On Estimation of Risk Premia in Linear Factor Models

Kewei Hou
Fisher College of Business
The Ohio State University*

Robert L. Kimmel
Fisher College of Business
The Ohio State University†

This Version: January 19, 2010

Abstract

We examine theoretical and econometric issues in the estimation of risk premia in a linear factor model, when the model is possibly misspecified. Common empirical methodologies can produce very misleading results. With unspanned factors and possible model misspecification, there are problems not just in estimating the risk premia, but even in defining them unambiguously. We show that, for a given set of test assets, the risk premium of an unspanned factor is very sensitive to the choice of other factors in the model. However, the risk premium of the projection of the unspanned factor onto the asset space is robust to the choice of other factors. The problem is greatly exacerbated in the presence of model misspecification, and can occur even when the unspanned components of the factors are very small (relative to the spanned components). These results highlight the importance of using factor-mimicking portfolios, rather than unspanned factors, in estimation of linear factor models.

JEL Classification: C13, G12.

1 Introduction

Linear factor models of expected returns have played a central role in financial economics for decades. Such models are sometimes motivated by theoretical considerations, as with the CAPM of Sharpe (1964) and Lintner (1965), the ICAPM of Merton (1973), the APT of Ross (1976), and the CCAPM of Breeden (1979). All of these models require that the excess returns of financial assets obey a linear relationship with their exposures to various sources of economic risk. However, linear factor models can also be empirically motivated. For example, the three-factor linear model of Fama and French (1996) includes two factors that do not play any role in traditional economic theories, but whose importance in asset pricing has nonetheless been observed in stock price data. Empirical studies of expected excess returns, such as Black, Jensen, and Scholes (1972), Fama and MacBeth (1973), Fama and French (1992), Jagannathan and Wang (1996), Daniel and Titman (1997), and Daniel, Titman, and Wei (2001) generally focus on either or both of the following questions: (1) What are the factors that explain the cross-section of expected returns of financial assets? and (2) What are the risk premia associated with those factors?

We examine this second question, the assignment and estimation of risk premia in linear factor models for expected excess returns, in the presence of possible misspecification, where “misspecification” means that the

*Columbus, OH 43210-1144. Phone: +1 614 292 0552. E-mail: hou.28@osu.edu.
†Columbus, OH 43210-1144. Phone: +1 614 292 1875. E-mail: kimmel.42@osu.edu.

factors do not fully explain the expected excess returns of all assets. Much of the literature on estimation of risk premia either assumes correct model specification, or focuses on very specific forms of misspecification. The former category includes Shanken (1992), who derives the distribution of risk premia estimates from a two-pass regression method under an assumption that the time series of returns and factor realizations are i.i.d., Kim (1995), who considers the effect of heteroskedasticity on risk premia estimates, Amihud, Christensen, and Mendelson (1992), who offer an alternative to traditional two-pass regression methodologies, and Kan and Zhou (1999), who examine the relative merits of estimation of risk premia (as well as other quantities) by regression methods or GMM methods. In the latter category, Roll and Ross (1994), inspired by the findings of Fama and French (1992), determine when a two-pass regression study is likely to estimate a risk premium of zero for a market factor. Kandel and Stambaugh (1995) consider a misspecified model with a single spanned factor, and find that two-pass OLS regression estimates of risk premia are highly sensitive to economically meaningless repackagings of the assets included in the study; they go on to show that GLS risk premium estimates are invariant to such repackaging. Kan and Zhang (1999b) and Kan and Zhang (1999a) focus on the effect of including "useless" factors, not correlated with any of the asset returns, in regression and GMM estimation procedures, and find that an estimate of the risk premium of the useless factor does not necessarily converge in probability to zero with increasingly long data samples, despite the fact that the factor has nothing to do with asset returns.

These studies focus mainly on econometric issues; by contrast, we find that misspecification makes it impossible even to define the risk premium of an unspanned factor unambiguously, let alone estimate it consistently from a panel of data. A spanned factor has a unique and well-defined risk premium, which can be estimated consistently by any one of several simple econometric techniques. By contrast, the risk premium of an unspanned factor depends on the other factors included in the model, and can be made to take on any arbitrarily specified value by appropriate selection of the other factors. Spanned factors can therefore be assigned a meaningful risk premium, independently of the other factors included in the model and the assets used in the study (provided a change in the set of assets does not cause the factor to cease to be spanned). By contrast, the risk premium assigned to an unspanned factor depends both on the other factors in the model and the assets used in a study, rather than being an intrinsic property of the factor itself. Furthermore, we find that the asymptotic variance of the risk premium estimate for a spanned factor is unaffected by misspecification, whereas the asymptotic variance for the risk premium estimate for an unspanned factor is increasing in the degree of misspecification.

To interpret these results, we note that the risk premium of a spanned factor is based directly on the expected excess returns of the assets which span it. The risk premium of an unspanned factor can be viewed as the sum of the risk premium of its spanned and unspanned components. The spanned component of an unspanned factor (i.e., its projection onto the asset space) is itself a spanned factor; like other spanned factors, its risk premium is the expected excess return of its spanning portfolio. However, there are by definition no assets whose expected excess returns reveal the risk premium of the unspanned component. Rather, this part is determined by extrapolation of the risk premium of the spanned component onto the unspanned component. This extrapolation is highly sensitive to changes in the factor set and in the asset space, and relies on an assumption of correct model specification. The validity of the risk premium of unspanned factors therefore relies on an assumption of model correctness, whereas that of spanned factors does not. This reliance on model correctness can have severe consequences. For example, in a given factor model, a particular factor may have a projection with a risk premium of zero, but a large (positive or negative) risk premium assigned to the unspanned component. If the model is in fact not correctly specified, this large risk premium may turn out to be entirely fictional. Even worse, a factor may have a projection with a positive risk premium, but an unspanned component with an even larger negative imputed risk premium. The net negative risk premium of the factor, based on an assumption of correct model specification, entirely obscures the positive risk premium of the projection, which is robust to misspecification. In addition to the problem of extrapolation, spanned and unspanned factors alike suffer from the problem of misattribution, by which best-fit procedures assign to factors risk premia that rightly belongs to missing factors, not included in the model. In the case of spanned factors, at least, it is possible to correct for misattribution. But we show both theoretically and by example, that for unspanned factors, misattribution and extrapolation together can result in hugely incorrect risk premia estimates.

Empirical studies in general do not distinguish between these two components of factor risk premia, and often interpret the risk premium of an unspanned factor as if it were an intrinsic property of the factor itself, rather than a quantity that is dependent both on the other factors included in the model and on the assets included in the empirical study. We argue that the risk premia of projected factors, being robust to misspecification, may provide more useful information than the usually reported risk premia of the unspanned factors themselves.

The rest of this paper is organized as follows. In Section 2, we examine the assignment of a vector of risk premia to a set of factors, and show that the risk premium of an unspanned factor is highly sensitive to the choice of the other factors included in the model (the problem of extrapolation).In Section 3, we focus on spanned factors, and show that typical best-fit econometric procedures incorrectly attribute risk premia from missing factors to those that are included in the model, sometimes to stunning effect. In Section 4, we offer an interpretation of the results of the previous sections, showing that the assignment of risk premia to unspanned factors effectively assigns shadow expected returns to the unspanned components; the risk premium of a factor then contains a component that is directly observed in the data, and a component that is extrapolated out onto the unspanned components. This latter component, which relies on correct model specification, can dominate the first, which does not. Section 5 considers the problem of estimation of risk premia from data, and finds that misspecification increases the asymptotic variance of estimates of risk premia for unspanned factors, but not for spanned factors. This section also provides an estimator of the risk premia vector for factor projections, and derives its asymptotic variance, which is unaffected by misspecification. Finally, Section 6 concludes.

2 Linear Factor Models and Risk Premia

In this section, we examine issues that arise in the definition of risk premia for factors in correctly specified models, i. e., in models which explain the expected returns of all assets. We find that whether or not a factor is spanned by the assets makes a critical difference in the definition of its risk premium. Specifically, we find that spanned factors always have the same risk premium, regardless of the other factors included in the model, and regardless of the assets the model is expected to price. By contrast, unspanned factors almost never have such a well-defined risk premium; the only exception is an unspanned factor that, by itself, prices all assets correctly; such an unspanned factor has a unique, well-defined risk premium. Any other unspanned factor can have essentially any risk premium at all, depending on the other factors included in the model, and the assets the model is expected to price. These results call into question statements that pervade the literature, of the form "we find that covariation with some macroeconomic variable carries a risk premium of so many basis points". Unless the macroeconomic variable in question is either spanned by the assets, or prices all assets correctly by itself (i. e., without the help of any other factors), such statements are essentially meaningless, as the risk premium of the factor can be any number at all, by appropriate choice of the other factors in the model and/or the assets used to evaluate the model.

2.1 Assumptions and Definitions

Throughout, we assume that both factors and asset excess returns (i. e., the payoffs of assets with an initial cost of zero) satisfy certain technical regularity conditions.

Definition 1. *An M-vector of asset payoffs Z and an N-vector of factor realizations F satisfy the regularity assumptions if all of the following conditions are satisfied:*

(1) Each asset payoff and each factor realization has finite variance.

(2) The $M \times M$ covariance matrix of the asset payoffs, Σ_{ZZ}, has full rank.

(3) The $N \times N$ covariance matrix of the factor realizations, Σ_{FF}, has full rank.

(4) The $N \times M$ covariance matrix between the factor realizations and asset payoffs, Σ_{FZ}, has rank N (which in turn requires $N \leq M$).

Assumption (1) simply ensures that the assets and factors have finite means, variances, and covariances. Assumption (2) ensures that there are no redundant assets, and Assumption (3) does the same for factors. Assumption (4) avoids a multicollinearity condition, in which the loadings of all assets on one factor cannot be strictly proportional to their loadings on another factor. If this assumption were violated, it would be impossible to tell which factor were driving the expected returns of the assets, since a risk premium associated with either factor would affect the asset returns in exactly the same way.

We denote the expected values of the factor realizations and the asset payoffs by μ_F and μ_Z, respectively; the means are necessarily finite if the full rank assumptions are satisfied. The transpose of Σ_{FZ} is denoted by Σ_{ZF}. When a set of M asset payoffs Z and N factor realizations F satisfy the regularity assumptions, there exist an $N \times M$ matrix of constants β and an M-vector of random variables ε such that the following conditions hold:

$$Z = \mu_Z + \beta^T (F - \mu_F) + \varepsilon \qquad \mathrm{E}[\varepsilon] = 0_{M\times 1} \qquad \mathrm{Cov}[F, \varepsilon] = 0_{N\times M}$$

By Assumption (3), which ensures that there are no redundant factors, the covariance matrix of the factors Σ_{FF} has full rank. One can then readily verify that there exists a unique β satisfying the above conditions, and that it is given by $\beta = \Sigma_{FF}^{-1}\Sigma_{FZ}$. By Assumptions (3) and (4), β is also of full rank.

We use the phrase *linear factor model* to mean simply a model that prices all the assets correctly, as per the following definition.

Definition 2. *A set of factor realizations F is called a linear factor model with respect to a set of asset payoffs Z if F and Z satisfy the regularity assumptions, and if:*

$$\mu_Z = \beta^T \gamma = \Sigma_{ZF}\Sigma_{FF}^{-1}\gamma \tag{2.1}$$

for some N-vector γ. The elements of γ are called the risk premia of the corresponding elements of F.

Note that whether a set of factors F is a linear factor model or not may depend on the assets under consideration; a set of factors that price some assets correctly but not others can be considered a linear factor model if attention is restricted only to those assets correctly priced.

(2.1) has the familiar interpretation that each asset has an expected excess return proportional to the exposure of that asset to various sources of economic risk. Linear factor models play a fundamental role in asset pricing theory, and with good reason; for example, the CAPM model of Sharpe (1964) and Lintner (1965) predicts that the market portfolio is a linear factor model for any set of assets; the ICAPM of Merton (1973) and the APT of Ross (1976) are linear factor models with potentially more than one factor. Mathematically, a linear factor model always exists, provided Z has a finite and full rank covariance matrix. For example, for a given Z, we can construct a model with only a single factor $F = \mu_Z^T \Sigma_{ZZ}^{-1} Z$ which tautologically constitutes a linear factor model for Z, with risk premium $\gamma = \mu_Z^T \Sigma_{ZZ}^{-1}\mu_Z$. However, this purely mechanical construction offers no economic insight or intuition; much research in financial economics focuses instead on constructing linear factor models in which the factors have simple intuitive interpretations. For example, Fama and French (1996) construct a three factor model in which the factors relate to firm size, book-to-market ratios, and comovement with the market return.

A major focus of the remainder of the paper is on the following questions:

(I) Does the risk premium of a factor depend on the other factors in the model?

(II) Does the risk premium of a factor depend on the assets used to evaluate the model?

(III) How should the risk premium of a factor be defined when it is *not* part of a linear factor model, i. e., when the factors collectively fail to price some assets correctly?

These questions all pertain to the definition of the risk premia of factors; we are also interested in interpretation of those same risk premia. Two additional questions are:

(I) When are the risk premia of factors *real*, in the sense that they correspond to actual excess returns of assets that can be purchased, as opposed to being simply artificial numbers needed to provide a best fit between expected returns and beta coefficients?

(II) What is the relation between the risk premia of a factor and the importance of the factor in determining expected returns?

The last point is often glossed over in empirical studies. It is routinely assumed that a statistically significant risk premium means a factor is important, and helps explain components of expected returns not explained by other factors in the model; similarly, if a factor has a risk premium estimated close to zero, then that factor is held to be unimportant, and not contributing to explanation of expected returns. We show both by example and by analysis that these two interpretations are completely false, and that the risk premium of of a factor has little to do with its power in explaining expected returns.

2.2 Risk Premium Example

We now consider an estimation procedure that is, if somewhat simplified, similar in spirit to studies commonly found in the finance literature. Specifically, we consider the Fama-French three factor model, and estimate the risk premia of the three factors (denoted by Fama and French as RMRF, SMB, and HML), using a two-pass regression methodology. In the first pass regressions, the time series of returns of each asset are regressed on the time series of the factor realizations, to obtain estimates of the beta coefficients of the assets on the factors:

$$(R_{i,t} - R_{f,t}) = \hat{\alpha}_i + \hat{\beta}_{i,RMRF} RMRF_t + \hat{\beta}_{i,SMB} SMB_t + \hat{\beta}_{i,HML} HML_t + \hat{\epsilon}_{i,t}$$

In the second pass, a cross-sectional regressions of the expected returns of the assets on their beta coefficients is performed:

$$\bar{R}_i = \hat{\gamma}_0 + \hat{\gamma}_{RMRF} \hat{\beta}_{i,RMRF} + \hat{\gamma}_{SMB} \hat{\beta}_{i,SMB} + \hat{\gamma}_{HML} \hat{\beta}_{i,HML} + \hat{\eta}_i$$

We use data from Ken French's website for the period July 1931 (prior to this time, returns for some of the assets are not available) until September 2009. The assets used are the 25 size and book-to-market portfolios. Data on the factor realizations and the risk-free rate of return were obtained from the same website; the factor realizations are already excess returns, but the risk-free rate was subtracted from the returns of each of the 25 assets, to make them excess returns, before running any regressions. For the cross-sectional regression, we use both OLS and GLS, in the latter case with the estimated covariance matrix of the asset excess returns used as the GLS weighting matrix.

As shown in Table 1, the estimated risk premia of $\hat{\gamma}_{SMB}$ and $\hat{\gamma}_{HML}$ are not dramatically different from the sample average returns of the factor portfolios. A stylized fact of the finance literature since the 1990s is that market risk provides no premium in terms of expected return; this result is widely cited and believed, despite its inconsistency with the existence of an apparently positive risk premium for the market portfolio. In this particular sample, the situation is much worse; if we take the $\hat{\gamma}_{RMRF}$ estimated from either the OLS or the

Factor	Cross-sectional Regression OLS	GLS	Sample Average
$\hat{\gamma}_0$	1.95% (0.32%)	1.45% (0.25%)	—
$\hat{\gamma}_{RMRF}$	-1.23% (0.31%)	-0.76% (0.24%)	0.63% (0.18%)
$\hat{\gamma}_{SMB}$	0.20% (0.05%)	0.32% (0.02%)	0.30% (0.11%)
$\hat{\gamma}_{HML}$	0.44% (0.07%)	0.40% (0.02%)	0.44% (0.12%)

Table 1

Two-Pass Regression Results for Fama-French Three-Factor Model—OLS and GLS

This table shows two-pass regression estimates of the risk premia of the three Fama-French factors, using monthly returns of the 25 size and book-to-market portfolios for the period from July 1931 to September 2009. Standard errors are shown in parentheses beneath each point estimate. The cross-sectional regression was done two ways: with OLS, and with GLS, using the estimated covariance matrix of the asset excess returns as the weighting matrix. The table also shows the sample averages of the factor portfolio excess returns, with standard errors. For the regression standard errors, no attempt was made to correct for the errors-in-variables problem that arises because the dependent variables in the cross-sectional regression are the beta coefficients estimated in the time series regressions. As shown, the risk premium associated with RMRF, estimated by either of the regression methodologies, is large and negative, whereas the sample average returns of the RMRF portfolio are large and positive.

GLS regression as the risk premium of the market, that risk premium is large and negative; i. e., exposure to market risk ought to *decrease* expected returns.

This result is grossly inconsistent with the time series properties of the market portfolio. The market has no exposure to any source of risk except itself; in a time series regression of the market portfolio on the three Fama-French factors, it has a loading of one on RMRF and zero on SMB and HML. If the risk premium of the market is actually negative, then the market portfolio itself ought to have an expected rate of return less than the risk-free rate, but the sample average is not only positive, but also statistically significant.

It is interesting to note that, although the three Fama-French factors are traded assets, the above procedure fails to exploit this fact, and treats them exactly the same way an untraded factor would be treated. We therefore augment the 25 size and book-to-market portfolios with the three factor portfolios. First, we conduct some analysis to get an idea of just how much this changes the investment opportunities available to investors. First, we regress the returns of the factor portfolios on the 25 other portfolios, to see how closely the returns of the factor portfolios can be replicated by the 25 portfolios included in the previous analysis:

$$RMRF_t = \hat{\alpha}_{RMRF} + \hat{\beta}_{RMRF,1}R_{1,t} + \ldots + \hat{\beta}_{RMRF,25}R_{25,t} + \hat{\epsilon}_{RMRF,t}$$
$$SMB_t = \hat{\alpha}_{SMB} + \hat{\beta}_{SMB,1}R_{1,t} + \ldots + \hat{\beta}_{SMB,25}R_{25,t} + \hat{\epsilon}_{SMB,t}$$
$$HML_t = \hat{\alpha}_{HML} + \hat{\beta}_{HML,1}R_{1,t} + \ldots + \hat{\beta}_{HML,25}R_{25,t} + \hat{\epsilon}_{HML,t}$$

We do not report all the results of these regressions, but rather only the R^2 statistics, in Table 2. As

Factor	RMRF	SMB	HML
R^2 (constant)	0.9922	0.9764	0.9613
R^2 (no constant)	0.9923	0.9766	0.9616

Table 2

R^2 statistics for Regressions of Fama-French Factors on Size and Book-to-Market Portfolios

This table shows the R^2 statistics from the regression of the time series of the Fama-French factor realizations on the time series of excess returns of the 25 size and book-to-market portfolios. The regressions were performed both with and without a constant term; for the regression without a constant, the R^2 statistic reported is the amount of the second moment, rather than the variance, explained by the independent variables. As shown, most of the variation (or in the case of the no-constant regression, second moment) is explained by the returns of the 25 portfolios. The R^2 statistic for the market portfolio is particularly high.

shown, the component of the three factors not spanned by the 25 portfolios is not particularly big; the R^2 statistics are all above 0.95, and for RMRF, it is above 0.99. An investor who could only trade in the 25 size and book-to-market portfolios would nonetheless be able to replicate the returns of the three factors fairly closely, by holding simple static portfolios of the 25 assets.

Another way to measure how much the investment opportunity set changes when the 25 size and book-to-market portfolios are augmented by the three factor portfolios, is to calculate point estimates of the Sharpe ratio that can be realized by the 25 assets, and by the 28 assets. In sample, these ratios are 0.3216 and 0.3338. So adding additional assets increases the Sharpe ratio that can be realized (as it must), but not by a very large amount; the increase is less than 4% of the Sharpe ratio estimated from the 25 assets. So by either measure, the R^2 statistics from Table 2, or the Sharpe ratios, it seems the addition of the three factor portfolios does not change the investment opportunity set by a large amount.

However, if we repeat the regressions of Table 1, but using the 28 assets instead of the 25, the results are dramatically different. Table 3 shows the results using both the original set of 25 assets, and the augmented set of 28 assets. As shown, the estimated risk premium of RMRF, using either OLS or GLS, changes very dramatically, and is much closer to the sample average excess return of RMRF. The rather small change in the test assets used to run the regressions has resulted in an extreme change in the estimated risk premium of RMRF.

The later sections examine analytically how it is possible for a relatively small change in the test assets to make such a large difference in the estimated risk premia. Two phenomena are at work here: *misattribution*, in which risk premia driven by factors that are missing from a model, are incorrectly attributed to factors that are included, and *extrapolation*, in which the unspanned components of factors are assigned risk premia for which there is no direct evidence in asset returns. Both of these phenomena are studied in detail in the subsequent sections.

2.3 Model Selection Example

We know examine the problem of model selection, by way of example, and show that a commonly used heuristic can be highly misleading. For this example, we use the same 25 size and book-to-market portfolios as test

Factor	Original Regression OLS	Original Regression GLS	Augmented Regression OLS	Augmented Regression GLS	Sample Average
$\hat{\gamma}_0$	1.95% (0.32%)	1.45% (0.25%)	0.22% (0.13%)	0.01% (0.01%)	—
$\hat{\gamma}_{RMRF}$	-1.23% (0.31%)	-0.76% (0.24%)	0.44% (0.13%)	0.62% (0.01%)	0.63% (0.18%)
$\hat{\gamma}_{SMB}$	0.20% (0.05%)	0.32% (0.02%)	0.16% (0.07%)	0.29% (0.01%)	0.30% (0.11%)
$\hat{\gamma}_{HML}$	0.44% (0.07%)	0.40% (0.02%)	0.40% (0.09%)	0.42% (0.01%)	0.44% (0.12%)

Table 3

Two-Pass Regression Results for Fama-French Three-Factor Model—OLS and GLS with Augmented Assets

This table shows two-pass regression estimates of the risk premia of the three Fama-French factors, using monthly returns of the 25 size and book-to-market portfolios for the period from July 1931 to September 2009, with both the original set of 25 size and book-to-market portfolios, and the augmented set of assets, consisting of those same 25 portfolios, but also the three factor portfolios. Standard errors are shown in parentheses beneath each point estimate. The cross-sectional regression was done two ways: with OLS, and with GLS, using the estimated covariance matrix of the asset excess returns as the weighting matrix. The table also shows the sample averages of the factor portfolio excess returns, with standard errors. For the regression standard errors, no attempt was made to correct for the errors-in-variables problem that arises because the dependent variables in the cross-sectional regression are the beta coefficients estimated in the time series regressions. As shown, the risk premia estimates for RMRF change very dramatically once the factor portfolios are included as assets as well; the estimated risk premium, using either OLS or GLS, is much closer to the sample average excess return of RMRF.

assets, but instead of the three Fama-French factors, use two-factor models, with contrived factors (i. e., the factors were deliberately engineered to bring out the points being emphasized).

Table 4 shows the results of the two-pass regressions; note that the OLS and GLS results are the same, the estimated $\hat{\gamma}_0$ coefficients for both models using both techniques are zero, and the standard errors are all zero. All of these phenomena occur because both models explain expected returns perfectly in sample. (If the errors-in-variables problem were tackled rather than ignored, the standard errors would not be zero.) Note that the first model has two factors, each with a positive risk premium; in the second model, the first factor has a positive risk premium, but the second factor has a risk premium of zero. A conventional interpretation of these results would be that, in the first model, both factors are needed to explain the expected returns, because both risk premia are positive (and infinitely statistically significant), whereas in the second model, the second factor is not needed to explain expected returns, since it has a risk premium of zero.

Both of these conclusions are intuitively appealing, and both of them are false. Table 4 shows the results of the two-pass regressions, but using only the first factor in each model this time. As shown, the first model still fits all expected returns perfectly, even though a factor with a statistically significant risk premium has been removed; in the second model, the fit has deteriorated, despite the fact that the factor that was removed had a risk premium of zero.

What drives both of these results is correlation between the two factors. When the second factor is removed, the loadings on the first factor change. In the first model, the change in loading on the first factor changes

Factor	First Model		Second Model	
	OLS	GLS	OLS	GLS
$\hat{\gamma}_0$	0.00% (0.00%)	0.00% (0.00%)	0.00% (0.00%)	0.00% (0.00%)
$\hat{\gamma}_{F_1}$	2.60% (0.00%)	2.60% (0.00%)	1.53% (0.00%)	1.53% (0.00%)
$\hat{\gamma}_{F_2}$	0.93% (0.00%)	0.93% (0.00%)	0.00% (0.00%)	0.00% (0.00%)

Table 4

Two-Pass Regression Results for Contrived Two-Factor Models—OLS and GLS

This table shows two-pass regression estimates of the risk premia on two different two-factor models, using monthly returns of the 25 size and book-to-market portfolios for the period from July 1931 to September 2009. Standard errors are shown in parentheses beneath each point estimate. The cross-sectional regression was done two ways: with OLS, and with GLS, using the estimated covariance matrix of the asset excess returns as the weighting matrix. The table also shows the sample averages of the factor portfolio excess returns, with standard errors. For the regression standard errors, no attempt was made to correct for the errors-in-variables problem that arises because the dependent variables in the cross-sectional regression are the beta coefficients estimated in the time series regressions. As shown, both models explain expected returns (at least in sample) perfectly; the OLS and GLS results are the same, the γ_0 coefficients are zero, and the standard errors are all zero. Note that the risk premium for each factor is positive in the first model, but the second model has the first factor with a positive risk premium, and the second factor with a zero risk premium.

the risk premium for each asset driven by the first factor, in an amount that exactly offsets the loss of the risk premium driven by the second factor. In the second model, the loadings on the first factor change, and therefore the risk premia driven by the first factor as well; but the loss of the second factor does not offset these changes, because the risk premium of the second factor was zero. Since the model fit the data perfectly before, and the expected returns predicted by the model change when the second factor is removed, the second model, upon removal of the second factor, fits the data less well.

2.4 Invariant Risk Premia

In some cases, it is possible to assign a unique risk premium to a factor, regardless of the other factors in a model or the assets used to evaluate the model. This section derives results on two cases.

We use the notation $\gamma(F, Z)$ to denote the risk premia vector when either the set of factors or the set of assets under consideration is not clear from the context. The vector of risk premia can be expressed in terms of the moments of the asset payoffs and factor realizations, as per the following Lemma.

Lemma 1. *If a set of factor realizations F is a linear factor model for a set of asset payoffs Z, then the risk premia vector γ is unique and is given by:*

$$\gamma = \gamma(F, Z) = \Sigma_{FF} \left(\Sigma_{FZ}\Sigma_{ZZ}^{-1}\Sigma_{ZF}\right)^{-1} \Sigma_{FZ}\Sigma_{ZZ}^{-1}\mu_Z \tag{2.2}$$

Proof: See Appendix.

It is a brief exercise to verify that if F is a linear factor model for Z, it is also a linear factor model for any other set of assets Y that spans a subspace of the space spanned by Z (provided Y and F satisfy the regularity

Factor	First Model		Second Model	
	OLS	GLS	OLS	GLS
$\hat{\gamma}_0$	0.00% (0.00%)	0.00% (0.00%)	– 0.18% (0.09%)	– 0.15% (0.20%)
$\hat{\gamma}_{F_1}$	2.60% (0.00%)	2.60% (0.00%)	1.23% (0.10%)	1.63% (0.28%)

Table 5

Two-Pass Regression Results for Contrived One-Factor Models—OLS and GLS

This table shows two-pass regression estimates of the risk premia on two different one-factor models, using monthly returns of the 25 size and book-to-market portfolios for the period from July 1931 to September 2009. Standard errors are shown in parentheses beneath each point estimate. The cross-sectional regression was done two ways: with OLS, and with GLS, using the estimated covariance matrix of the asset excess returns as the weighting matrix. The table also shows the sample averages of the factor portfolio excess returns, with standard errors. For the regression standard errors, no attempt was made to correct for the errors-in-variables problem that arises because the dependent variables in the cross-sectional regression are the beta coefficients estimated in the time series regressions. The single factor for each model is simply the first factor from the models analyzed in Table 4. As shown, the first model continues to fit the data perfectly, despite removal of a factor with a statistically significant risk premium, and the second model fits the data less well, even though the factor that was removed had a risk premium of zero.

assumptions), with $\gamma(F,Y) = \gamma(F,Z)$. It is also straightforward to verify that if F is a linear factor model for Z and G is a set of factors that spans the same space as F, then G is also a linear factor model for Z.

Although we have defined risk premia for any set of factors and any set of assets, as long as they satisfy the regularity assumptions and the factors price the assets correctly, we will be interested in determining when the risk premium assigned to an individual factor is invariant to the additional or removal of other factors, and also when this risk premium is invariant to changes in the asset space. When a factor is spanned by the assets, (2.2) always assigns the same risk premium to that factor, regardless of the other factors included. This result is an immediate consequence of an argument presented in Shanken (1992). However, it is possible to prove a more general result, as per the following theorem.

Theorem 1. *Let F and G be two sets of factor realizations, and let Z be a set of asset payoffs, such that F and Z satisfy the regularity assumptions, and G and Z also satisfy the regularity assumptions. For any vectors x, y, and z such that:*

$$x^T(F-\mu_F) = y^T(G-\mu_G) = z^T(Z-\mu_Z)$$

the risk premia vectors $\gamma(F,Z)$ and $\gamma(G,Z)$ satisfy:

$$x^T\gamma(F,Z) = y^T\gamma(G,Z) = z^T\mu_Z = x^T\Gamma_F^T\mu_Z = y^T\Gamma_G^T\mu_Z$$

Proof: See Appendix.

Theorem 1 states that any spanned factor, or any spanned linear combination of factors, must have the same risk premium, regardless of the other factors included in the model. Furthermore, the risk premium of a spanned factor (or spanned linear combination of factors) is equal to the expected excess return of its projection onto the asset space. Since this value does not depend on the other factors included in a factor set, it is meaningful to speak of the risk premium of a spanned factor, independently of the other factors needed to form a linear factor model.

We also note that, in an economy free of arbitrage opportunities, the risk premium of a spanned factor is invariant to changes in the asset space (provided the factor does not cease to be spanned as a result of the change). If the same factor is spanned by two different sets of assets, its projections onto the two asset spaces must have the same expected excess return, and therefore the same risk premium as specified by Theorem 1. The risk premium of a spanned factor is therefore not only invariant to the other factors included in a model, but also to the particular choice of spanning assets. Furthermore, if a factor is itself a traded asset [e.g., the excess return of the market portfolio in the CAPM of Sharpe (1964) and Lintner (1965)], since it spans itself, its risk premium is simply equal to its expected excess return.[1] The well-known finding of Fama and French (1992), that the risk premium of the market portfolio is approximately zero or slightly negative during a sample period when the market portfolio experienced large positive returns, can thus be called into question purely on theoretical grounds; the risk premium of the market portfolio in any well-specified linear factor model must be equal to its expected excess return. This phenomenon is further studied by Roll and Ross (1994).

The next theorem shows that, if F is a linear factor model for Z, then the risk premia assigned to the elements of F are invariant to the introduction of additional factors, provided the new factors do not lead to a violation of the regularity assumptions. Such factors, are, of course, unnecessary,[2] since F is already a linear factor model, but an econometrician most likely does not know a priori which factors are necessary and which are not. The risk premia assigned are also invariant to the removal of factors, provided the remaining factors still constitute a linear factor model.

Theorem 2. *Let F be linear factor model for Z, and let G be a set of factors such that G and Z satisfy the regularity assumptions. If there exists a matrix Ψ such that $(F - \mu_F) = \Psi^T (G - \mu_G)$, then G is a linear factor model for Z and $\gamma(F, Z) = \Psi^T \gamma(G, Z)$.*

Proof: See appendix.

In Theorem 2, G is a set of factors that effectively includes F, and possibly some other new factors as well. The risk premia for the factors in F are unaffected by the introduction of the new factors, provided that the new factors do not result in a violation of the regularity assumptions. Of particular note is the case where F contains a single factor. Since, by assumption, F is a linear factor model, that single factor must explain all asset returns, and its risk premium is defined in Lemma 1. Any other factor is therefore unnecessary. However, if other factors are introduced, the risk premium of the original factor remains unchanged. The single factor contained in F therefore has an unambiguous and well-defined risk premium.

This section has therefore described two cases where a factor has a unique, well-defined risk premium:

[1] This point is made by Shanken (1992), who in two-pass regression estimation constrains the risk premium of a spanned factor (but not necessarily spanned linear combinations of factors) to be equal to the expected return of its spanning portfolio. As we shall see, this constraint is unnecessary provided an appropriate GLS weighting matrix is used; in this case, the results of the estimation procedure automatically satisfy the constraint. See, for example, Kandel and Stambaugh (1995), who consider two-pass regression estimation of a model with a single spanned factor.

[2] Kan and Zhang (1999a) and Kan and Zhang (1999b) discuss GMM and regression estimation, respectively, in models with "useless" factors, where "useless" is defined as being uncorrelated with any of the assets under study. Such factors are specifically precluded here, since they result in violations of the regularity assumptions. Rather, the factors considered here are still correlated with the asset excess returns (from the full rank assumptions), but are considered unnecessary because, after their removal, the remaining factors are still a linear factor model.

(i) The factor is spanned by the assets.

(ii) The factor is a linear factor model all by itself, i. e., it explains all asset returns without the help of any other factors.

In these two cases, it makes sense to speak of the risk premium of the factor. One may wonder whether these results can be extended to a broader class of factors; as shown in the next section, the answer is a resounding "no". If a factor does not satisfy either of these two conditions (i. e., it is not spanned, and it does not explain all expected returns by itself), then its risk premium is not even a well-defined concept. Such a factor can have any risk premium at all, by appropriate introduction of other factors into the model.

2.5 Extrapolation

To understand the problem with assigning a unique risk premium to an unspanned factor, we express such a factor as the sum of a spanned and unspanned component.

The β matrix describes the projection of the assets Z onto the space of factors F; we shall often wish to refer to the reverse projection, of the factors F onto the space of assets Z. Under the regularity assumptions, there exist an $M \times N$ matrix of constants Γ and an N-vector of random variables η, such that the following conditions hold:

$$F = \mu_F + \Gamma^T \left(Z - \mu_Z\right) + \eta \qquad \mathrm{E}\left[\eta\right] = 0_{Nx1} \qquad \mathrm{Cov}\left[Z, \eta\right] = 0_{M\times N} \tag{2.3}$$

The matrices β and Γ are related by the following identities:

$$\beta = \Sigma_{FF}^{-1} \Gamma^T \Sigma_{ZZ} \qquad \Gamma = \Sigma_{ZZ}^{-1} \beta^T \Sigma_{FF}$$

from which it follows that Γ is of rank N. We denote by P the projection of the factors onto the asset space:

$$P = \Gamma^T Z$$

and denote by Σ_{PP} and Σ_{PZ} the covariance matrix of the projection P, and the covariance between the projection P and the assets Z. We refer to a factor F_i, with $1 \leq i \leq N$, as "spanned" if $\mathrm{Var}\left[\eta_i\right] = 0$ (i. e., a spanned factor may differ from its projection only by a constant). Similarly, a linear combination of factors $w^T F$ is considered spanned if $\mathrm{Var}\left(w^T \eta\right) = 0$. From the definition of the projection P:

$$\Sigma_{PZ} = \Sigma_{FZ} \qquad \Sigma_{PP} = \Sigma_{FZ} \Sigma_{ZZ}^{-1} \Sigma_{ZF}$$

The β coefficients of the assets on F and P (denoted by β_F and β_P, respectively) are related as follows:

$$\beta_F = \Sigma_{FF}^{-1} \Sigma_{FZ} = \left(\Sigma_{PP} + \Sigma_{\eta\eta}\right)^{-1} \Sigma_{FZ} = \left(I + \Sigma_{PP}^{-1} \Sigma_{\eta\eta}\right)^{-1} \Sigma_{PP}^{-1} \Sigma_{PZ} = \left(I + \Sigma_{PP}^{-1} \Sigma_{\eta\eta}\right)^{-1} \beta_P$$

It follows immediately that F is a linear factor model for Z, if and only if P is a linear factor model for Z, and the risk premia of the two are related by:

$$\gamma(F,Z) = \Sigma_{FF}\Sigma_{PP}^{-1}\Gamma^T\mu_Z = \Sigma_{FF}\Sigma_{PP}^{-1}\gamma(P,Z) = \underbrace{\gamma(P,Z)}_{\text{Risk premia of spanned components}} + \underbrace{\Sigma_{\eta\eta}\Sigma_{PP}^{-1}\gamma(P,Z)}_{\text{Extrapolation Component}} \tag{2.4}$$

When factors are not spanned by assets, the risk premia of those factors can be expressed as the sum of two components. The first is the risk premium of factor-mimicking portfolios. This component is an actual risk premium that an investor can realize by purchasing the appropriate securities. The second component is, in effect, an extrapolation of the risk premium of the spanned components of the factors (i. e., the factor mimicking portfolios) to the unspanned components. This component does not correspond to an actual risk premium that can be experienced by investors; by definition, investors cannot purchase the unspanned components of the factors, without the introduction of additional assets to complete the market. Furthermore, even if such assets are introduced, the risk premia of the unspanned components may well be very different numbers than what is assigned by (2.4). The risk premia assigned by (2.4) are numbers that result in a perfect linear relation between expected returns and beta coefficients on the factors; by assumption, the model prices all assets correctly. However, to interpret the risk premia from (2.4) as real risk premia that investors can experience by purchasing assets requires the assumption that, first, additional assets exist that complete the market, and second, that the model prices those additional assets correctly.

We refer to this phenomenon as *extrapolation*; even in a model that fits expected returns perfectly, if the factors are unspanned, some component of the risk premia assigned may not be the result of extrapolation from the spanned components of the factors to the unspanned components, rather than real risk premia that investors can experience by holding assets.

From (2.4), we see that two linear factor models with the same projection onto the asset space do not necessarily have the same vector of risk premia. The following theorem shows how much the vector of risk premia can vary for different models with the same projection, i. e., how bad the extrapolation component can be.

Theorem 3. *Let Z be a vector of M asset payoffs, and let F be a set of N factors such that F is a linear factor model for Z with projection P onto the asset space. The vector of risk premia $\gamma(F,Z)$ is either within the open half-space of $\mathbb{R}^N$ defined by:*

$$\gamma(P,Z)^T\,\Sigma_{PP}^{-1}\gamma(F,Z) > \gamma(P,Z)^T\,\Sigma_{PP}^{-1}\gamma(P,Z)$$

or equal to the point $\gamma(P,Z)$ (which lies on the boundary of the half-space). Conversely, let $P = \Gamma^T Z$ be a set of N factors such that P is a linear factor model for Z. Given any N-vector γ that satisfies either:

$$\gamma(P,Z)^T\,\Sigma_{PP}^{-1}\gamma > \gamma(P,Z)^T\,\Sigma_{PP}^{-1}\gamma(P,Z)$$

or $\gamma = \gamma(P,Z)$, there exists a linear factor model F for Z with N factors, such that F has the projection P onto the asset space, and $\gamma = \gamma(F,Z)$.

Proof: See Appendix.

As previously noted, the vector of risk premia for a given set of factors F and a given set of assets Z is unique and well-defined whenever F and Z satisfy the regularity assumptions, and in particular when F is

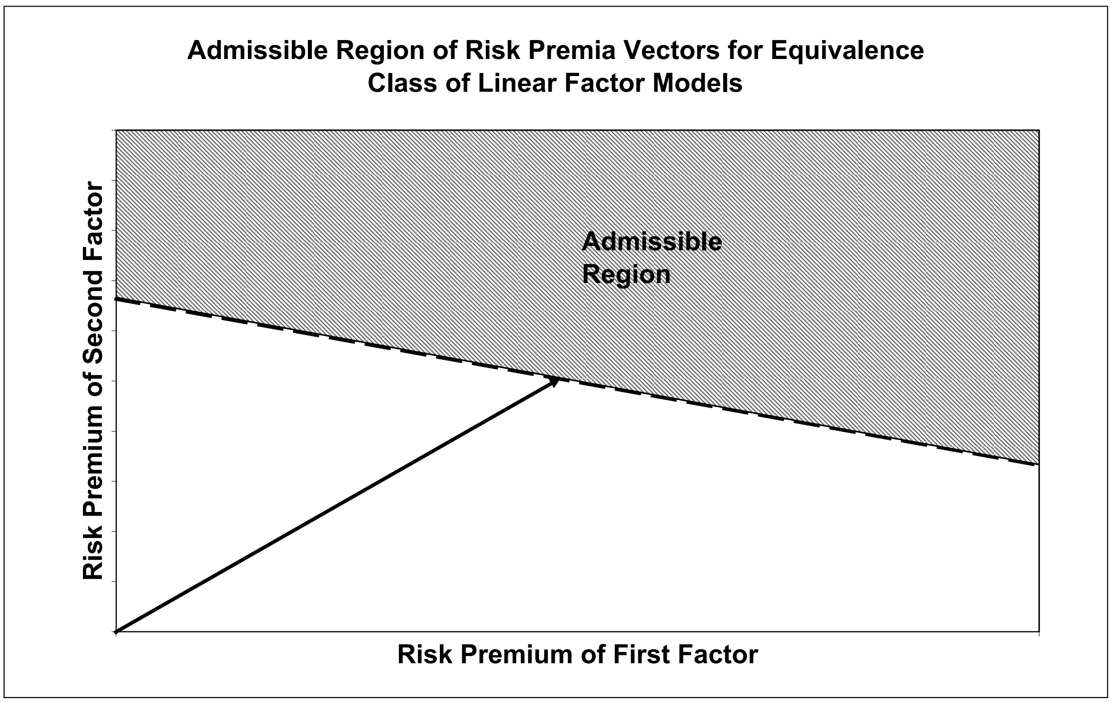

Figure 1
This figure shows a stylized representation of the admissible region for the vector of risk premia for a two factor model. The arrow shows the location of $\gamma(P, Z)$, which lies on the boundary of the half space. Any linear factor model for the assets Z with the projection P has a vector of risk premia within the labeled admissible region. The only point on the boundary that is included in the admissible region is $\gamma(P, Z)$ itself. Conversely, for any given point γ within the admissible region, there exists a linear factor model with projection P onto the assets with risk premia vector equal to γ.

a linear factor model for Z. However, varying the unspanned components of F (while holding the spanned components fixed) can result in assignment of a nearly arbitrary vector of risk premia to F. As shown in Theorem 3, models with the same projection onto the asset space can have very different risk premia; any point within an open half-space is achievable by varying the unspanned components associated with each factor. Figure 1 provides a graphical illustration of the subset of $\mathbb{R}^2$ that can be achieved in a two-factor model by adding unspanned components to the factors. The point where the arrow meets the boundary of the half-space is the risk premium vector of the spanned components of the two factors. However, by adding noise (uncorrelated with any asset returns) to the factors, it is possible to make the risk premia of the two (now unspanned factors) be anywhere in the half-space. It is not possible to assign arbitrary values to the two factors simultaneously (e. g., it is not possible to make both have a risk premium of zero), but considerable flexibility remains nonetheless. Either of the two factors can be made, for example, to have a risk premium of zero; the risk premium of either factor can be made to be any arbitrarily large negative number, provided the risk premium of the other factor is a sufficiently large positive number. This flexibility is driven by choice of the $\Sigma_{\eta\eta}$ matrix in (2.4). In the general case, the risk premia vector of a model with M unspanned factors can be made to vary completely arbitrarily in $M-1$ dimensions, and some flexibility still remains in the one other dimension.

The extrapolation component of the risk premium vector assigned by (2.4) can therefore be extreme; for example, in an economy with relatively modest risk premia offered by the traded assets, the risk premia of explanatory factors could still be very large, due to the presence of unspanned components in the factors.

The previous result holds the spanned components of all factors fixed, and asks how the risk premia of the factors change as unspanned components are added to those factors. In the next result, we examine how the risk premium of an unspanned factor can depend on the other factors in the model. Provided the first factor does not explain all assets all by itself, its risk premium can take on any arbitrary value at all, depending on the other factors.

Theorem 4. *Let Z be a vector of M asset payoffs, and let F be a vector of N factor realizations, such that F and Z satisfy the regularity assumptions, but such that F is not a linear factor model for Z. Let P be the projection of F onto the assets Z. Let γ be any N-vector such that $w^T\left[\gamma-\gamma\left(P,Z\right)\right]=0$ for any N-vector w with $w^T\Sigma_{\eta\eta}w=0$. Then there exists a linear factor model H for Z such that $\left(F-\mu_F\right)=\Psi^T\left(H-\mu_H\right)$ for some matrix Ψ, and such that $\gamma=\Psi^T\gamma\left(H,Z\right)$.*

Proof: See Appendix.

Theorem reffactorExpansionTheorem stands in Sharpe contrast to Theorem 1. In the latter case, it is shown that the risk premia assigned to spanned factors are completely invariant to the addition of new factors, so it makes sense to refer to the "risk premium" of such a factor without any additional context. By contrast, unspanned factors can be made to have any risk premia at all (provided they do not already explain all asset returns) by appropriate introduction of other factors. For example, if a set of factors is not a linear factor model, and all of the factors are unspanned, then appropriate choice of an additional factor that completes the model (i. e., turns it into a linear factor model that prices all assets correctly), can make the risk premia of the existing factors all equal to zero. Choice of a different factor that completes the model can make the risk premia of the existing factors all equal to some arbitrary and very large number.

Theorem reffactorExpansionTheorem therefore makes it clear that results of the previous section cannot be extended to any broader class of factors; if an individual factor (or a set of factors) is unspanned and fails to explain the expected returns of all assets, then the "risk premia" of those factors are not well-defined. They can be any arbitrary numbers at all; a set of factors that is included in one linear factor model, can have a completely different risk premium vector than when the exact same set of factors is included in a different linear factor model. As in the punch line of many lawyer jokes, the risk premium of an unspanned factor can be "anything you want it to be".

3 Misspecified Models

The previous section focuses on definition of risk premia in correctly specified linear factor models, and finds several key results. First, the risk premium of a spanned factor is invariant to the other factors included in a model. Second, the risk premium of an unspanned factor that by itself prices all assets correctly, also has a well-defined risk premium that does not depend on the other factors in the model. Third, an unspanned factor that does not price all assets by itself can have any arbitrary risk premium at all, depending on which other factors it is bundled together with. This is a result of the extrapolation of the risk premia of spanned components of factors (which are well-defined) to the unspanned components.

3.1 Spanned Factors in Misspecified Models—Misattribution

The results of the previous section generally apply to linear factor models (i. e., models that price all assets correctly), and ignore any estimation issue. In practice, we may not have identified a set of factors that price all assets correctly, so that our model is misspecified; furthermore, even if we have identified a linear factor model, we may be ignorant of the fact because of estimation error, or sampling variation in the data. We may nonetheless be interested in determining, as best as possible under the circumstances, what the risk premia of the factors in a misspecified model are.

The question of how to define the risk premia of factors in a model that is (possibly) misspecified is hardly ever addressed. A typical methodology is to use an econometric procedure that may or may not converge asymptotically to the risk premia defined by (2.2) if the model is correctly specified; whatever numbers that econometric procedure produces are then interpreted as the "risk premia" of the factors, without stating clearly how exactly risk premia are defined in a misspecified model. Rather than defining the quantities sought and then developing an econometric procedure designed to estimate those quantities as accurately as possible given the noise in the data, the econometric procedure is essentially allowed to define the quantity being sought when the model is misspecified.

If the factors are spanned, there is a relatively simple and unambiguous way to define the risk premia of such factors in a model which is (or which may be) misspecified. From the previous section, we know that the risk premium of a spanned factor must be equal to the expected excess return of the factor-mimicking portfolio, in any correctly specified model. There would therefore seem little reason to define it as anything else in a potentially misspecified model; if we define the risk premium of a spanned factor as something other than the expected excess return of the factor mimicking-portfolio in a potentially misspecified model, then the risk premium found will have to be revised as soon as we discover the additional factors that are currently missing from the model. And yet the finance literature is full of examples of methodologies that assign to spanned factors something different than the factor-mimicking portfolio excess returns; for example, in the previous section, a two-pass regression methodology was applied to the Fama-French factors and test portfolios four different ways, and in all cases, produced risk premia estimates that were at least slightly different (and sometimes dramatically different) than the sample average of the factor returns.

Such a result generally follows by applying some best-fit procedure (usually OLS or GLS regression) to the expected returns and factor loadings. Although regression is an econometric technique, it is here applied to a problem that is only partly econometric. There may fail to be a perfect linear relation between expected returns and factor loadings because of sampling variation in the data. However, there may also fail to be a perfect linear relation because of model misspecification; even if we have an extremely long data sample, in which the expected returns and factor loadings of all assets can be estimated extremely accurately, the fit between the two will not be perfect, necessitating some kind of best-fit procedure, if the model is missing some necessary factors. In this case, a procedure designed to solve econometric problems is instead used to solve a model misspecification problem.

The result is often what we call *misattribution*. Restricting attention to spanned factors, the (in-sample) expected return of assets can be expressed as:

$$\bar{Z}_i = \hat{\alpha}_i + \hat{\beta}_i^T \bar{F} \tag{3.1}$$

The $\hat{\beta}$ coefficients are the estimated loadings of the assets on the factors; the $\hat{\alpha}$ coefficients are the estimated amount of each asset's average return that is not explained by the factors. In a correctly specified model, the true α coefficients are zero, but the estimated $\hat{\alpha}$ coefficients can still fail to be zero because of estimation error.

If we denote the average (across all assets) estimated beta and estimated expected excess return as $\bar{\hat{\beta}}$ and $\bar{\bar{Z}}$, respectively, then the risk premia estimated by a best-fit procedure such as regression are:

$$\hat{\gamma} = \left(\bar{\hat{\beta}}\Omega^{-1}\bar{\hat{\beta}}^T\right)^{-1}\bar{\hat{\beta}}\Omega^{-1}\bar{\bar{Z}} \tag{3.2}$$

where Ω^{-1} is a GLS weighting matrix. By substituting (3.1) into (3.2), we find:

$$\hat{\gamma} = \underbrace{\bar{F}}_{\text{Risk premium of spanned factor}} + \underbrace{\left(\bar{\hat{\beta}}\Omega^{-1}\bar{\hat{\beta}}^T\right)^{-1}\bar{\hat{\beta}}\Omega^{-1}\hat{\alpha}}_{\text{Misattribution component}}$$

The first component on the right-hand side is the sample average of the returns of the factor portfolio; we know that the risk premia of the factors in a correctly specified model must be the expected excess returns, so the sample averages of those returns seems like a reasonable way to estimate the population value.

We refer to the second component on the right-hand side as the *misattribution* component. There are two reasons the $\hat{\alpha}$ vector might be different than zero; one is sampling variation, and the other is model misspecification. In either case, a best-fit procedure misattributes the effect of sampling variation and/or model misspecification to the risk premia of the factors included in the model.

To see this, suppose for simplicity that we have a very long data series, and the expected returns and factor loadings are all estimated extremely precisely. Then any non-zero $\hat{\alpha}$ can only be due to model misspecification, i. e., risk premia from factors that are missing from the model. This non-zero vector is then multiplied by a projection matrix, and the result attributed to factors that are included in the model. If, for example, assets with a high loading on one of the factors also have high loading on a factor that is missing from the model, then the best-fit procedure wrongly attributes the risk premium from the missing factor to the factor that is included in the model.

This misattribution is a consequence of use of a best-fit procedure. The estimated risk premia do indeed provide the best-fit between the expected returns and the factor loadings, but are different from the values that we know, from theoretical results, they must have. Use of a best-fit procedure in cross-sectional regressions where the independent variables are factor loadings is overfitting by construction.

3.2 Fama-French Example Revisited

At this point, it is worth revisiting the numeric example of the previous section, with the Fama-French three-factor model and the 25 size and book-to-market portfolios.

The two-pass cross-sectional regression procedure was applied four different ways. OLS and GLS were used, and the 25 assets were used alone, and also augmented by the three factor portfolios. The Fama-French factors are traded assets that investors can buy, but when the 25 assets are used as the test assets, the factors are treated as if they are unspanned. In other words, the procedure fails to take advantage of the fact that the factors are traded; if the factors were instead three untraded macroeconomic variables, then the procedure could still be carried out in exactly the same way. Therefore, when only the 25 (instead of the 28) test assets are used, the factors are effectively unspanned. An investor who is constrained to purchase only the 25 size and book-to-market portfolios cannot perfectly replicate the returns of the three factor portfolios. By contrast, when the 28 test assets are used, the factors are spanned.

Since extrapolation arises only when factors are not spanned, only the misattribution issue arises when the 28 test assets are used. Table 3 shows the regression results. The misattribution effect is quite small when GLS is used, but much larger when OLS is used. The difference between the estimated risk premia using GLS and the sample average excess return is approximately one basis point per month; for OLS, the difference is between four and nineteen basis points for the three factors. The use of GLS instead of OLS, as recommended by, for example, Kandel and Stambaugh (1995), would seem to be well-supported by this example.

The situation is quite different when only the 25 test assets are used. For these results, the factors are effectively treated as unspanned, so both misattribution and extrapolation can occur. The combined effect of these two phenomena for the RMRF factor are huge; the estimated risk premium differs from the sample average by 186 basis points when OLS is used, and 139 basis points when GLS is used. The combined effect of misattribution and extrapolation produce a risk premium estimate that is vastly different from what a simple theoretical argument tells us it must be.

3.3 Unspanned Factors

Throughout the remainder, it is necessary to define the risk premia of factors, not necessarily spanned, in a model that may be misspecified. For unspanned factors, we lack the type of invariance result we have for spanned factors (i. e., the risk premium of spanned factors does not depend on the other factors included in the model), and know from the previous section that no such result is possible. We therefore use the following definition, but recognize that there is a certain degree of arbitrariness in it.

Definition 3. *Let F be a set of factors, and let Z be a set of assets such that F and Z satisfy the regularity assumptions. Then the vector of risk premia of F with respect to Z is defined by* (2.4)*, whether or not F is a linear factor model for Z.*

When F is a linear factor model for Z, the motivation for defining the risk premia vector in this way is clear; with this choice, the factors describe the expected excess returns of all assets perfectly. When F is not a linear factor model for Z, simply applying the definition from (2.4) may seem somewhat arbitrary. However, a researcher will not necessarily know a priori whether a particular set of factors is a linear factor model for a given set of assets, and statistical tests to that point may well prove inconclusive. It is therefore useful to have a definition that can be applied in the face of uncertainty, but that is consistent with (2.4) when F is a linear factor model for Z. Furthermore, as shown in Section 5, this definition is equivalent to that obtained by a two-pass GLS regression technique. The risk premia vector γ minimizes the distance between the predicted and actual vectors of expected excess returns, if distance is measured with respect to the matrix Σ_{ZZ}^{-1}:

$$\gamma = \underset{\gamma_0 \in \mathbb{R}^N}{\operatorname{argmin}} \left(\mu_Z - \beta^T \gamma_0\right)^T \Sigma_{ZZ}^{-1} \left(\mu_Z - \beta^T \gamma_0\right)$$

This definition is invariant to the linear transformations of the asset space considered by Kandel and Stambaugh (1995); for choices of the distance matrix other than Σ_{ZZ}^{-1}, this is not necessarily the case. Finally, if the vector μ_Z is estimated from a time series of i.i.d. observations of Z, its covariance matrix is given by $\frac{1}{n}\Sigma_{ZZ}$, with n equal to the number of observations. The inverse of this matrix is therefore a reasonable choice of distance matrix. Both Shanken (1992) and (in the case of a single spanned factor) Kandel and Stambaugh (1995) point out the desirable econometric features of using Σ_{ZZ}^{-1} as a GLS weighting matrix in the second pass when estimating γ by two-pass linear regression.

With a risk premia vector defined whether or not a set of factors is a linear factor model for a given set of assets, it is convenient to define the deviations of the actual expected excess returns of a set of assets from the fitted expected excess returns:

$$\varepsilon\left(F, Z\right) = \mu_Z - \beta^T \gamma\left(F, Z\right)$$

A set of factors F is thus a linear factor model for a set of M assets Z if and only if F and Z satisfy the regularity assumptions and $\varepsilon\left(F, Z\right) = 0_{M \times 1}$.

4 Shadow Returns

One might be interested in unspanned factors because markets are incomplete, or because it is impossible or impractical to identify enough assets to span a particular set of factors. Focusing on the latter case, we consider the "unspanned" components of a set of factors to be in fact spanned, but by assets not observed by the econometrician. For a set of factors F and a set of assets Z that satisfy the regularity assumptions, let $K \leq N$ be the rank of $\Sigma_{\eta\eta}$. We assume the existence of a set of asset payoffs Y, with finite and full-rank covariance matrix Σ_{YY} and mean μ_Y, such that:

$$(\eta - \mu_\eta) = \Psi^T (Y - \mu_Y)$$

for some $K \times N$ matrix Ψ with rank K. It follows that:

$$\begin{aligned} \Sigma_{\eta\eta} &= \Psi^T \Sigma_{YY} \Psi \\ \Sigma_{YF} &= \Sigma_{Y\eta} = \Sigma_{YY} \Psi \end{aligned}$$

Y and Z are uncorrelated (since η and Z are uncorrelated), so F and the union of Z and Y satisfy the regularity assumptions. Since the factors F are spanned by Z and Y, we have:

$$\gamma_0 = \gamma\left(F, \begin{bmatrix} Z \\ Y \end{bmatrix}\right) = \begin{bmatrix} \Gamma \\ \Psi \end{bmatrix}^T \begin{bmatrix} \mu_Z \\ \mu_Y \end{bmatrix} = \Gamma^T \mu_Z + \Psi^T \mu_Y$$

This value is the solution of the minimization problem:

$$\gamma_0 = \underset{\gamma \in \mathbb{R}^N}{\operatorname{argmin}} \left[\left(\mu_Z - \beta_Z^T \gamma\right)^T \Sigma_{ZZ}^{-1} \left(\mu_Z - \beta_Z^T \gamma\right) + \left(\mu_Y - \beta_Y^T \gamma\right)^T \Sigma_{YY}^{-1} \left(\mu_Y - \beta_Y^T \gamma\right) \right]$$

where $\beta_Z = \Sigma_{FF}^{-1} \Sigma_{FZ}$ and $\beta_Y = \Sigma_{FF}^{-1} \Sigma_{FY}$. However, since the expected excess returns of the assets Y are not observed, we can only calculate the vector of risk premia for F using the assets Z, as per (2.4). If we set the two risk premium definitions (one requiring knowledge of μ_Y, and therefore infeasible) equal to each other, we find:

$$\begin{aligned} \gamma\left(F, \begin{bmatrix} Z \\ Y \end{bmatrix}\right) &= \gamma(F, Z) \\ \Gamma^T \mu_Z + \Psi^T \mu_Y &= \Gamma^T \mu_Z + \Sigma_{\eta\eta} \Sigma_{PP}^{-1} \Gamma^T \mu_Z \\ \mu_Y &= \left(\Psi \Psi^T\right)^{-1} \Psi \Sigma_{\eta\eta} \Sigma_{PP}^{-1} \Gamma^T \mu_Z = \Sigma_{YY} \Psi \Sigma_{PP}^{-1} \Gamma^T \mu_Z \end{aligned} \tag{4.1}$$

We can therefore interpret the assignment of risk premia to F when μ_Z is observed (but μ_Y is not) as assigning shadow expected excess returns to the unspanned components of F (i.e., to Y). From (4.1), these shadow values are unique, whether or not F is a linear factor model for Z. If we attempt to predict the expected excess returns of the assets Y using the risk premia vector $\gamma(F, Z)$, we find:

$$\beta_Y^T \gamma(F, Z) = \Sigma_{YF} \Sigma_{FF}^{-1} \gamma(F, Z) = \Sigma_{YY} \Psi \Sigma_{PP}^{-1} \Gamma^T \mu_Z$$

Note that the expected excess returns predicted by this equation are the same as the shadow prices assigned by (4.1), whether F is a linear factor model for Z or not. If F is a linear factor model for Z, and the actual expected excess returns of Y are equal to the values predicted by (4.1), then F is a linear factor model for the union of Z and Y (F may fail to be a linear factor model for Y, since there is no guarantee that F and Y satisfy the full rank assumptions).

Of course, if the actual value of μ_Y is something different than the shadow value specified in (4.1), then the unobservability of Y results in a different (and incorrect) assignment of a vector of risk premia to F. Since the econometrician who does not observe the assets Y has no information concerning the true value of μ_Y, other than perhaps a theoretically motivated belief that a particular set of factors constitutes a linear factor model, it is not clear that the shadow values assigned are meaningful. One possible way to characterize the shadow expected excess returns is in terms of the mean-variance efficient portfolios formed from P and Y. By premultiplication of both sides of (4.1) by $(Y-\mu_Y)^T\,\Sigma_{YY}^{-1}$, we find:

$$(Y-\mu_Y)^T\,\Sigma_{YY}^{-1}\mu_Y = (Y-\mu_Y)^T\,\Psi\Sigma_{PP}^{-1}\Gamma^T\mu_Z = (\eta-\mu_\eta)^T\,\Sigma_{PP}^{-1}\Gamma^T\mu_Z$$

If the shadow value of μ_Y is in fact the true value, then $Y^T\Sigma_{YY}^{-1}\mu_Y$ is the mean-variance efficient portfolio formed from the assets Y, and $P^T\Sigma_{PP}^{-1}\Gamma^T\mu_Z$ is the mean-variance efficient portfolio formed from the assets P. In other words, the linear combination of factors $F^T\Sigma_{PP}^{-1}\Gamma^T\mu_Z$ has a projection onto Z of $P^T\Sigma_{PP}^{-1}\Gamma^T\mu_Z$, and a projection onto Y of $(\eta-\mu_\eta)^T\,\Sigma_{PP}^{-1}\Gamma^T\mu_Z$. If the shadow value of μ_Y is equal to the true (but unobserved) expected excess returns of Y, then both of these projections are mean-variance efficient within the spaces spanned by Z and Y, respectively. The risk premia assigned to F therefore implies shadow values of μ_Y that ensure the same linear combination of factors with a mean-variance efficient projection onto P, also has a mean-variance efficient projection onto Y.

The shadow values of μ_Y, of course, do not necessarily bear any relation to the true values. Consider a set of assets Z and a set of spanned factors $P=\Gamma^T Z$, and a set of assets Y, uncorrelated with P. It is possible to construct a set of N factors F with projection P onto Z, such that the shadow value of μ_Y is equal to any arbitrarily chosen non-zero value; if the rank of Ψ is less than N, then any arbitrarily chosen value of μ_Y, including zero, is achievable.

Theorem 5. *Let Z be a set of M assets, and let $P=\Gamma^T Z$ be a set of N factors spanned by Z, such that Z and P satisfy the full rank assumptions. Let Y be a set of $K\leq N$ assets with a finite and full rank covariance matrix Σ_{YY}, uncorrelated with Z. Let μ_0 be any K-vector. If $K<N$ or $\mu_0\neq 0$, then there exists a set of N factors F with projection P onto Z, such that the shadow value of μ_Y from* (4.1) *is equal to μ_0.*

Proof: See Appendix.

If the allegedly unspanned components of a set of factors are in fact spanned, then, as per the results of the previous sections, the risk premia vector (under the assumption that all assets are observed) is unique, well-defined, and invariant to the addition or removal of assets or other factors, so long as the regularity assumptions are satisfied (and as long as the removal of an asset does not cause a previously spanned factor to become unspanned). By contrast, the risk premia definition of (2.4) implicitly assigns to the "unspanned" components a set of shadow expected excess returns that results in the closest fit of the assets Z to the factors F, without regard to the true value of μ_Y (which is unobserved). Theorem 5 shows how arbitrary this assignment can be; essentially any shadow value can be assigned simply by repackaging the spanned and unspanned components of the factors. Use of "unspanned" factors, which are in fact spanned by other assets not included in a study, therefore tends to result in overfitting; unless the shadow value of μ_Y is exactly equal to the true value, the risk premia assigned to a set of factors F by (2.4) predicts expected excess returns for the assets Z that are a closer fit to the observed expected returns than those predicted by the true risk premia.

Empirical studies generally do not distinguish between the two types of risk premia, i.e., that which is due to μ_Z and that which is due to μ_Y. Since the assignment of risk premia to μ_Y is very sensitive to misspecification, recognition of such a distinction would be useful; reporting the risk premia of factor projections, in addition to or instead of the risk premia of the factors themselves, would be a useful guide to the reliability of the risk premia. The risk premia of the projected factors, as already noted, are given by:

$$\gamma\left(P,Z\right)=\Gamma^{T}\mu_{Z}$$

Considering the factors collectively, the maximum squared Sharpe ratio offered by the factors (treating them as if they were traded assets) is given by:

$$\begin{aligned}\gamma^{T}\left(F,Z\right)\Sigma_{FF}^{-1}\gamma\left(F,Z\right)&=\mu_{Z}^{T}\Gamma\Sigma_{PP}^{-1}\Sigma_{FF}\Sigma_{PP}^{-1}\Gamma^{T}\mu_{Z}\\&=\mu_{Z}^{T}\Gamma\Sigma_{PP}^{-1}\Gamma^{T}\mu_{Z}+\mu_{Z}^{T}\Gamma\Sigma_{PP}^{-1}\Sigma_{\eta\eta}\Sigma_{PP}^{-1}\Gamma^{T}\mu_{Z}\end{aligned}\tag{4.2}$$

By contrast, the maximum squared Sharpe ratio offered by the factor projections (treating them as if they were traded assets) is given by:

$$\gamma^{T}\left(P,Z\right)\Sigma_{PP}^{-1}\gamma\left(P,Z\right)=\mu_{Z}^{T}\Gamma\Sigma_{PP}^{-1}\Gamma^{T}\mu_{Z}\tag{4.3}$$

It is a fairly common practice to examine the difference between this latter quantity and the maximum squared Sharpe ratio offered by the assets (given by $\mu_{Z}^{T}\Sigma_{ZZ}^{-1}\mu_{Z}$); typical GMM formulations and the method of Gibbons, Ross, and Shanken (1989) test the hypothesis that this difference is zero (in which case the factors are a linear factor model). However, the difference between the quantities in (4.2) and (4.3) is rarely, if ever, examined; nor is the difference between $\gamma\left(F,Z\right)$ and $\gamma\left(P,Z\right)$. Although there would appear to be little point in performing a statistical test to see if the two are different, the difference provides insight into how much of the risk premia assigned to a set of factors is directly observed in the asset returns, and how much is the result of extrapolation (which relies on model correctness). The risk premia $\gamma\left(P,Z\right)$ may be viewed as the product of a policy of conservatism, reflecting only risk premium for which the asset returns provide direct evidence. The risk premium of $\gamma\left(F,Z\right)$, by contrast, may be viewed as the result of a policy of optimism, in which possibly large risk premia, for which there are no direct evidence, are attributed to unobserved assets in an effort to fit the data better. Should the model turn out to be misspecified, much of the risk premia assigned to the factors by $\gamma\left(F,Z\right)$ may turn out to be fictional. By contrast, the risk premia assigned to the projection factors $\gamma\left(P,Z\right)$ are robust to misspecification.

The next section considers the problem of estimating risk premia of various types of factors, including the projections of unspanned factors.

5 Estimation

We now consider the problem of estimation of risk premia for a set of factors from a time series of observations of the factor realizations and asset payoffs. Such a problem is considered by Black, Jensen, and Scholes (1972) and by Fama and MacBeth (1973), who employ a two-pass regression method; Shanken (1992) derives the asymptotic distribution of the vector of risk premia estimated by such a method, but under an assumption that the factors constitute a linear factor model for the assets, and Kim (1995) modifies this analysis to account for heteroskedasticity. However, the combination of unspanned factors and possible misspecification adds some additional complexity to the analysis.

We take the estimation approach of replacing mean vectors and covariance matrices in the definition of the risk premia vector by their sample counterparts. This approach is straightforward, and also produces the same estimates as a two-pass regression technique, provided an appropriate GLS weighting matrix is used. The β coefficients for a set of assets can be estimated in a first pass regression of the time series of Z onto F:

$$\widehat{\beta} = \widehat{\Sigma}_{FF}^{-1}\widehat{\Sigma}_{FZ}$$

The risk premia estimates can then be estimated in a second pass regression of $\widehat{\mu}_Z$ on $\widehat{\beta}$:

$$\widehat{\gamma}(F,Z) = \left(\widehat{\beta}\widehat{\Sigma}_{ZZ}^{-1}\widehat{\beta}^T\right)^{-1}\left(\widehat{\beta}\widehat{\Sigma}_{ZZ}^{-1}\widehat{\mu}_Z\right)$$

where the GLS weighting matrix $\widehat{\Sigma}_{ZZ}^{-1}$ is used. If the joint distribution of F and Z is such that the sample estimates of Σ_{FF}, Σ_{FZ}, Σ_{ZZ}, and μ_Z are consistent, then:

$$\begin{aligned}
\operatorname{plim}\widehat{\gamma}(F,Z) &= \left(\beta\Sigma_{ZZ}^{-1}\beta^T\right)^{-1}\left(\beta\Sigma_{ZZ}^{-1}\mu_Z\right) \\
&= \left(\Sigma_{FF}^{-1}\Sigma_{PP}\Sigma_{FF}^{-1}\right)^{-1}\left(\Sigma_{FF}^{-1}\Sigma_{FZ}\Sigma_{ZZ}^{-1}\mu_Z\right) \\
&= \Sigma_{FF}\Sigma_{PP}^{-1}\Gamma^T\mu_Z = \gamma(F,Z)
\end{aligned}$$

This method of estimation is therefore consistent (provided, as stated above, the estimates of Σ_{FF}, Σ_{FZ}, Σ_{ZZ}, and μ_Z are consistent). If F is a linear factor model for Z, then it is consistent with other choices of weighting matrix as well, although in this case the results of the second regression are subject to the small sample problems discussed in Kandel and Stambaugh (1995). Furthermore, we note that no constant is included in the second regression, since we are dealing with excess returns rather than returns. If a constant is included in this regression, the results are subject to the Kandel and Stambaugh (1995) small sample problem even if the GLS weighting matrix Σ_{ZZ}^{-1} is used. In the case of spanned factors, although the two-pass regression estimator with GLS weighing matrix of $\widehat{\Sigma}_{ZZ}^{-1}$ is consistent, is is unnecessarily complicated; in this case, the risk premia are $\gamma = \Gamma^T\mu_Z$, which is more simply estimated by regressing the time series of factor realizations on the asset returns, and then multiplying the estimated coefficients by an estimate of μ_Z. We simply take the approach of replacing the matrices Σ_{FF}, Σ_{FZ}, Σ_{ZZ}, and μ_Z in the definition of the risk premia vector by their sample counterparts, since both of the two regression schemes described above yield the same result anyway.

First, we consider the case of spanned factors. The vector of risk premia is then:

$$\gamma(F, Z) = \Gamma^T \mu_Z = \Sigma_{FZ} \Sigma_{ZZ}^{-1} \mu_Z$$

Since the factors are spanned, the matrix Γ can be estimated without error; the coefficients of Γ are the coefficients of a regression of F onto Z, as per (2.3), but in which the variance of each element of η is zero. The only variation in estimation is therefore due to variation in the estimation of μ_Z, which can be estimated by its sample mean:

$$\widehat{\gamma}(F, Z) = \widehat{\Sigma}_{FZ} \widehat{\Sigma}_{ZZ}^{-1} \widehat{\mu}_Z = \widehat{\Gamma}^T \widehat{\mu}_Z = \Gamma^T \widehat{\mu}_Z \tag{5.1}$$

Under an assumption that the observations of Z are i.i.d. distributed, the variance of this estimator has a particularly simple form.

Theorem 6. *Let Z be a set of assets and let $P = \Gamma^T Z$ be a set of factors such that P and Z satisfy the regularity assumptions. Let $\widehat{\gamma}(P, Z)$ be the estimate of $\gamma(P, Z)$, as per* (5.1). *If the time series of observations of F and Z is i.i.d., then $\widehat{\gamma}(P, Z)$ is a consistent estimate of $\gamma(P, Z)$, with asymptotic variance given by:*

$$AVar\left[\widehat{\gamma}(P, Z)\right] = \Sigma_{PP}$$

Proof: See Appendix.

The asymptotic variance of an individual factor does not depend on the other factors included in the factor set; furthermore, as long as the factors are spanned, the asymptotic variance of the risk premia vector is completely invariant to the choice of the spanning assets Z. This result can be viewed as a multivariate generalization of the result of Kandel and Stambaugh (1995), who consider a single spanned factor and expected returns (rather than expected excess returns) and find that the risk premium estimate differs from the zero-beta rate by the expected return of the factor's spanning portfolio when the GLS weighting matrix is used. Since we consider expected excess returns, the zero-beta rate is equal to zero, and the risk premia for factors are thus the expected excess return of the factors' spanning portfolios. No other characteristic of the assets or factors is relevant for the determination of the factor risk premia.

Before considering the fully general case of factors with unspanned components, we examine estimation of the risk premia of the projection P of a set of (not necessarily spanned) factors F onto the assets Z. If F and Z satisfy the regularity assumptions, then so do P and Z, and if F is a linear factor model for Z, then so is P. However, in this case, the coefficients Γ are estimated with error. The risk premia of the projections of the factors are given by:

$$\gamma(P, Z) = \Gamma^T \mu_Z = \Sigma_{FZ} \Sigma_{ZZ}^{-1} \mu_Z$$

This value can be estimated by:

$$\widehat{\gamma}(P, Z) = \widehat{\Gamma}^T \widehat{\mu}_Z = \widehat{\Sigma}_{FZ} \widehat{\Sigma}_{ZZ}^{-1} \widehat{\mu}_Z \tag{5.2}$$

The uncertainty in the estimation of Γ leads to a larger asymptotic variance than when P is observed directly.

Theorem 7. *Let F be a set of factors and Z a set of assets, such that F and Z satisfy the regularity assumptions. Let P be the projection of F onto Z. Let $\widehat{\gamma}(P,Z)$ be the estimate of $\gamma(P,Z)$, as per* (5.2). *If the time series of observations of P and Z is i.i.d. with a multivariate Gaussian distribution, then $\widehat{\gamma}(P,Z)$ is a consistent estimate of $\gamma(P,Z)$, with asymptotic variance given by:*

$$AVar\left[\widehat{\gamma}(P,Z)\right] = \left(\mu_Z^T \Sigma_{ZZ}^{-1} \mu_Z\right) \Sigma_{\eta\eta} + \Sigma_{PP}$$

Proof: See Appendix.

Of course, Theorem 6 may be viewed as a special case of Theorem 7. The presence of unspanned components necessarily increases the asymptotic variance of the risk premia estimates, unless the expected excess return of every asset is equal to zero. Note that the asymptotic variance of the risk premia vector depends only on the covariance matrices of the spanned and unspanned components, and on the maximum squared Sharpe ratio offered by the assets Z. As in the case of Theorem 6, the asymptotic variance of the risk premium of an individual factor does not depend on the other factors included in the factor set. In particular, if an individual factor has no unspanned component, the estimate is not affected by the presence of unspanned components in other factors included in the factor set.

In the fully general case of unspanned factors, the risk premia vector is given by:

$$\gamma(F,Z) = \Sigma_{FF} \Sigma_{PP}^{-1} \gamma(P,Z) = \Sigma_{FF} \left(\Sigma_{FZ} \Sigma_{ZZ}^{-1} \Sigma_{ZF}\right)^{-1} \Sigma_{FZ} \Sigma_{ZZ}^{-1} \mu_Z$$

In addition to the variation due to the estimation of $\gamma(P,Z)$, the result is premultiplied by $\Sigma_{FF}\Sigma_{PP}^{-1}$, which must also be estimated. The estimate is therefore given by:

$$\widehat{\gamma}(F,Z) = \widehat{\Sigma}_{FF} \left(\widehat{\Sigma}_{FZ} \widehat{\Sigma}_{ZZ}^{-1} \widehat{\Sigma}_{ZF}\right)^{-1} \widehat{\Sigma}_{FZ} \widehat{\Sigma}_{ZZ}^{-1} \widehat{\mu}_Z \tag{5.3}$$

If F is a linear factor model for Z, then the asymptotic variance of $\widehat{\gamma}(F,Z)$ takes on a particularly simple form.

Theorem 8. *Let F be a linear factor model for Z, and let $\widehat{\gamma}(F,Z)$ be the estimate of $\gamma(F,Z)$, as per* (5.3). *If F and Z are i.i.d. with a multivariate Gaussian distribution, then $\widehat{\gamma}(F,Z)$ is a consistent estimate of $\gamma(F,Z)$, with asymptotic variance given by:*

$$AVar\left[\widehat{\gamma}(F,Z)\right] = \Sigma_{FF} + \left[1 + \gamma^T(F,Z) \Sigma_{FF}^{-1} \gamma(F,Z)\right] \left(\Sigma_{\eta\eta} + \Sigma_{\eta\eta} \Sigma_{PP}^{-1} \Sigma_{\eta\eta}\right)$$

The asymptotic variance can also be expressed in terms of μ_Z instead of $\gamma(F,Z)$:

$$AVar\left[\widehat{\gamma}(F,Z)\right] = \Sigma_{FF} + \left[1 + \mu_Z^T \Gamma \Sigma_{PP}^{-1} \Gamma^T \mu_Z + \mu_Z^T \Gamma \Sigma_{PP}^{-1} \Sigma_{\eta\eta} \Sigma_{PP}^{-1} \Gamma^T \mu_Z\right] \left(\Sigma_{\eta\eta} + \Sigma_{\eta\eta} \Sigma_{PP}^{-1} \Sigma_{\eta\eta}\right)$$

Proof: See Appendix.

A similar expression appears in Shanken (1992), who considers the estimation of risk premia by two-pass regression when dealing with expected returns (as opposed to expected excess returns). The incremental variance incurred by choosing F rather than P as the set of factors is given by:

$$AVar[\widehat{\gamma}(F,Z)] - AVar[\widehat{\gamma}(P,Z)] = \left[1 + \gamma^T(F,Z)\Sigma_{FF}^{-1}\gamma(F,Z)\right]\left(\Sigma_{\eta\eta}\Sigma_{PP}^{-1}\Sigma_{\eta\eta}\right) + \left[2 + \gamma^T(F,Z)\Sigma_{FF}^{-1}\Sigma_{\eta\eta}\Sigma_{FF}^{-1}\gamma(F,Z)\right]\Sigma_{\eta\eta}$$

or, expressed in terms of μ_Z instead of $\gamma(F,Z)$:

$$AVar[\widehat{\gamma}(F,Z)] - AVar[\widehat{\gamma}(P,Z)] = \left(1 + \mu_Z^T\Gamma\Sigma_{PP}^{-1}\Gamma^T\mu_Z + \mu_Z^T\Gamma\Sigma_{PP}^{-1}\Sigma_{\eta\eta}\Sigma_{PP}^{-1}\Gamma^T\mu_Z\right)\left(\Sigma_{\eta\eta}\Sigma_{PP}^{-1}\Sigma_{\eta\eta}\right) + \left(2 + \mu_Z^T\Gamma\Sigma_{PP}^{-1}\Sigma_{\eta\eta}\Sigma_{PP}^{-1}\Gamma^T\mu_Z\right)\Sigma_{\eta\eta}$$

The difference is clearly positive as long as $\Sigma_{\eta\eta}$ is not equal to zero. The presence of unspanned components causes $\widehat{\gamma}(F,Z)$ to have larger variance than $\widehat{\gamma}(P,Z)$, even if $\gamma(F,Z) = \gamma(P,Z)$. Although the variance of the risk premium estimate for an individual spanned factor is unaffected by the inclusion of unspanned components in other factors, it should also be noted that the asymptotic variance of the risk premium estimate of an individual factor with an unspanned component is, in contrast with the two previous cases, dependent on the other factors included in the factor set.

When F is not a linear factor model for Z, there are additional sources of variance.

Theorem 9. *Let F be a set of factors and Z be a set of assets such that F and Z satisfy the regularity assumptions, and let $\widehat{\gamma}(F,Z)$ be the estimate of $\gamma(F,Z)$, as per (5.3). If F and Z are i.i.d. with a multivariate Gaussian distribution, then $\widehat{\gamma}(F,Z)$ is a consistent estimate of $\gamma(F,Z)$, with asymptotic variance given by:*

$$AVar[\widehat{\gamma}(F,Z)] = \Sigma_{FF} + \left[\Sigma_{\eta\eta} + \Sigma_{\eta\eta}\Sigma_{PP}^{-1}\Sigma_{\eta\eta}\right]\left(1 + \mu_Z^T\Sigma_{ZZ}^{-1}\mu_Z + \mu_Z^T\Gamma\Sigma_{PP}^{-1}\Sigma_{\eta\eta}\Sigma_{PP}^{-1}\Gamma^T\mu_Z\right) + \left[\Sigma_{\eta\eta}\Sigma_{PP}^{-1}\Sigma_{\eta\eta} + \Sigma_{\eta\eta}\Sigma_{PP}^{-1}\Sigma_{\eta\eta}\Sigma_{PP}^{-1}\Sigma_{\eta\eta}\right]\left(\mu_Z^T\Sigma_{ZZ}^{-1}\mu_Z - \mu_Z^T\Gamma\Sigma_{PP}^{-1}\Gamma^T\mu_Z\right)$$

Proof: See Appendix.

To distinguish between the two versions of the asymptotic variance of $\widehat{\gamma}(F,Z)$ (derived under differing assumptions about the asset expected excess returns), we denote the asymptotic variance from Theorem 8 by $AVar_{LF}$, since it is derived under the assumption that F is a linear factor model for Z. The difference between the two asymptotic variances (from Theorems 9 and 8) is then given by:

$$AVar[\widehat{\gamma}(F,Z)] - AVar_{LF}[\widehat{\gamma}(F,Z)] = \begin{bmatrix} \Sigma_{\eta\eta} + 2\Sigma_{\eta\eta}\Sigma_{PP}^{-1}\Sigma_{\eta\eta} \\ +\Sigma_{\eta\eta}\Sigma_{PP}^{-1}\Sigma_{\eta\eta}\Sigma_{PP}^{-1}\Sigma_{\eta\eta} \end{bmatrix} \varepsilon^T(F,Z)\Sigma_{ZZ}^{-1}\varepsilon(F,Z) \tag{5.4}$$

The right-hand side is always positive whenever $\Sigma_{\eta\eta} \neq 0$ and $\varepsilon(F,Z) \neq 0$. Holding the covariance matrices and $\gamma(F,Z)$ fixed, (5.4) indicates the degree to which misspecification (i.e., that F is not a linear factor model for Z) affects the variance of the estimate of $\gamma(F,Z)$. The greater the misspecification (i.e., the larger the magnitude of $\varepsilon^T(F,Z)\Sigma_{ZZ}^{-1}\varepsilon(F,Z)$), the larger is the asymptotic variance of the risk premia vector.

It is worth noting a few points on hypothesis testing. A common practice is to test whether a given factor is "priced," in the sense that it has a risk premium different from zero. When risk premia are estimated using a two-pass regression technique, the t-statistic of a risk premium estimate is often used to determine whether the corresponding factor is priced; Chen, Kan, and Zhang (1999) offer a criticism of this approach based on the statistical properties of the risk premia estimates. However, quite apart from any econometric issues, this approach is conceptually flawed. If a factor is spanned, its risk premium is equal to the expected excess return of the spanning portfolio. This result does not depend on what other factors are included in a factor set, whether those factors constitute a linear factor model for the assets, etc. For example, given a set of $M > 1$ assets Z, consider the factors F defined as:

$$F^T = \mu_F^T + (Z - \mu_Z)^T \Sigma_{ZZ}^{-1} \begin{bmatrix} \mu_Z & w \end{bmatrix}$$

where w is an M-vector with $w \neq \mu_Z k$ for any constant k. It is a brief exercise to verify that F is indeed a linear factor model for Z, and that the risk premia of the two factors are $\mu_Z^T \Sigma_{ZZ}^{-1} \mu_Z$ and $\mu_Z^T \Sigma_{ZZ}^{-1} w$, respectively. However, one quickly verifies that the first factor is by itself a linear factor model for Z, with the same risk premium. If $\mu_Z^T \Sigma_{ZZ}^{-1} w \neq 0$, then the second factor has a non-zero risk premium despite its non-importance, and a consistent estimation technique allows an econometrician, given enough data, to reject the hypothesis that the risk premium of this second factor is zero. Despite this finding that the factor is "priced," it has absolutely no relevance, given the other factors in the model, for asset pricing.

Similarly, let w be an M-vector such that $\mu_Z^T \Sigma_{ZZ}^{-1} w = 0$ and $w \neq \mu_Z k$ for any value of k. Then define the factors:

$$F^T = (Z - \mu_Z)^T \Sigma_{ZZ}^{-1} \begin{bmatrix} \mu_Z - w & w \end{bmatrix}$$

One also readily verifies that this set of factors is a linear factor model for Z. However, if the second factor, which by definition has a risk premium of zero, is removed, the remaining factor is not a linear factor model. One can in fact make the fit of the first factor arbitrarily bad (i.e., the value of $\varepsilon^T (F, Z) \Sigma_{ZZ}^{-1} \varepsilon (F, Z)$ is arbitrarily large) simply by choosing w such that $w^T \Sigma_{ZZ}^{-1} w$ is large enough. For sufficiently large w, the first factor does a very poor job predicting the expected excess returns of the assets, but when the second factor (with its risk premium of zero) is added, the model predicts the expected excess returns of all assets perfectly. However, conventional statistical tests will fail to reject the hypothesis that the risk premium of the second factor is zero (since the hypothesis is true) 95% of the time. Despite its importance in asset pricing, the second factor is not "priced."

Whether the risk premium of a spanned factor is equal to zero or not has essentially nothing to do with its importance in a model for expected excess returns. As per the results of earlier sections, holding a set of assets fixed, the risk premium of an unspanned factor is not even uniquely defined, unless the factor is a linear factor model by itself. However, even if the risk premium of a particular factor is "correct," in the sense that it is spanned by additional assets (not included in the study) with actual expected excess returns equal to the shadow values, the use of t-statistics of risk premia, or similar techniques, in model selection is fundamentally flawed, because it tests the wrong hypothesis. An appropriate hypothesis to test is not whether the risk premium of a factor is equal to zero, but whether the removal of the factor from the model results in a statistically significant change in the fitted values of the expected excess returns of the assets.

6 Conclusion

We have examined the problem of definition and estimation of risk premia for linear factor models in the presence of misspecification. The results show that factors that are spanned by the assets under study can be assigned risk premia that are invariant to the choice of other factors included in the model, and also to the choice of assets under study; furthermore, model misspecification does not increase the asymptotic variance of estimates of such risk premia. By contrast, it is impossible even to define unambiguously the risk premium of an unspanned factor in the presence of model misspecification; such a risk premium is dependent on the other factors included in a model, and also on the assets under study. Furthermore, even holding the set of factors fixed, misspecification leads to larger asymptotic variance for estimates of the risk premia of unspanned factors. The risk premium of a spanned factor can be considered an intrinsic property of the factor itself, as it is invariant to the model in which the factor is placed and the assets used to span it. By contrast, the risk premium of an unspanned factor is only meaningful in the context of the other factors in the model and the assets included in a study.

If it is not possible or practical to extend the set of assets included in a study so that all factors are spanned, then a researcher may nonetheless avoid the hard consequences of potential misspecification on factor risk premia by focusing instead on the risk premia of the projections of the factors onto the asset space. These projections, being spanned factors themselves, are completely immune to the troubles associated with unspanned factors, although, since the projections onto the asset space are estimated with error, the asymptotic variance of the risk premia are larger than they would be if the projections were estimated without error. However, the asymptotic variances of the risk premia estimates of the projections are still smaller than the asymptotic variances of the risk premia estimates for the factors themselves. If a model is possibly misspecified, distinguishing between the risk premia of the factors and their projections provides an indication of how much of a factor risk premium is due to extrapolation, which is sensitive to misspecification, rather than direct observation, which is not. The risk premia of the projections are completely robust to misspecification, and may be viewed as the result of a policy of conservatism, reflecting only risk premium for which there is direct evidence from asset returns. By contrast, the risk premia of unspanned factors are highly sensitive to misspecification, and may be viewed as the result of a policy of optimism, reflecting extrapolated components of risk premium for which there is no direct evidence in asset return data. The robustness of an empirical study can thus be improved by reporting the risk premia of factor projections, in addition to or instead of the risk premia of unspanned factors.

References

Amihud, Y., B. J. Christensen, and H. Mendelson (1992): "Further Evidence on the Risk-Return Relationship," Working paper, New York University and Stanford University.

Black, F., M. Jensen, and M. Scholes (1972): "The Capital Asset Pricing Model: Some Empirical Results," in *Studies in the Theory of Capital Markets*, ed. by M. Jensen, pp. 79–121. Praeger.

Breeden, D. (1979): "An Intertemporal Asset Pricing Model with Stochastic Consumption and Investment Opportunities," *Journal of Financial Economics*, 7, 265–296.

Chen, N.-F., R. Kan, and C. Zhang (1999): "A Critique of the Use of T-Ratios in Model Selection," Working paper, Hong Kong University of Science and Technology and University of Toronto.

Daniel, K., and S. Titman (1997): "Evidence on the Characteristics of Cross-Sectional Variation in Stock Returns," *Journal of Finance*, 52, 1–33.

Daniel, K., S. Titman, and K. J. Wei (2001): "Explaining the Cross-Section of Stock Returns in Japan: Factors or Characteristics?," *Journal of Finance*, 55, 743–766.

Fama, E. F., and K. R. French (1992): "The Cross-Section of Expected Stock Returns," *Journal of Finance*, 47, 427–465.

——— (1996): "Multifactor Explanations of Asset Pricing Anomalies," *Journal of Finance*, 51, 55–84.

Fama, E. F., and J. D. MacBeth (1973): "Risk, Return, and Equilibrium: Empirical Tests," *Journal of Political Economy*, 81, 607–636.

Gibbons, M. R., S. A. Ross, and J. Shanken (1989): "A Test of the Efficiency of a Given Portfolio," *Econometrica*, 57, 1121–1152.

Jagannathan, R., and Z. Wang (1996): "The Conditional CAPM and the Cross-Section of Expected Returns," *Journal of Finance*, 51, 3–53.

Kan, R., and C. Zhang (1999a): "GMM Tests of Stochastic Discount Factor Models with Useless Factors," *Journal of Financial Economics*, 54, 103–127.

Kan, R., and C. Zhang (1999b): "Two-Pass Tests of Asset Pricing Models with Useless Factors," *Journal of Finance*, 54, 203–235.

Kan, R., and G. Zhou (1999): "A Critique of the Stochastic Discount Factor Methodolgy," *Journal of Finance*, 54, 1221–1248.

Kandel, S., and R. F. Stambaugh (1995): "Portfolio Inefficiency and the Cross-Section of Expected Returns," *Journal of Finance*, 50, 157–184.

Kim, D. (1995): "The Errors in the Variables Problem in the Cross-Section of Expected Stock Returns," *Journal of Finance*, 50, 1605–1634.

Lintner, J. (1965): "The Valuation of Risky Assets and the Selection of Risky Investments in the Portfolios and Capital Budgets," *Review of Economics and Statistics*, 47, 13–37.

Merton, R. C. (1973): "An Intertemporal Capital Asset Pricing Model," *Econometrica*, 41, 867–887.

Roll, R., and S. A. Ross (1994): "On the Cross-Sectional Relation Between Expected Returns and Betas," *Journal of Finance*, 49, 101–121.

Ross, S. A. (1976): "The Arbitrage Theory of Capital Asset Pricing," *Journal of Economic Theory*, 13, 341–360.

Shanken, J. (1992): "On the Estimation of Beta-Pricing Models," *Review of Financial Studies*, 5, 1–33.

Sharpe, W. F. (1964): "Capital Asset Prices: A Theory of Market Equilibrium," *Journal of Finance*, 19, 425–442.

Appendix A Proofs

A.1 Proof of Lemma 1

Since F is a linear factor model for Z, we have:

$$\mu_Z = \beta^T \gamma = \Sigma_{ZF} \Sigma_{FF}^{-1} \gamma \tag{A.1}$$

From the regularity assumptions, the quantity $\Sigma_{FF} \left(\Sigma_{FZ} \Sigma_{ZZ}^{-1} \Sigma_{ZF} \right)^{-1} \Sigma_{FZ} \Sigma_{ZZ}^{-1}$ is well-defined and unique. Premultiplying both sides of (A.1) by this quantity yields:

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

The last expression is the desired result.

A.2 Proof of Theorem 1

We show that the other four quantities in the chain of equalities are all equal to $z^T \mu_Z$. Beginning with F, the risk premia vector is:

$$\begin{aligned} \gamma(F, Z) &= \Sigma_{FF} \left(\Sigma_{FZ} \Sigma_{ZZ}^{-1} \Sigma_{ZF} \right)^{-1} \Sigma_{FZ} \Sigma_{ZZ}^{-1} \mu_Z = \Sigma_{FF} \Sigma_{PP}^{-1} \Gamma_F^T \mu_Z \\ &= \left(\Sigma_{PP} + \Sigma_{\eta\eta} \right) \Sigma_{PP}^{-1} \Gamma_F^T \mu_Z = \Gamma_F^T \mu_Z + \Sigma_{\eta\eta} \Sigma_{PP}^{-1} \Gamma_F^T \mu_Z \end{aligned} \tag{A.2}$$

We premultiply both sides of the equation describing the projection of F onto Z by x^T:

$$x^T \left(F - \mu_F \right) = x^T \Gamma_F^T \left(Z - \mu_Z \right) + x^T \eta$$

Since η and Z are uncorrelated, and by assumption $x^T \left(F - \mu_F \right) = z^T \left(Z - \mu_Z \right)$, it must be the case that:

$$\begin{aligned} x^T \eta &= 0 \\ z^T &= x^T \Gamma_F^T \end{aligned}$$

From the last expression, we have $z^T \mu_Z = x^T \Gamma_F^T \mu_Z$. Premultiplying both sides of (A.2) by x^T, we find:

$$x^T \gamma(F, Z) = x^T \Gamma_F^T \mu_Z + x^T \Sigma_{\eta\eta} \Sigma_{PP}^{-1} \Gamma_F^T \mu_Z = z^T \mu_Z$$

Proceeding analogously with G in place of F establishes the remaining identities.

A.3 Proof of Theorem 3

Consider the deviation of the vector $\gamma(F,Z)$ from $\gamma(P,Z)$:

$$\gamma(F,Z) - \gamma(P,Z) = \Sigma_{\eta\eta}\Sigma_{PP}^{-1}\gamma(P,Z) \tag{A.3}$$

If we premultiply both sides of (A.3) by $\gamma(P,Z)^T\,\Sigma_{PP}^{-1}$, we find:

$$\gamma(P,Z)^T\,\Sigma_{PP}^{-1}\left[\gamma(F,Z) - \gamma(P,Z)\right] = \gamma(P,Z)^T\,\Sigma_{PP}^{-1}\Sigma_{\eta\eta}\Sigma_{PP}^{-1}\gamma(P,Z)$$

The right-hand side is clearly non-negative, since $\Sigma_{\eta\eta}$ is positive semidefinite. If the right-hand side is zero, then the right-hand side of (A.3) is also zero, and $\gamma(F,Z) = \gamma(P,Z)$. If the right-hand side is positive, we have:

$$\begin{aligned}\gamma(P,Z)^T\,\Sigma_{PP}^{-1}\left[\gamma(F,Z) - \gamma(P,Z)\right] &> 0\\ \gamma(P,Z)^T\,\Sigma_{PP}^{-1}\gamma(F,Z) &> \gamma(P,Z)^T\,\Sigma_{PP}^{-1}\gamma(P,Z)\end{aligned}$$

To show the other direction of the implication, we take as given a vector γ falling within the specified region (i.e., either $\gamma = \gamma(P,Z)$ or $\gamma(P,Z)^T\,\Sigma_{PP}^{-1}\gamma > \gamma(P,Z)^T\,\Sigma_{PP}^{-1}\gamma(P,Z)$), and construct a linear factor model F such that $\gamma(F,Z) = \gamma$. If $\gamma = \gamma(P,Z)$, then we can trivially take $F = P$. If $\gamma \neq \gamma(P,Z)$, then we must choose $\Sigma_{\eta\eta}$ such that:[3]

$$\gamma(F,Z) = \gamma(P,Z) + \Sigma_{\eta\eta}\Sigma_{PP}^{-1}\gamma(P,Z) = \gamma$$

If $N = 1$, we can trivially choose any random variable η such that:

$$Var[\eta] = \Sigma_{\eta\eta} = \frac{\Sigma_{PP}^{-1}\gamma(P,Z)}{\gamma - \gamma(P,Z)}$$

Since:

$$\left[\gamma - \gamma(P,Z)\right]^T\,\Sigma_{PP}^{-1}\gamma(P,Z) > 0$$

it must be the case that either:

$$\gamma - \gamma(P,Z) > 0 \text{ and } \Sigma_{PP}^{-1}\gamma(P,Z) > 0$$

or:

$$\gamma - \gamma(P,Z) < 0 \text{ and } \Sigma_{PP}^{-1}\gamma(P,Z) < 0$$

[3] We assume the probability space contains at least N random variables with finite variance that are uncorrelated with Z and each other.

In either case, the required value of $\Sigma_{\eta\eta}$ is positive.

For $N > 1$, rather than choosing $\Sigma_{\eta\eta}$ directly, we consider a non-singular transformation of this matrix. The required identity can be expressed as:

$$\gamma - \gamma(P,Z) = \Sigma_{\eta\eta}\Sigma_{PP}^{-1}\gamma(P,Z) = K\left[K^{-1}\Sigma_{\eta\eta}\left(K^{T}\right)^{-1}\right]K^{T}\Sigma_{PP}^{-1}\gamma(P,Z) \tag{A.4}$$

where K is an arbitrarily specified non-singular $N \times N$ matrix. We choose for K:

$$K = \begin{bmatrix} \frac{\Sigma_{PP}^{-1}\gamma(P,Z)}{\gamma(P,Z)^{T}\Sigma_{PP}^{-1}\Sigma_{PP}^{-1}\gamma(P,Z)} & \varepsilon & H \end{bmatrix}$$

where ε is defined as:

$$\varepsilon = [\gamma - \gamma(P,Z)] - \Sigma_{PP}^{-1}\gamma(P,Z)\frac{\gamma(P,Z)^{T}\Sigma_{PP}^{-1}[\gamma - \gamma(P,Z)]}{\gamma(P,Z)^{T}\Sigma_{PP}^{-1}\Sigma_{PP}^{-1}\gamma(P,Z)}$$

and H is an $N \times (N-2)$ matrix orthogonal to the first two columns of K:

$$\begin{aligned} \gamma(P,Z)^{T}\Sigma_{PP}^{-1}H &= 0 \\ \varepsilon^{T}H &= 0 \end{aligned}$$

(If $N = 2$, then H is empty and K consists only of the first two columns.) Note that ε is also orthogonal to the first column of K:

$$\frac{\gamma(P,Z)^{T}\Sigma_{PP}^{-1}\varepsilon}{\gamma(P,Z)^{T}\Sigma_{PP}^{-1}\Sigma_{PP}^{-1}\gamma(P,Z)} = \frac{\gamma(P,Z)^{T}\Sigma_{PP}^{-1}\left([\gamma - \gamma(P,Z)] - \Sigma_{PP}^{-1}\gamma(P,Z)\frac{\gamma(P,Z)^{T}\Sigma_{PP}^{-1}[\gamma-\gamma(P,Z)]}{\gamma(P,Z)^{T}\Sigma_{PP}^{-1}\Sigma_{PP}^{-1}\gamma(P,Z)}\right)}{\gamma(P,Z)^{T}\Sigma_{PP}^{-1}\Sigma_{PP}^{-1}\gamma(P,Z)} = 0$$

We define the matrix Ω:

$$\Omega = K^{-1}\Sigma_{\eta\eta}\left(K^{T}\right)^{-1}$$

We can now rewrite (A.4) as:

$$\gamma - \gamma(P,Z) = K\Omega K^{T}\Sigma_{PP}^{-1}\gamma(P,Z)$$

where Ω is a positive semidefinite matrix. Substituting in the definition of K, we find:

$$\gamma - \gamma(P,Z) = \begin{bmatrix} \frac{\Sigma_{PP}^{-1}\gamma(P,Z)}{\gamma(P,Z)^{T}\Sigma_{PP}^{-1}\Sigma_{PP}^{-1}\gamma(P,Z)} & \varepsilon & H \end{bmatrix}\Omega\begin{bmatrix} 1 \\ 0 \\ 0_{(N-2)\times 1} \end{bmatrix}$$

We choose Ω such that:

$$\Omega \begin{bmatrix} 1 \\ 0 \\ 0_{(N-2)\times 1} \end{bmatrix} = \begin{bmatrix} \gamma\left(P,Z\right)^T \Sigma_{PP}^{-1}\left[\gamma - \gamma\left(P,Z\right)\right] \\ 1 \\ 0_{(N-2)\times 1} \end{bmatrix}$$

We then have:

$$\begin{aligned} \gamma - \gamma\left(P,Z\right) &= \begin{bmatrix} \frac{\Sigma_{PP}^{-1}\gamma(P,Z)}{\gamma(P,Z)^T\Sigma_{PP}^{-1}\Sigma_{PP}^{-1}\gamma(P,Z)} & \varepsilon & H \end{bmatrix} \begin{bmatrix} \gamma\left(P,Z\right)^T \Sigma_{PP}^{-1}\left[\gamma - \gamma\left(P,Z\right)\right] \\ 1 \\ 0_{(N-2)\times 1} \end{bmatrix} \\ &= \frac{\Sigma_{PP}^{-1}\gamma\left(P,Z\right)}{\gamma\left(P,Z\right)^T \Sigma_{PP}^{-1}\Sigma_{PP}^{-1}\gamma\left(P,Z\right)}\gamma\left(P,Z\right)^T \Sigma_{PP}^{-1}\left[\gamma - \gamma\left(P,Z\right)\right] + \varepsilon \\ &= \gamma - \gamma\left(P,Z\right) \end{aligned} \tag{A.5}$$

By assumption, $\gamma\left(P,Z\right)^T \Sigma_{PP}^{-1}\left[\gamma - \gamma\left(P,Z\right)\right]$ is positive, so the matrix Ω can be chosen to be positive semidefinite. Since $\Sigma_{\eta\eta} = K\Omega K^T$ and K is non-singular, $\Sigma_{\eta\eta}$ is also positive semidefinite; from (A.5), this choice of $\Sigma_{\eta\eta}$ generates the specified vector of risk premia γ.

A.4 Proof of Theorem 4

Let G be a set of additional factors, such that the union of F and G, called H, is a linear factor model for Z. Consider the partitioned model:

$$H = \begin{bmatrix} F \\ G \end{bmatrix} = \begin{bmatrix} \Gamma_F^T \\ \Gamma_G^T \end{bmatrix}\left(Z - \mu_Z\right) + \begin{bmatrix} \eta \\ \delta \end{bmatrix} = \begin{bmatrix} P \\ Q \end{bmatrix} + \begin{bmatrix} \eta \\ \delta \end{bmatrix}$$

where η and δ are uncorrelated with Z. The vector of risk premia for the projection of H onto the asset space is given by:

$$\gamma\left(\begin{bmatrix} P \\ Q \end{bmatrix}, Z\right) = \begin{bmatrix} \Gamma_F^T \\ \Gamma_G^T \end{bmatrix}\mu_Z$$

The vector of risk premia for H is given by:

$$\begin{aligned} \gamma\left(H,Z\right) &= \gamma\left(\begin{bmatrix} P \\ Q \end{bmatrix}, Z\right) + \begin{bmatrix} \Sigma_{\eta\eta} & \Sigma_{\eta\delta} \\ \Sigma_{\delta\eta} & \Sigma_{\delta\delta} \end{bmatrix}\begin{bmatrix} \Gamma_F^T\Sigma_{ZZ}\Gamma_F & \Gamma_F^T\Sigma_{ZZ}\Gamma_G \\ \Gamma_G^T\Sigma_{ZZ}\Gamma_F & \Gamma_G^T\Sigma_{ZZ}\Gamma_G \end{bmatrix}^{-1}\begin{bmatrix} \Gamma_F^T \\ \Gamma_G^T \end{bmatrix}\mu_Z \\ &= \gamma\left(\begin{bmatrix} P \\ Q \end{bmatrix}, Z\right) + \begin{bmatrix} \Sigma_{\eta\eta} & \Sigma_{\eta\delta} \\ \Sigma_{\delta\eta} & \Sigma_{\delta\delta} \end{bmatrix}\Phi\mu_Z \end{aligned} \tag{A.6}$$

where:

$$\Phi = \begin{bmatrix} \left[\Gamma_F^T \Sigma_{ZZ} \Gamma_F - \Gamma_F^T \Sigma_{ZZ} \Gamma_G \left(\Gamma_G^T \Sigma_{ZZ} \Gamma_G\right)^{-1} \Gamma_G^T \Sigma_{ZZ} \Gamma_F\right]^{-1} \left[\Gamma_F^T - \Gamma_F^T \Sigma_{ZZ} \Gamma_G \left(\Gamma_G^T \Sigma_{ZZ} \Gamma_G\right)^{-1} \Gamma_G^T\right] \\ \left[\Gamma_G^T \Sigma_{ZZ} \Gamma_G - \Gamma_G^T \Sigma_{ZZ} \Gamma_F \left(\Gamma_F^T \Sigma_{ZZ} \Gamma_F\right)^{-1} \Gamma_F^T \Sigma_{ZZ} \Gamma_G\right]^{-1} \left[\Gamma_G^T - \Gamma_G^T \Sigma_{ZZ} \Gamma_F \left(\Gamma_F^T \Sigma_{ZZ} \Gamma_F\right)^{-1} \Gamma_F^T\right] \end{bmatrix}$$

Since H is a linear factor model, it must be the case that:

$$\mu_Z = \beta_H^T \gamma(H,Z) = \Sigma_{ZH} \Sigma_{HH}^{-1} \gamma(H,Z) = \begin{bmatrix} \Sigma_{ZF} \\ \Sigma_{ZG} \end{bmatrix} \Sigma_{HH}^{-1} \gamma(H,Z)$$
$$= \Sigma_{ZF} \lambda_F + \Sigma_{ZG} \lambda_G = \Sigma_{ZZ} \Gamma_F \lambda_F + \Sigma_{ZZ} \Gamma_G \lambda_G$$

for some λ_F and λ_G. Plugging this value for μ_Z into (A.6), we find:

$$\gamma(H,Z) - \gamma\left(\begin{bmatrix} P \\ Q \end{bmatrix}, Z\right) = \begin{bmatrix} \Sigma_{\eta\eta} \lambda_F + \Sigma_{\eta\delta} \lambda_G \\ \Sigma_{\delta\eta} \lambda_F + \Sigma_{\delta\delta} \lambda_G \end{bmatrix}$$

Consider the case in which the factors G have no unspanned components (i. e., $\delta = 0$). In this case, we have:

$$\gamma(H,Z) - \gamma\left(\begin{bmatrix} P \\ Q \end{bmatrix}, Z\right) = \begin{bmatrix} \Sigma_{\eta\eta} \lambda_F \\ 0 \end{bmatrix}$$

The matrix Ψ in this case is given by:

$$\Psi = \begin{bmatrix} I_N \\ 0_{K \times N} \end{bmatrix}$$

where K is the number of additional factors contained in G. Then the desired relation is:

$$\gamma = \Psi^T \gamma(H,Z) = \Psi^T \gamma\left(\begin{bmatrix} P \\ Q \end{bmatrix}, Z\right) + \Psi^T \begin{bmatrix} \Sigma_{\eta\eta} \lambda_F \\ 0 \end{bmatrix} = \gamma(P,Z) + \Sigma_{\eta\eta} \lambda_F$$
$$\gamma - \gamma(P,Z) = \Sigma_{\eta\eta} \lambda_F$$

Let L be the rank of $\Sigma_{\eta\eta}$. Let $w_1, ..., w_{N-L}$ be $N-L$ linearly independent vectors such that $w_i^T \Sigma_{\eta\eta} w_i = 0$ for each $1 \le i \le N-L$, and let $v_1, ..., v_L$ be a set of L vectors linearly independent of each other and of $w_1, ..., w_{L-N}$. By assumption, $w_i^T (\gamma - \gamma(P,Z)) = 0$ for each $1 \le i \le N-L$. $\gamma - \gamma(P,Z)$ can therefore be expressed as a linear combination of the vectors $v_1, ..., v_L$; furthermore, each v_i, $1 \le i \le L$ can be expressed as $v_i = \Sigma_{\eta\eta} x$ for some vector x. There therefore exists a vector λ such that:

$$\gamma - \gamma(P,Z) = \Sigma_{\eta\eta} \lambda \tag{A.7}$$

We therefore need only choose a set of factors G, such that each factor in G is spanned, the union of F and G is a linear factor model for Z, and the value of λ_F corresponding to this choice of G is equal to λ. We choose:

$$G = \left(\mu_Z^T \Sigma_{ZZ}^{-1} - \lambda^T \Gamma_F^T\right)(Z - \mu_Z) \tag{A.8}$$

G is clearly spanned by the assets Z. Σ_{HH} can only be singular if the factor G can be expressed as a linear combination of P. Suppose this is the case for some N-vector x:

$$\begin{aligned}
\left(\mu_Z^T \Sigma_{ZZ}^{-1} - \lambda^T \Gamma_F^T\right)(Z - \mu_Z) = x^T P = x^T \Gamma^T (Z - \mu_Z) \\
\left(\mu_Z^T \Sigma_{ZZ}^{-1} - \lambda^T \Gamma_F^T - x^T \Gamma^T\right) = 0 \\
\mu_Z &= \Sigma_{ZZ} \Gamma_F (\lambda + x) = \Sigma_{ZF} (\lambda + x) \\
&= \Sigma_{ZF} \Sigma_{FF}^{-1} \Sigma_{FF} (\lambda + x) = \beta_F^T \Sigma_{FF} (\lambda + x)
\end{aligned}$$

The last result shows that, contrary to the assumptions of the theorem, F is a linear factor model for Z. Σ_{HH} is therefore non-singular. Similarly, Σ_{HZ} is given by:

$$\Sigma_{HZ} = \begin{bmatrix} \Sigma_{FZ} \\ \Sigma_{GZ} \end{bmatrix} = \begin{bmatrix} \Sigma_{FZ} \\ \left(\mu_Z^T \Sigma_{ZZ}^{-1} - \lambda^T \Gamma_F^T\right) \Sigma_{ZZ} \end{bmatrix} = \begin{bmatrix} \Sigma_{FZ} \\ \left(\mu_Z^T - \lambda^T \Sigma_{FZ}\right) \end{bmatrix}$$

For this matrix to have rank N (rather than $N+1$), there must exist an N-vector x and constant y such that either $x \neq 0$ or $y \neq 0$ and:

$$\begin{bmatrix} x^T & y \end{bmatrix} \Sigma_{HZ} = \begin{bmatrix} x^T & y \end{bmatrix} \begin{bmatrix} \Sigma_{FZ} \\ \left(\mu_Z^T - \lambda^T \Sigma_{FZ}\right) \end{bmatrix} = x^T \Sigma_{FZ} + y \left(\mu_Z^T - \lambda^T \Sigma_{FZ}\right) = 0$$

If $y = 0$, and there exists such an $x \neq 0$, then F and Z, contrary to the assumptions of the theorem, do not satisfy the regularity assumptions. But if $y \neq 0$, then the last equation becomes:

$$\begin{aligned}
x^T \Sigma_{FZ} + y \left(\mu_Z^T - \lambda^T \Sigma_{FZ}\right) = 0 \\
\mu_Z y &= \Sigma_{ZF} \lambda y - \Sigma_{ZF} x \\
\mu_Z &= \Sigma_{ZF} \left(\lambda - \frac{x}{y}\right) = \beta_F^T \Sigma_{FF} \left(\lambda - \frac{x}{y}\right)
\end{aligned}$$

F is then, contrary to the theorem assumptions, a linear factor model for Z. It must therefore be the case that Σ_{HZ} has rank $N+1$. H and Z therefore satisfy the regularity assumptions. Finally, we note that:

$$\begin{aligned}
\Sigma_{ZH} &= \begin{bmatrix} \Sigma_{ZF} & (\mu_Z - \Sigma_{ZF} \lambda) \end{bmatrix} \\
\Sigma_{ZH} \begin{bmatrix} \lambda \\ 1 \end{bmatrix} &= \begin{bmatrix} \Sigma_{ZF} & (\mu_Z - \Sigma_{ZF} \lambda) \end{bmatrix} \begin{bmatrix} \lambda \\ 1 \end{bmatrix} = \Sigma_{ZF} \lambda + \mu_Z - \Sigma_{ZF} \lambda = \mu_Z
\end{aligned}$$

$$\mu_Z = \Sigma_{ZH} \begin{bmatrix} \lambda \\ 1 \end{bmatrix} = \Sigma_{ZH}\Sigma_{HH}^{-1}\Sigma_{HH} \begin{bmatrix} \lambda \\ 1 \end{bmatrix} = \beta_H^T \Sigma_{HH} \begin{bmatrix} \lambda \\ 1 \end{bmatrix}$$

So H is a linear factor model for Z. Construction of H therefore requires selection of a vector λ that satisfies (A.7), and then selecting G according to (A.8).

A.5 Proof of Theorem 2

Since F is a linear factor model for Z, we have:

$$\mu_Z = \beta_F^T \gamma(F, Z) = \Sigma_{ZF}\Sigma_{FF}^{-1}\gamma(F, Z)$$

This expression can be manipulated to yield:

$$\mu_Z = \Sigma_{ZF}\Sigma_{FF}^{-1}\gamma(F, Z) = \Sigma_{ZG}\Psi\Sigma_{FF}^{-1}\gamma(F, Z) = \Sigma_{ZG}\Sigma_{GG}^{-1}\Sigma_{GG}\Psi\Sigma_{FF}^{-1}\gamma(F, Z) = \beta_G^T\Sigma_{GG}\Psi\Sigma_{FF}^{-1}\gamma(F, Z)$$

From the last expression, G is a linear factor model for Z with:

$$\gamma(G, Z) = \Sigma_{GG}\Psi\Sigma_{FF}^{-1}\gamma(F, Z) = \Sigma_{GG}\Psi\left(\Psi^T\Sigma_{GG}\Psi\right)^{-1}\gamma(F, Z)$$

Premultiplication of both sides by Ψ^T yields the desired result:

$$\Psi^T\gamma(G, Z) = \Psi^T\Sigma_{GG}\Psi\left(\Psi^T\Sigma_{GG}\Psi\right)^{-1}\gamma(F, Z) = \gamma(F, Z)$$

A.6 Proof of Theorem 5

The task at hand is to choose Ψ so that:

$$\mu_0 = \left(\Psi\Psi^T\right)^{-1}\Psi\Sigma_{\eta\eta}\Sigma_{PP}^{-1}\Gamma^T\mu_Z \tag{A.9}$$

But recall:

$$\Sigma_{\eta\eta} = \Psi^T\Sigma_{YY}\Psi$$

(A.9) therefore simplifies to:

$$\begin{aligned} \mu_0 &= \left(\Psi\Psi^T\right)^{-1}\Psi\left(\Psi^T\Sigma_{YY}\Psi\right)\Sigma_{PP}^{-1}\Gamma^T\mu_Z = \Sigma_{YY}\Psi\Sigma_{PP}^{-1}\Gamma^T\mu_Z \\ \Sigma_{YY}^{-1}\mu_0 &= \Psi\Sigma_{PP}^{-1}\Gamma^T\mu_Z \end{aligned}$$

The left-hand side of the last equation is a K-vector, and the right-hand side is a $K \times N$ matrix multiplied by an N-vector. Since $K \leq N$, we can choose a full-rank matrix Ψ that satisfies the equation.

A.7 Proof of Theorem 6

Under the assumptions of the theorem, the sample mean $\widehat{\mu}_Z$ of the excess return vector Z is a consistent estimator with asymptotic variance Σ_{ZZ}. The sample estimate is:

$$\widehat{\gamma}(P,Z) = \widehat{\Gamma}^T \widehat{\mu}_Z$$

Note, however, that $\widehat{\Gamma} = \Gamma$; the factors P are spanned, and their projection onto the assets can therefore be estimated without error. The asymptotic variance of the estimator is therefore:

$$AVar\left[\widehat{\gamma}(P,Z)\right] = AVar\left[\Gamma^T \widehat{\mu}_Z\right] = \Gamma^T AVar\left[\widehat{\mu}_Z\right]\Gamma = \Gamma^T \Sigma_{ZZ}\Gamma = \Sigma_{PP}$$

A.8 Proof of Theorem 7

The estimate is given by:

$$\widehat{\gamma}(P,Z) = \widehat{\Gamma}^T \widehat{\mu}_Z = \widehat{\Sigma}_{FZ}\widehat{\Sigma}_{ZZ}^{-1}\widehat{\mu}_Z$$

A first-order Taylor expansion is given by:

$$\widehat{\gamma}(P,Z) - \gamma(P,Z) \approx \left(\widehat{\Sigma}_{FZ} - \Sigma_{FZ}\right)\Sigma_{ZZ}^{-1}\mu_Z - \Gamma^T\left(\widehat{\Sigma}_{ZZ} - \Sigma_{ZZ}\right)\Sigma_{ZZ}^{-1}\mu_Z + \Gamma^T\left(\widehat{\mu}_Z - \mu_Z\right)$$

We denote the three terms on the right-hand side as $\widehat{\gamma}_1$, $\widehat{\gamma}_2$ and $\widehat{\gamma}_3$. The asymptotic covariance between each pair of terms can be found element by element as follows:

$$\begin{aligned}
ACovar\left([\widehat{\gamma}_1]_{i1}, [\widehat{\gamma}_1]_{j1}\right) &= ACovar\left(\sum_{k=1}^{M}\left[\widehat{\Sigma}_{FZ} - \Sigma_{FZ}\right]_{ik}\left[\Sigma_{ZZ}^{-1}\mu_Z\right]_{k1}, \sum_{l=1}^{M}\left[\widehat{\Sigma}_{FZ} - \Sigma_{FZ}\right]_{jl}\left[\Sigma_{ZZ}^{-1}\mu_Z\right]_{l1}\right) \\
&= \sum_{k=1}^{M}\sum_{l=1}^{M}\left[\Sigma_{ZZ}^{-1}\mu_Z\right]_{k1}\left[\Sigma_{ZZ}^{-1}\mu_Z\right]_{l1} ACovar\left(\left[\widehat{\Sigma}_{FZ} - \Sigma_{FZ}\right]_{ik}, \left[\widehat{\Sigma}_{FZ} - \Sigma_{FZ}\right]_{jl}\right) \\
&= \sum_{k=1}^{M}\sum_{l=1}^{M}\left[\Sigma_{ZZ}^{-1}\mu_Z\right]_{k1}\left[\Sigma_{ZZ}^{-1}\mu_Z\right]_{l1}\left([\Sigma_{FF}]_{ij}[\Sigma_{ZZ}]_{kl} + [\Sigma_{FZ}]_{il}[\Sigma_{ZF}]_{kj}\right) \\
&= [\Sigma_{FF}]_{ij}\left[\mu_Z^T\Sigma_{ZZ}^{-1}\mu_Z\right] + \left[\Gamma^T\mu_Z\right]_{i1}\left[\Gamma^T\mu_Z\right]_{j1}
\end{aligned}$$

Proceeding similarly with the other pairs, we find:

$$\begin{aligned}
ACovar\left([\widehat{\gamma}_1]_{i1}, [\widehat{\gamma}_2]_{j1}\right) &= [\Sigma_{PP}]_{ij}\left[\mu_Z^T\Sigma_{ZZ}^{-1}\mu_Z\right] + \left[\Gamma^T\mu_Z\right]_{i1}\left[\Gamma^T\mu_Z\right]_{j1} \\
ACovar\left([\widehat{\gamma}_1]_{i1}, [\widehat{\gamma}_3]_{j1}\right) &= 0 \\
ACovar\left([\widehat{\gamma}_2]_{i1}, [\widehat{\gamma}_2]_{j1}\right) &= [\Sigma_{PP}]_{ij}\left[\mu_Z^T\Sigma_{ZZ}^{-1}\mu_Z\right] + \left[\Gamma^T\mu_Z\right]_{i1}\left[\Gamma^T\mu_Z\right]_{j1} \\
ACovar\left([\widehat{\gamma}_2]_{i1}, [\widehat{\gamma}_3]_{j1}\right) &= 0
\end{aligned}$$

$$ACovar\left([\widehat{\gamma}_3]_{i1}, [\widehat{\gamma}_3]_{j1}\right) = [\Sigma_{PP}]_{ij}$$

Putting together the above results, we find:

$$ACovar\left([\widehat{\gamma}(P,Z)]_i, [\widehat{\gamma}(P,Z)]_j\right) = [\Sigma_{PP}]_{ij} + [\Sigma_{\eta\eta}]_{ij}\left[\mu_Z^T \Sigma_{ZZ}^{-1}\mu_Z\right]$$

The desired result follows immediately.

A.9 Proof of Theorem 8

The estimator is:

$$\widehat{\gamma}(F,Z) = \widehat{\Sigma}_{FF}\widehat{\Sigma}_{PP}^{-1}\widehat{\Gamma}^T\widehat{\mu}_Z = \widehat{\Sigma}_{FF}\left(\widehat{\Sigma}_{FZ}\widehat{\Sigma}_{ZZ}^{-1}\widehat{\Sigma}_{ZF}\right)^{-1}\widehat{\Sigma}_{FZ}\widehat{\Sigma}_{ZZ}^{-1}\widehat{\mu}_Z$$

A first order Taylor expansion is:

$$\begin{aligned}
\widehat{\gamma}(F,Z) - \gamma(F,Z) \approx &\left(\widehat{\Sigma}_{FF} - \Sigma_{FF}\right)\Sigma_{PP}^{-1}\Gamma^T\mu_Z \\
&- \Sigma_{FF}\Sigma_{PP}^{-1}\left(\widehat{\Sigma}_{FZ} - \Sigma_{FZ}\right)\Gamma\Sigma_{PP}^{-1}\Gamma^T\mu_Z \\
&+ \Sigma_{FF}\Sigma_{PP}^{-1}\Gamma^T\left(\widehat{\Sigma}_{ZZ} - \Sigma_{ZZ}\right)\Gamma\Sigma_{PP}^{-1}\Gamma^T\mu_Z \\
&- \Sigma_{FF}\Sigma_{PP}^{-1}\Gamma^T\left(\widehat{\Sigma}_{ZF} - \Sigma_{ZF}\right)\Sigma_{PP}^{-1}\Gamma^T\mu_Z \\
&+ \Sigma_{FF}\Sigma_{PP}^{-1}\left(\widehat{\Sigma}_{FZ} - \Sigma_{FZ}\right)\Sigma_{ZZ}^{-1}\mu_Z \\
&- \Sigma_{FF}\Sigma_{PP}^{-1}\Gamma^T\left(\widehat{\Sigma}_{ZZ} - \Sigma_{ZZ}\right)\Sigma_{ZZ}^{-1}\mu_Z \\
&+ \Sigma_{FF}\Sigma_{PP}^{-1}\Gamma^T\left(\widehat{\mu}_Z - \mu_Z\right)
\end{aligned}$$

However, since F is a linear factor model for Z, we have $\mu_Z = \beta^T\gamma(F,Z)$. Making this substitution, four of the seven terms on the right-hand side cancel:

$$\begin{aligned}
\widehat{\gamma}(F,Z) - \gamma(F,Z) \approx &\left(\widehat{\Sigma}_{FF} - \Sigma_{FF}\right)\Sigma_{FF}^{-1}\gamma(F,Z) \\
&- \Sigma_{FF}\Sigma_{PP}^{-1}\Gamma^T\left(\widehat{\Sigma}_{ZF} - \Sigma_{ZF}\right)\Sigma_{FF}^{-1}\gamma(F,Z) \\
&+ \Sigma_{FF}\Sigma_{PP}^{-1}\Gamma^T\left(\widehat{\mu}_Z - \mu_Z\right)
\end{aligned}$$

Denoting the terms on the right-hand side by $\widehat{\gamma}_1$ through $\widehat{\gamma}_3$, and applying the technique used for the previous theorem, the asymptotic covariances between each pair of terms is given by:

$$\begin{aligned}
ACovar\left([\widehat{\gamma}_1]_i, [\widehat{\gamma}_1]_j\right) &= [\Sigma_{FF}]_{ij}\left[\gamma^T(F,Z)\Sigma_{FF}^{-1}\gamma(F,Z)\right] + [\gamma(F,Z)]_{i1}[\gamma(F,Z)]_{j1} \\
ACovar\left([\widehat{\gamma}_1]_i, [\widehat{\gamma}_2]_j\right) &= [\Sigma_{FF}]_{ij}\left[\gamma^T(F,Z)\Sigma_{FF}^{-1}\gamma(F,Z)\right] + [\gamma(F,Z)]_{i1}[\gamma(F,Z)]_{j1}
\end{aligned}$$

$$ACovar\left([\widehat{\gamma}_1]_i, [\widehat{\gamma}_3]_j\right) = 0$$

$$ACovar\left([\widehat{\gamma}_2]_i, [\widehat{\gamma}_2]_j\right) = \left[\Sigma_{FF}\Sigma_{PP}^{-1}\Sigma_{FF}\right]_{ij}\left[\gamma^T(F,Z)\,\Sigma_{FF}^{-1}\gamma(F,Z)\right] + [\gamma(F,Z)]_{i1}\,[\gamma(F,Z)]_{j1}$$

$$ACovar\left([\widehat{\gamma}_2]_i, [\widehat{\gamma}_3]_j\right) = 0$$

$$ACovar\left([\widehat{\gamma}_3]_i, [\widehat{\gamma}_3]_j\right) = \left[\Sigma_{FF}\Sigma_{PP}^{-1}\Sigma_{FF}\right]_{ij}$$

Adding up all the terms, we find:

$$\begin{aligned} ACovar\left([\widehat{\gamma}(F,Z)]_i, [\widehat{\gamma}(F,Z)]_j\right) &= [\Sigma_{FF}]_{ij} + \\ &\left[\Sigma_{FF}\Sigma_{PP}^{-1}\Sigma_{FF} - \Sigma_{FF}\right]_{ij}\left[1 + \gamma^T(F,Z)\,\Sigma_{FF}^{-1}\gamma(F,Z)\right] \\ &= [\Sigma_{FF}]_{ij} + \\ &\left[\Sigma_{\eta\eta} + \Sigma_{\eta\eta}\Sigma_{PP}^{-1}\Sigma_{\eta\eta}\right]_{ij}\left[1 + \gamma^T(F,Z)\,\Sigma_{FF}^{-1}\gamma(F,Z)\right] \end{aligned}$$

This last expression yields the first desired result. To express this asymptotic covariance in terms of μ_Z instead of $\gamma(F,Z)$, we substitute in $\gamma(F,Z) = \Sigma_{FF}\Sigma_{PP}^{-1}\Gamma^T\mu_Z$:

$$\begin{aligned} ACovar\left([\widehat{\gamma}(F,Z)]_i, [\widehat{\gamma}(F,Z)]_j\right) &= [\Sigma_{FF}]_{ij} + \left[\Sigma_{\eta\eta} + \Sigma_{\eta\eta}\Sigma_{PP}^{-1}\Sigma_{\eta\eta}\right]_{ij}\left[1 + \mu_Z^T\Gamma\Sigma_{PP}^{-1}\Sigma_{FF}\Sigma_{PP}^{-1}\Gamma^T\mu_Z\right] \\ &= [\Sigma_{FF}]_{ij} + \left[\Sigma_{\eta\eta} + \Sigma_{\eta\eta}\Sigma_{PP}^{-1}\Sigma_{\eta\eta}\right]_{ij}\left[1 + \mu_Z^T\Gamma\Sigma_{PP}^{-1}\Gamma^T\mu_Z + \mu_Z^T\Gamma\Sigma_{PP}^{-1}\Sigma_{\eta\eta}\Sigma_{PP}^{-1}\Gamma^T\mu_Z\right] \end{aligned}$$

A.10 Proof of Theorem 9

The first order Taylor expansion can be expressed as:

$$\begin{aligned} \widehat{\gamma}(F,Z) - \gamma(F,Z) \approx &\left(\widehat{\Sigma}_{FF} - \Sigma_{FF}\right)\Sigma_{PP}^{-1}\Gamma^T\mu_Z \\ &+ \Sigma_{FF}\Sigma_{PP}^{-1}\Gamma^T\left(\widehat{\Sigma}_{ZZ} - \Sigma_{ZZ}\right)\left(\Gamma\Sigma_{PP}^{-1}\Gamma^T - \Sigma_{ZZ}^{-1}\right)\mu_Z \\ &- \Sigma_{FF}\Sigma_{PP}^{-1}\left(\widehat{\Sigma}_{FZ} - \Sigma_{FZ}\right)\left(\Gamma\Sigma_{PP}^{-1}\Gamma^T - \Sigma_{ZZ}^{-1}\right)\mu_Z \\ &- \Sigma_{FF}\Sigma_{PP}^{-1}\Gamma^T\left(\widehat{\Sigma}_{ZF} - \Sigma_{ZF}\right)\Sigma_{PP}^{-1}\Gamma^T\mu_Z \\ &+ \Sigma_{FF}\Sigma_{PP}^{-1}\Gamma^T\left(\widehat{\mu}_Z - \mu_Z\right) \end{aligned}$$

Denoting the terms on the right-hand side by $\widehat{\gamma}_1$ through $\widehat{\gamma}_5$, and applying the technique used for the previous theorems, the asymptotic covariances between each pair of terms is given by:

$$\begin{aligned} ACovar\left([\widehat{\gamma}_1]_i, [\widehat{\gamma}_1]_j\right) &= [\Sigma_{FF}]_{ij}\left[\mu_Z^T\Gamma\Sigma_{PP}^{-1}\Sigma_{FF}\Sigma_{PP}^{-1}\Gamma^T\mu_Z\right] \\ &+ \left[\Sigma_{FF}\Sigma_{PP}^{-1}\Gamma^T\mu_Z\right]_{i1}\left[\Sigma_{FF}\Sigma_{PP}^{-1}\Gamma^T\mu_Z\right]_{j1} \\ ACovar\left([\widehat{\gamma}_1]_i, [\widehat{\gamma}_2]_j\right) &= 0 \end{aligned}$$

$$
\begin{aligned}
ACovar\left([\widehat{\gamma}_1]_i, [\widehat{\gamma}_3]_j\right) &= 0 \\
ACovar\left([\widehat{\gamma}_1]_i, [\widehat{\gamma}_4]_j\right) &= [\Sigma_{FF}]_{ij} \left[\mu_Z^T \Gamma \Sigma_{PP}^{-1} \Sigma_{FF} \Sigma_{PP}^{-1} \Gamma^T \mu_Z\right] \\
&+ \left[\Sigma_{FF} \Sigma_{PP}^{-1} \Gamma^T \mu_Z\right]_{i1} \left[\Sigma_{FF} \Sigma_{PP}^{-1} \Gamma^T \mu_Z\right]_{j1} \\
ACovar\left([\widehat{\gamma}_1]_i, [\widehat{\gamma}_5]_j\right) &= 0 \\
ACovar\left([\widehat{\gamma}_2]_i, [\widehat{\gamma}_2]_j\right) &= \left[\Sigma_{FF} \Sigma_{PP}^{-1} \Sigma_{FF}\right]_{ij} \left[\mu_Z^T \Sigma_{ZZ}^{-1} \mu_Z - \mu_Z^T \Gamma \Sigma_{PP}^{-1} \Gamma^T \mu_Z\right] \\
ACovar\left([\widehat{\gamma}_2]_i, [\widehat{\gamma}_3]_j\right) &= \left[\Sigma_{FF} \Sigma_{PP}^{-1} \Sigma_{FF}\right]_{ij} \left[\mu_Z^T \Sigma_{ZZ}^{-1} \mu_Z - \mu_Z^T \Gamma \Sigma_{PP}^{-1} \Gamma^T \mu_Z\right] \\
ACovar\left([\widehat{\gamma}_2]_i, [\widehat{\gamma}_4]_j\right) &= 0 \\
ACovar\left([\widehat{\gamma}_2]_i, [\widehat{\gamma}_5]_j\right) &= 0 \\
ACovar\left([\widehat{\gamma}_3]_i, [\widehat{\gamma}_3]_j\right) &= \left[\Sigma_{FF} \Sigma_{PP}^{-1} \Sigma_{FF} \Sigma_{PP}^{-1} \Sigma_{FF}\right]_{ij} \left[\mu_Z^T \left(\Sigma_{ZZ}^{-1} - \Gamma \Sigma_{PP}^{-1} \Gamma^T\right) \mu_Z\right] \\
ACovar\left([\widehat{\gamma}_3]_i, [\widehat{\gamma}_4]_j\right) &= 0 \\
ACovar\left([\widehat{\gamma}_3]_i, [\widehat{\gamma}_5]_j\right) &= 0 \\
ACovar\left([\widehat{\gamma}_4]_i, [\widehat{\gamma}_4]_j\right) &= \left[\Sigma_{FF} \Sigma_{PP}^{-1} \Sigma_{FF}\right]_{ij} \left(\mu_Z^T \Gamma \Sigma_{PP}^{-1} \Sigma_{FF} \Sigma_{PP}^{-1} \Gamma^T \mu_Z\right) \\
&+ \left[\Sigma_{FF} \Sigma_{PP}^{-1} \Gamma^T \mu_Z\right]_{i1} \left[\Sigma_{FF} \Sigma_{PP}^{-1} \Gamma^T \mu_Z\right]_{j1} \\
ACovar\left([\widehat{\gamma}_4]_i, [\widehat{\gamma}_5]_j\right) &= 0 \\
ACovar\left([\widehat{\gamma}_5]_i, [\widehat{\gamma}_5]_j\right) &= \left[\Sigma_{FF} \Sigma_{PP}^{-1} \Sigma_{FF}\right]_{ij}
\end{aligned}
$$

Summing up the terms, we find:

$$
\begin{aligned}
ACovar\left([\widehat{\gamma}(F,Z)]_i, [\widehat{\gamma}(F,Z)]_j\right) &= -[\Sigma_{FF}]_{ij} \left[\mu_Z^T \Gamma \Sigma_{PP}^{-1} \Sigma_{FF} \Sigma_{PP}^{-1} \Gamma^T \mu_Z\right] \\
&+ \left[\Sigma_{FF} \Sigma_{PP}^{-1} \Sigma_{FF}\right]_{ij} \left[1 - \mu_Z^T \Sigma_{ZZ}^{-1} \mu_Z + \mu_Z^T \Gamma \Sigma_{PP}^{-1} \Gamma^T \mu_Z\right] \\
&+ \left[\Sigma_{FF} \Sigma_{PP}^{-1} \Sigma_{FF} \Sigma_{PP}^{-1} \Sigma_{FF}\right]_{ij} \left[\mu_Z^T \Sigma_{ZZ}^{-1} \mu_Z - \mu_Z^T \Gamma \Sigma_{PP}^{-1} \Gamma^T \mu_Z\right] \\
&= [\Sigma_{FF}]_{ij} + \left[\Sigma_{\eta\eta} + \Sigma_{\eta\eta} \Sigma_{PP}^{-1} \Sigma_{\eta\eta}\right]_{ij} \left[1 + \mu_Z^T \Sigma_{ZZ}^{-1} \mu_Z + \mu_Z^T \Gamma \Sigma_{PP}^{-1} \Sigma_{\eta\eta} \Sigma_{PP}^{-1} \Gamma^T \mu_Z\right] \\
&+ \left[\Sigma_{\eta\eta} \Sigma_{PP}^{-1} \Sigma_{\eta\eta} + \Sigma_{\eta\eta} \Sigma_{PP}^{-1} \Sigma_{\eta\eta} \Sigma_{PP}^{-1} \Sigma_{\eta\eta}\right]_{ij} \left[\mu_Z^T \Sigma_{ZZ}^{-1} \mu_Z - \mu_Z^T \Gamma \Sigma_{PP}^{-1} \Gamma^T \mu_Z\right]
\end{aligned}
$$

The last expression yields the desired result.