquad R_2^\star = \frac{3}{2}R_1 - \frac{1}{2}R_2.$$

Assets $R_1^\star$ and $R_2^\star$ offer exactly the same investment opportunities as R_1 and R_2.

A rotation-invariant model evaluation technique returns the same score if applied to assets R_1 and R_2 or $R_1^\star$ and $R_2^\star$.

Typical OLS cross-sectional regression methods are *not* rotation invariant.

Kandel and Stambaugh (1995) show shocking examples of how badly non-rotation invariant techniques can be manipulated to provide any desired result.

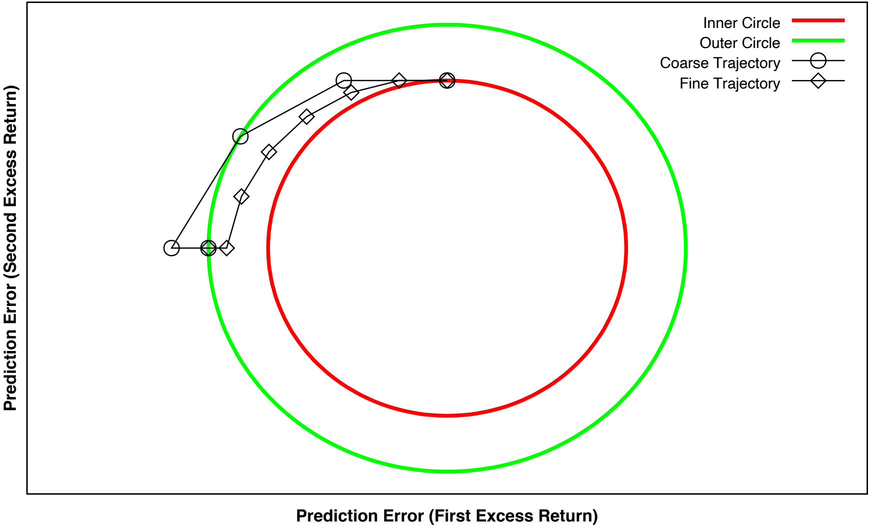
Asset rotation invariance is a consequence of the two guiding principles (with a slightly annoying technical exception, if two models are exactly tied on a primary model evaluation criterion).

We start with a special case, of two assets with equal variance of return, and zero correlation.

The job is to show that, if the two guiding principles are respected, then a model with a lower sum of squared prediction errors has a better fit.

Plot prediction errors on a graph, one asset on each axis—models with the same sum of squared prediction errors fall on a circle around the origin. Larger radius means larger sum of squared prediction error.

1. Show that, if the two guiding principles are respected, every model on a circle with a smaller radius, has a better fit than every model on a circle with larger radius.
2. Do this, by following, segment by segment, a spiral trajectory from a point on the inner circle to a point on the outer circle.


Inner Circle
Outer Circle
Coarse Trajectory
Fine Trajectory
Prediction Error (Second Excess Return)
Prediction Error (First Excess Return)

By following a sufficiently fine spiral trajectory from a point on the inner circle to a point on the outer circle, the goodness-of-fit must decrease along each segment.

Because it decreases along each segment, the point on the outer circle has worse goodness-of-fit than the point on the inner circle.

This is true regardless of which points we choose on the two circles.

It follows that every point on the inner circle, has better goodness-of-fit, than every point on the outer circle.

This is true for any two circles with different radii.

Extend to many assets, uncorrelated returns, equal variance.

Achieved by orthogonal axis rotation.

Instead of circles, we have spheres (in three dimensions), or hyperspheres (if four or more dimensions). We use “hypersphere” throughout, regardless of number of dimensions.

Choose a point on the inner hypersphere, and a point on the outer hypersphere.

Work with excess returns,

$$Z = R - R_f.$$

Axis rotation—choose portfolios of the original assets, such that excess returns are given by

$$Z^\dagger = C^T Z,$$

where

$$CC^T = I.$$

The new excess returns also have equal variance, and are uncorrelated.

Also note that the sum of squared prediction errors for the alternate assets is the same as for the original assets.

Prediction errors are

$$\eta_Z \equiv \hat{\mu}_Z - \mu_Z.$$

Then

$$\eta_Z^{\dagger} = C^T \eta_Z,$$

and

$$\left(\eta_Z^{\dagger}\right)^T \eta_Z^{\dagger} = \eta_Z^T C C^T \eta_Z = \eta_Z^T \eta_Z.$$

Let the first column of C be proportional to the observation errors on the inner hypersphere.

Let the second column of C be the component of the observation errors on the outer hypersphere, that is orthogonal to the first column of C.

Consider a model that has the prediction errors corresponding to the point on the first hypersphere. Recall that the alternate assets have excess returns

$$Z^{\dagger} = C^{T} Z.$$

Then the prediction error for the first alternate excess return is non-zero, but the prediction errors for all the other alternate excess returns are zero.

Now consider a model that has prediction errors corresponding to the point on the second hypersphere. This model has a non-zero prediction error for the first and second alternate assets, and prediction errors of zero for all the other alternate assets.

The “spiral” proof can now be applied to the alternate excess returns, instead of the original excess returns.

Both the endpoints, and every point along the spiral trajectory, have zero prediction errors for all the alternate assets except the first two.

Since the sum of squared prediction errors for the alternate assets is the same as for the original assets, it follows that every model on the outer hypersphere has a worse goodness-of-fit measure than every model on the inner hypersphere.

This is the case for all choices of inner and outer hypersphere, even if they are very close together.

This, in the many asset case, with identical variance and zero correlation, larger sum of squared prediction errors means worse goodness-of-fit (provided we adhere to the two guiding principles).

The key result can be extended to the many asset case with arbitrary variances and covariances, with slight modification.

Let Σ_{ZZ} be the covariance matrix of the excess returns, with spectral decomposition

$$\Sigma_{ZZ} = C\Lambda C^T.$$

This time, it is not the sum of squared prediction errors that determines the model ranking, but rather, the quantity

$$\eta^T \Sigma_{ZZ}^{-1} \eta.$$

We can construct an alternate set of alternate assets,

$$Z^\star = \Lambda^{-\frac{1}{2}} C^T Z.$$

The covariance matrix of the alternate excess returns is the identity matrix.

The prediction errors for the alternate assets are

$$\eta^\star = \Lambda^{-\frac{1}{2}} C^T \eta,$$

so that

$$\eta = C \Lambda^{\frac{1}{2}} \eta^\star,$$

The quantity

$$\eta^T \Sigma_{ZZ}^{-1} \eta$$

is therefore equal to

$$(\eta^\star)^T \eta^\star,$$

that is, the sum of the squared prediction errors for the transformed assets.

So consider two models, one with prediction errors η_X and one with prediction errors η_Y. If

$$\eta_X^T \Sigma_{ZZ}^{-1} \eta_X < \eta_Y^T \Sigma_{ZZ}^{-1} \eta_Y,$$

then the model with prediction errors η_X has a smaller sum of squared prediction errors when applied to the alternate assets.

But these alternate assets are uncorrelated with each other, and each one has a variance of one. So if the two guiding principles are respected, then whichever has the lower sum of squared prediction errors (for the alternate assets) has the better goodness-of-fit.

But this directly implies that whichever model has the lower quantity,

$$\eta^T \Sigma_{ZZ}^{-1} \eta,$$

has the better goodness-of-fit.

We can now put it all together, and derive the asset rotation invariance principle. Suppose there are two models, with the quantities

$$\eta_X^T \Sigma_{ZZ}^{-1} \eta_X \qquad \text{and} \qquad \eta_Y^T \Sigma_{ZZ}^{-1} \eta_Y$$

determining their relative goodness-of-fit.

We construct alternate set of assets,

$$Z^\star = \Omega Z.$$

The covariance matrix for the alternate assets is

$$\Sigma_{Z^\star Z^\star} = \Omega \Sigma_{ZZ} \Omega^T,$$

and the prediction errors for the alternate assets for the two models are

$$\eta_X^\star = \Omega \eta_X \qquad \text{and} \qquad \eta_Y^\star = \Omega \eta_Y.$$

So we have

$$(\eta_X^\star)^T \Sigma_{Z^\star Z^\star}^{-1} \eta_X^\star = \eta_X^T \Sigma_{ZZ}^{-1} \eta_X \qquad \text{and} \qquad (\eta_Y^\star)^T \Sigma_{Z^\star Z^\star}^{-1} \eta_Y^\star = \eta_Y^T \Sigma_{ZZ}^{-1} \eta_Y.$$

There are the quantities that determine which model has better goodness-of-fit—they are preserved through asset rotation!

The exception is if two models are “tied”, that is, if

$$\eta_X^T \Sigma_{ZZ}^{-1} \eta_X = \eta_Y^T \Sigma_{ZZ}^{-1} \eta_Y.$$

In this case, a goodness-of-fit measure does not have to preserve the ordering of the two models through asset rotation.

If a model evaluation technique is both asset rotation and factor rotation invariant, there is not much choice in how to evaluate models with traded factors.

Begin with factors F and excess asset returns Z. Then

$$\mathrm{E}\left[F\right] = \mu_F, \qquad \mathrm{E}\left[Z\right] = \mu_Z,$$
$$\mathrm{Var}\left[F\right] = \Sigma_{FF}, \qquad \mathrm{Var}\left[Z\right] = \Sigma_{ZZ}, \qquad \text{and} \qquad \mathrm{Cov}\left[F, Z^T\right] = \Sigma_{FZ}.$$

However, since the factors are traded, we have $F = \Gamma Z$ for some Γ, which means that

$$\mu_F = \Psi \mu_Z, \qquad \Sigma_{FZ} = \Psi \Sigma_{ZZ}, \qquad \text{and} \qquad \Sigma_{FF} = \Psi \Sigma_{ZZ} \Psi^T.$$

The middle condition is equivalent to

$$\Psi = \Sigma_{FZ} \Sigma_{ZZ}^{-1}.$$

Introduce new factors, risk premia, and excess returns,

$$F^\star = \left(\Sigma_{FZ}\Sigma_{ZZ}^{-1}\Sigma_{ZF}\right)^{-\frac{1}{2}} F, \qquad \gamma^\star = \left(\Sigma_{FZ}\Sigma_{ZZ}^{-1}\Sigma_{ZF}\right)^{-\frac{1}{2}} \gamma,$$
$$\text{and} \qquad Z^\star = \Sigma_{ZZ}^{-\frac{1}{2}} Z,$$

where the matrix "inverse square root" operations are interpreted to mean the unique symmetric matrices such that

$$\left(\Sigma_{FZ}\Sigma_{ZZ}^{-1}\Sigma_{ZF}\right)^{-\frac{1}{2}} \left(\Sigma_{FZ}\Sigma_{ZZ}^{-1}\Sigma_{ZF}\right)^{-\frac{1}{2}} = \left(\Sigma_{FZ}\Sigma_{ZZ}^{-1}\Sigma_{ZF}\right)^{-1}$$

and

$$\Sigma_{ZZ}^{-\frac{1}{2}}\Sigma_{ZZ}^{-\frac{1}{2}} = \Sigma_{ZZ}^{-1}.$$

Then

$$\mathrm{E}\left[F^\star\right] = \left(\Sigma_{FZ}\Sigma_{ZZ}^{-1}\Sigma_{ZF}\right)^{-\frac{1}{2}} \Sigma_{FZ}\Sigma_{ZZ}^{-1}\mu_Z, \qquad \mathrm{E}\left[Z^\star\right] = \Sigma_{ZZ}^{-\frac{1}{2}}\mu_Z,$$
$$\mathrm{Var}\left[F^\star\right] = I, \qquad \mathrm{Var}\left[Z^\star\right] = I,$$
$$\text{and} \qquad \mathrm{Cov}\left[F^\star, \left(Z^\star\right)^T\right] = \left(\Sigma_{FZ}\Sigma_{ZZ}^{-1}\Sigma_{ZF}\right)^{-\frac{1}{2}} \Sigma_{FZ}\Sigma_{ZZ}^{-\frac{1}{2}}.$$

Make another transformation,

$$F^{\star\star} = F^{\star}, \qquad \gamma^{\star\star} = \gamma^{\star},$$

$$\text{and} \qquad Z^{\star\star} = \begin{bmatrix} \left(\Sigma_{FZ}\Sigma_{ZZ}^{-1}\Sigma_{ZF}\right)^{-\frac{1}{2}} \Sigma_{FZ}\Sigma_{ZZ}^{-\frac{1}{2}} \\ \\ C \end{bmatrix} Z^{\star}$$

where

$$C\Sigma_{ZZ}^{-\frac{1}{2}}\Sigma_{ZF} = 0 \qquad \text{and} \qquad CC^{T} = I.$$

Then

$$\mathrm{E}\left[F^{\star\star}\right] = \left(\Sigma_{FZ}\Sigma_{ZZ}^{-1}\Sigma_{ZF}\right)^{-\frac{1}{2}} \Sigma_{FZ}\Sigma_{ZZ}^{-1}\mu_{Z},$$

$$\mathrm{E}\left[Z^{\star\star}\right] = \begin{bmatrix} \left(\Sigma_{FZ}\Sigma_{ZZ}^{-1}\Sigma_{ZF}\right)^{-\frac{1}{2}} \Sigma_{FZ}\Sigma_{ZZ}^{-1}\mu_{Z} \\ \\ C\Sigma_{ZZ}^{-\frac{1}{2}}\mu_{Z} \end{bmatrix},$$

$$\mathrm{Var}\left[F^{\star\star}\right] = I, \qquad \mathrm{Var}\left[Z^{\star\star}\right] = I,$$

$$\text{and} \qquad \mathrm{Cov}\left[F^{\star\star}, \left(Z^{\star\star}\right)^{T}\right] = \begin{bmatrix} I_N & 0_{N\times(M-N)} \end{bmatrix}$$

At this point, it is perhaps worth noting that both the factors and the excess returns have been orthogonalised, and that the N factors are simply the first N excess returns.

Since the previous two transformations have worked out so well, we will try one more.

$$F^{\star\star\star} = \begin{bmatrix} \frac{\mu_Z^T \Sigma_{ZZ}^{-1} \Sigma_{ZF} \left(\Sigma_{FZ} \Sigma_{ZZ}^{-1} \Sigma_{ZF}\right)^{-\frac{1}{2}}}{\sqrt{\mu_Z^T \Sigma_{ZZ}^{-1} \Sigma_{ZF} \left(\Sigma_{FZ} \Sigma_{ZZ}^{-1} \Sigma_{ZF}\right)^{-1} \Sigma_{FZ} \Sigma_{ZZ}^{-1} \mu_Z}} \\ \\ D_1 \end{bmatrix} F^{\star\star}$$

and $$\gamma^{\star\star\star} = \begin{bmatrix} \frac{\mu_Z^T \Sigma_{ZZ}^{-1} \Sigma_{ZF} \left(\Sigma_{FZ} \Sigma_{ZZ}^{-1} \Sigma_{ZF}\right)^{-\frac{1}{2}}}{\sqrt{\mu_Z^T \Sigma_{ZZ}^{-1} \Sigma_{ZF} \left(\Sigma_{FZ} \Sigma_{ZZ}^{-1} \Sigma_{ZF}\right)^{-1} \Sigma_{FZ} \Sigma_{ZZ}^{-1} \mu_Z}} \\ \\ D_1 \end{bmatrix} \gamma^{\star\star}$$

where

$$D_1 \left(\Sigma_{FZ} \Sigma_{ZZ}^{-1} \Sigma_{ZF}\right)^{-\frac{1}{2}} \Sigma_{FZ} \Sigma_{ZZ}^{-1} \mu_Z = 0 \qquad \text{and} \qquad D_1 D_1^T = I.$$

Simultaneously, we transform

$$Z^{\star\star\star} = \begin{bmatrix} \dfrac{\mu_Z^T \Sigma_{ZZ}^{-1} \Sigma_{ZF} \left(\Sigma_{FZ} \Sigma_{ZZ}^{-1} \Sigma_{ZF} \right)^{-\frac{1}{2}}}{\sqrt{\mu_Z^T \Sigma_{ZZ}^{-1} \Sigma_{ZF} \left(\Sigma_{FZ} \Sigma_{ZZ}^{-1} \Sigma_{ZF} \right)^{-1} \Sigma_{FZ} \Sigma_{ZZ}^{-1} \mu_Z}} \\ \\ D_1 \\ \\ \dfrac{\mu_Z^T \Sigma_{ZZ}^{-\frac{1}{2}} C^T}{\sqrt{\mu_Z^T \Sigma_{ZZ}^{-\frac{1}{2}} C^T C \Sigma_{ZZ}^{-\frac{1}{2}} \mu_Z}} \\ \\ D_2 \end{bmatrix} Z^{\star\star}$$

where

$$D_2 C \Sigma_{ZZ}^{-\frac{1}{2}} \mu_Z = 0 \qquad \text{and} \qquad D_2 D_2^T = I.$$

Then

$$\mathrm{E}\left[F^{\star\star\star}\right] = \begin{bmatrix} \sqrt{\mu_Z^T \Sigma_{ZZ}^{-1} \Sigma_{ZF} \left(\Sigma_{FZ} \Sigma_{ZZ}^{-1} \Sigma_{ZF}\right)^{-1} \Sigma_{FZ} \Sigma_{ZZ}^{-1} \mu_Z} \\ 0_{(N-1)\times 1} \end{bmatrix},$$

$$\mathrm{E}\left[Z^{\star\star\star}\right] = \begin{bmatrix} \sqrt{\mu_Z^T \Sigma_{ZZ}^{-1} \Sigma_{ZF} \left(\Sigma_{FZ} \Sigma_{ZZ}^{-1} \Sigma_{ZF}\right)^{-1} \Sigma_{FZ} \Sigma_{ZZ}^{-1} \mu_Z} \\ 0_{(N-1)\times 1} \\ \sqrt{\mu_Z^T \Sigma_{ZZ}^{-\frac{1}{2}} C^T C \Sigma_{ZZ}^{-\frac{1}{2}} \mu_Z} \\ 0_{(M-N-1)\times 1} \end{bmatrix},$$

$$\mathrm{Var}\left[F^{\star\star\star}\right] = I, \qquad \mathrm{Var}\left[Z^{\star\star\star}\right] = I,$$

and $$\mathrm{Cov}\left[F^{\star\star\star}, \left(Z^{\star\star\star}\right)^T\right] = \begin{bmatrix} I_N & 0_{N\times(M-N)} . \end{bmatrix}$$

The important thing to note is that the first and second moments (including the cross-moments) of $F^{\star\star\star}$ and $Z^{\star\star\star}$ are characterised completely by only two numbers—the excess returns of two of the (transformed) assets.

One of these assets is a portfolio of the factors; the other has excess returns that are uncorrelated with any of the factors.

It is not immediately obvious, but the sum of the squares of the expected returns of the two assets does not depend on the factors—effectively, only one of the numbers is specific to the model being considered. The other number depends only on the properties of the assets. This follows from the invariance of the quantity

$$\mu_Z^T \Sigma_{ZZ}^{-1} \mu_Z$$

to rotation of the assets.

It follows that

$$\mu_Z^T \Sigma_{ZZ}^{-1} \mu_Z = \mu_Z^T \Sigma_{ZZ}^{-1} \Sigma_{ZF} \left(\Sigma_{FZ} \Sigma_{ZZ}^{-1} \Sigma_{ZF} \right)^{-1} \Sigma_{FZ} \Sigma_{ZZ}^{-1} \mu_Z + \mu_Z^T \Sigma_{ZZ}^{-\frac{1}{2}} C^T C \Sigma_{ZZ}^{-\frac{1}{2}} \mu_Z,$$

and that we can rewrite the expected excess returns (after the three rotations) as

$$\mathrm{E}\left[Z^{\star\star\star} \right] = \begin{bmatrix} \sqrt{\mu_Z^T \Sigma_{ZZ}^{-1} \Sigma_{ZF} \left(\Sigma_{FZ} \Sigma_{ZZ}^{-1} \Sigma_{ZF} \right)^{-1} \Sigma_{FZ} \Sigma_{ZZ}^{-1} \mu_Z} \\ 0_{(N-1)\times 1} \\ \sqrt{\mu_Z^T \Sigma_{ZZ}^{-1} \mu_Z - \mu_Z^T \Sigma_{ZZ}^{-1} \Sigma_{ZF} \left(\Sigma_{FZ} \Sigma_{ZZ}^{-1} \Sigma_{ZF} \right)^{-1} \Sigma_{FZ} \Sigma_{ZZ}^{-1} \mu_Z} \\ 0_{(M-N-1)\times 1} \end{bmatrix}.$$

Note that the only two distinct numbers that appear in the moment expressions can be interpreted as follows:

1. The maximum Sharpe ratio that can be achieved with the factor portfolios
2. The maximum Sharpe ratio that can be achieved with assets that are uncorrelated with the factor portfolios

If an evaluation technique is invariant to asset rotation and factor rotation, it must produce the same result when applied to $F^{\star\star\star}$ and $Z^{\star\star\star}$, as when it is applied to F and Z.

The first and second moments of $F^{\star\star\star}$ and $Z^{\star\star\star}$ are completely characterised by the maximum Sharpe ratio provided by all assets, and the maximum Sharpe ratio provided by the factors.

How can an evaluation technique depend on something other than the Sharpe ratios achievable with all assets, and with the factors only?

The optimal choice of risk premium also follows, if the two guiding principles are respected.

The beta coefficients of the excess returns on the factors (after the three sets of rotations) are

$$\beta^{\star\star\star} = \begin{bmatrix} I_N \\ \\ 0_{N\times(M-N)} \end{bmatrix}.$$

The predicted excess returns are therefore

$$\hat{\mu}_Z^{\star\star\star} = \begin{bmatrix} \gamma^{\star\star\star} \\ \\ 0_{N\times(M-N)} \end{bmatrix}.$$

But since the transformed excess returns $Z^{\star\star\star}$ have equal variance (one) and are uncorrelated with each other, we know that the optimal choice of $\gamma^{\star\star\star}$ is the one that minimises the sum of squared prediction errors.

We can't do anything about the last $M - N$ assets; the model predicts zero expected excess return for those assets, regardless of the choice of risk premia.

However, the expected excess returns of the first N assets are all matched perfectly if

$$\gamma^{\star\star\star} = \begin{bmatrix} \sqrt{\mu_Z^T \Sigma_{ZZ}^{-1} \Sigma_{ZF} \left(\Sigma_{FZ} \Sigma_{ZZ}^{-1} \Sigma_{ZF} \right)^{-1} \Sigma_{FZ} \Sigma_{ZZ}^{-1} \mu_Z} \\ \\ 0_{(N-1)\times 1} \end{bmatrix}.$$

Working backwards, we find this requires that

$$\gamma^{\star\star} = \left(\Sigma_{FZ}\Sigma_{ZZ}^{-1}\Sigma_{ZF}\right)^{-1}\Sigma_{FZ}\Sigma_{ZZ}^{-1}\mu_Z$$

and

$$\gamma = \Sigma_{FZ}\Sigma_{ZZ}^{-1}\mu_Z = \Psi\mu_Z = \mu_F.$$

So the optimal choice of the risk premia is simply to set the value for each factor equal to the expected excess return of that factor (recalling that the factors are assumed to be traded).

So there is an optimal choice of the risk premia (equal to the expected excess returns of the factors), and goodness-of-fit must be based on the maximum Sharpe ratio achievable with the factors—these results follow from the two guiding principles.

What changes when the factors are not traded?

The exact same transformations can still be applied, but there is a difference—the covariance matrix of the factors is no longer identity. (The means can also be different than the means of the corresponding excess returns.)

We can express the factors as

$$F = \alpha + \Psi Z + \epsilon, \qquad \mathrm{E}\left[\epsilon\right] = 0 \qquad \text{and} \qquad \mathrm{Cov}\left[Z,\epsilon^{T}\right] = 0.$$

A consequence of the above is that

$$\Psi = \Sigma_{FZ}\Sigma_{ZZ}^{-1}.$$

We can instead focus on a model based on the factors

$$F^{\dagger} = \Psi Z.$$

Since these factors are traded, the results derived above apply, the optimal risk premia are

$$\gamma^{\dagger} = \Sigma_{FZ}\Sigma_{ZZ}^{-1}\mu_{Z}.$$

Furthermore, holding the excess returns fixed, the best model is the one with the highest possible Sharpe ratio using the factor portfolios.

We also have

$$\Sigma_{F^{\dagger}F^{\dagger}} = \Sigma_{FZ}\Sigma_{ZZ}^{-1}\Sigma_{ZF}.$$

Beta coefficients are

$$\beta^{\dagger} = \Sigma_{ZF} \left(\Sigma_{FZ} \Sigma_{ZZ}^{-1} \Sigma_{ZF} \right)^{-1} .$$

The expected returns predicted by the alternate factors are

$$\mu_{Z}^{\dagger} = \Sigma_{ZF} \left(\Sigma_{FZ} \Sigma_{ZZ}^{-1} \Sigma_{ZF} \right)^{-1} \Sigma_{FZ} \Sigma_{ZZ}^{-1} \mu_{Z}.$$

Returning to the original model,

$$\Sigma_{FF} = \Sigma_{FZ} \Sigma_{ZZ}^{-1} \Sigma_{ZF} + \Sigma_{\epsilon\epsilon},$$

with

$$\Sigma_{\epsilon\epsilon} \equiv \mathrm{Cov} \left[\epsilon, \epsilon^{T} \right] .$$

(This matrix need not be positive definite, and could even be zero.)

The beta coefficients of the assets with respect to the factors are

$$\beta = \beta^{\dagger}\left[I + \Sigma_{\epsilon\epsilon}\left(\Sigma_{FZ}\Sigma_{ZZ}^{-1}\Sigma_{ZF}\right)^{-1}\right]^{-1},$$

and the expected excess returns predicted by the model are

$$\hat{\mu}_{Z} = \beta^{\dagger}\left[I + \Sigma_{\epsilon\epsilon}\left(\Sigma_{FZ}\Sigma_{ZZ}^{-1}\Sigma_{ZF}\right)^{-1}\right]^{-1}\gamma.$$

The two models, one based on F and the other based on $F^{\dagger}$, make the same predictions if

$$\gamma^{\dagger} = \left[I + \Sigma_{\epsilon\epsilon}\left(\Sigma_{FZ}\Sigma_{ZZ}^{-1}\Sigma_{ZF}\right)^{-1}\right]^{-1}\gamma,$$

or, equivalently, if

$$\gamma = \left[I + \Sigma_{\epsilon\epsilon}\left(\Sigma_{FZ}\Sigma_{ZZ}^{-1}\Sigma_{ZF}\right)^{-1}\right]\gamma^{\dagger}.$$

However, recall the optimal choice of the risk premia for the factors $F^{\dagger}$,

$$\gamma^{\dagger} = \Sigma_{FZ}\Sigma_{ZZ}^{-1}\mu_{Z}.$$

The model based on the factors F therefore makes the same predictions when

$$\gamma = \left[I + \Sigma_{\epsilon\epsilon}\left(\Sigma_{FZ}\Sigma_{ZZ}^{-1}\Sigma_{ZF}\right)^{-1}\right]\Sigma_{FZ}\Sigma_{ZZ}^{-1}\mu_{Z}.$$

This choice of risk premia corresponds with that in Hou and Kimmel (2020), although these authors do not offer a rigorous justification for their choice.

Conclusions

1. Unique and robust method for assignment of risk premia follows from two guiding principles—this method was previously found in literature, but now more rigorously justified.
2. Unique (to within monotonic transformation and "tie-breaking" criterion) method for model evaluation also follows from two guiding principles—based on maximum Sharpe ratio achievable using factor-mimicking portfolios.
3. Methods found in extant literature frequently do not satisfy the two guiding principles.

谢谢!