A Skeptical Appraisal of Robust Asset Pricing Tests

Tim A. Kroencke[‡] Julian Thimme[§]

This version: June 4, 2021

Abstract

We analyze the size and power of a large number of "robust" asset pricing tests, investigating the hypothesis that the price of risk of a candidate factor is equal to zero. Different from earlier studies, our bootstrap approach puts all tests on an equal footing and focuses on sample sizes comparable to standard applications in asset pricing research. Thus, our paper provides guidance for researchers about which method to use. We find that the classic Fama-MacBeth/Shanken approach does not over-reject useless factors and provides a reasonable balance between size and power. In contrast, some of the "robust" methods suffer from poor power in realistic sample sizes, especially in situations where the asset pricing model is mildly misspecified.

Keywords: Asset pricing tests · factor models · small samples

JEL: C12 · C31 · C58 · G12

[‡]University of Neuchâtel, Switzerland, Email: tim.kroencke@unine.ch.
[§]Karlsruhe Institute of Technology, Germany, Email: thimme@kit.edu.

1 Introduction

The trade-off between risk and return is one of the most intuitive principles in finance. Asset pricing theory implies that expected excess returns should be proportional to the asset's risk, as quantified by its covariance with a risk factor. For several decades, researchers have argued that certain candidate factors are empirically important, following the testing strategy proposed by Fama and MacBeth (1973).[1] In a first step, they estimate the factor exposures (or betas) of a set of test assets, thereby assessing the risk inherent in the asset returns. In a second step, they run a cross-sectional regression of returns on betas to test if riskier assets have higher returns on average. While this approach is very intuitive, it has been criticized from several angles and alternative estimators have been suggested. Our paper provides a comprehensive assessment of the small sample properties of a large set of estimators for linear models for the cross-section of expected asset returns.

The tests we consider are supposed to allow conclusions about the hypothesis that the price of risk (the slope coefficient in the cross-sectional regression) is equal to zero. The rejection of this hypothesis is then interpreted as evidence that a candidate risk factor is "priced" and helps to explain risk premia. A common pattern in different testing strategies is to estimate the standard error of the slope coefficient estimator and see if the t-ratio is large enough to reject the hypothesis of a zero price of risk. It has been noted that standard errors should account for the fact that the factor exposures are estimated imprecisely in the first stage. Popular textbooks and summary papers in asset pricing carefully discuss ways to do so (e.g., Cochrane, 2005; Burnside, 2011; Campbell, 2017; Ferson, 2019).

However, Kan and Zhang (1999) illustrate that in certain situations, the classic textbook approach rejects the null hypothesis too often. This problem can occur when the risk factor is "useless", referring to the case where the factor is uncorrelated with all test assets

[1]Examples include, Breeden, Gibbons and Litzenberger (1989), Jagannathan and Wang (1996), Lettau and Ludvigson (2001), Ait-Sahalia, Parker and Yogo (2004), Campbell and Vuolteenaho (2004), Parker and Julliard (2005), Santos and Veronesi (2006), Jagannathan and Wang (2007), Lustig and Verdelhan (2007), Lewellen, Nagel and Shanken (2010), Savov (2011), Menkhoff, Sarno, Schmeling and Schrimpf (2012), Boguth and Kuehn (2013), Adrian, Etula and Muir (2014), Lettau, Maggiori and Weber (2014), He, Kelly and Manela (2017), Kroencke (2017), Campbell, Giglio, Polk and Turley (2018), Barroso, Boons and Karehnke (2021) among many others.

and has a true price of risk of zero. A different but related issue is concerned with model misspecification. Traditional inference is based on the assumption that the candidate asset pricing model is perfectly specified (under the null, there is no pricing error in population). However, in applications, asset pricing models are always imperfect, meaning that they are expected to have at least some pricing error for some of the considered test assets (see Fama and French, 2015; Hou et al., 2015). This might lead to incorrect standard error estimates even when a factor is useful (Kan et al., 2013). These problems have triggered a large literature on "robust" asset pricing tests, suggesting alternative ways to perform inferences on the price of risk.

In this paper, we provide a skeptical appraisal of this literature. For this purpose we put ourselves in the shoes of an applied asset pricing researcher (rather than an econometrician) and consider realistic sample sizes (rather than asymptotics) and time series properties of real returns (rather than simulated data under special distributional assumptions). So far, there is little guidance in the literature on which method should be chosen in which situation. Running all alternative approaches is tedious and new robust approaches are continuously added to the literature, while older robust approaches remain unchallenged. Moreover, we show that this literature can be misleading, as some of the robust asset pricing tests are actually not more robust than the traditional approach in empirically relevant applications. However, the robust methods often sacrifice power to detect *useful* factors, in particular when the asset pricing model comes even with only a mild form of misspecification. Some "robust" methods are in fact surprisingly "unrobust".

For illustration, in Table 1, we consider 16 more or less robust asset pricing tests for two factors: a useful and a useless one. The test assets are the popular 25 portfolios sorted by size and book-to-market (Fama and French, 1993) and further details about the data can be found in the caption of the table. The experiment is set up such that the "useful" factor is expected to be "priced" in the cross-section of stock returns. The "useless" factor is "unpriced" by construction. As it is difficult to digest the numerical output of all 16 tests, the table only provides the conclusions on the hypothesis that the price of risk is zero (H0: $\lambda = 0$). We find that the alternative testing approaches often "agree" with inference based on

Table 1: Comparing Approaches: Testing an Empirical and a Useless Factor

This table provides inference on the price of risk (λ, in %) from the cross-sectional regression:

$$\bar{R}^e_i = (\lambda_0 +)\hat{\beta}_i \lambda + a_i$$

where $\bar{R}^e_i$ is the mean excess return of 25 Fama-French portfolios, λ_0 is an intercept (common pricing error), $\hat{\beta}_i$ is the risk factor loading, cs, R^2 is the cross-sectional coefficient of determination, MAE is the mean absolute error of the cross-sectional pricing errors (a_i). The "useful" risk factor is the tangency portfolio of the three Fama-French factors. The useless factor is uncorrelated with all test assets and has the same mean and standard deviation as the empirical factor. Under the null hypothesis, the factor sample mean is equal to the price of risk, $\mu_f = \lambda$. Below we report the conclusion of different tests of the hypothesis H0: $\lambda = 0$, as well as the conclusion of tests on the SDF loading, $SDF_t = 1 - (F_t - \mu_f)b$, H0: $b = 0$. The SDF-based tests use the identity matrix ($W = I$), or the covariance matrix of the test asset excess returns ($W = S^{-1}$) as the weighting matrix. All tests are conducted at the 5% significance level. The data are sampled monthly between Jan 1963 and Dec 2019 ($T = 672$).

sample size, $T = 672, N = 25$		**useful**	**useless**
$\hat{\mu}_f$		0.72	0.70
$\hat{\lambda}$, imposing zero intercept		0.71	-11.62
cs, R^2		0.45	-5.15
MAE		0.10	0.39
Price of risk: $reject\ H0(5\%) : \lambda = 0?\ zero\ intercept$			
$FMB - OLS$	Fama and MacBeth (1973)	yes	yes
$FMB - Shanken$	Shanken (1992)	yes	no
$FMB - GMM$	Cochrane (2005)	yes	no
$Bayesian - FMB$	Bryzgalova et al. (2020)	yes	no
$Robust\ GRS - FAR$	Kleibergen and Zhan (2020)	no	no
Price of risk: $reject\ H0(5\%) : \lambda = 0?\ with\ intercept$			
$FMB - OLS$	Fama and MacBeth (1973)	yes	no
$FMB - Shanken$	Shanken (1992)	yes	no
$FMB - GMM$	Burnside (2011)	yes	no
$Robust\ FMB$	Kan et al. (2013)	yes	no
$Bayesian - FMB$	Bryzgalova et al. (2020)	yes	no
$Three\ pass\ method$	Giglio and Xiu (2021)	no	no
$Robust\ GRS - FAR$	Kleibergen and Zhan (2020)	no	no
SDF loading: $reject\ H0(5\%)\ b = 0?$			
$SDF, W = I,\ zero\ intercept$	Cochrane (2005)	yes	no
$SDF, W = I,\ with\ intercept$	Burnside (2011)	yes	no
$SDF, W = S^{-1},\ with\ i.$	Burnside (2011)	yes	no
$Robust\ SDF, W = S^{-1},\ w.i.$	Gospodinov et al. (2014)	yes	no

the textbook approach (i.e., $FMB - Shanken/GMM$), but they sometimes also "disagree". The robust asset pricing tests also disagree with one another on the conclusion that should be drawn.

To provide a comprehensive and nuanced assessment of the practical usefulness of the alternative approaches, we conduct an extensive bootstrap simulation experiment. In a first step, in line with Kan and Zhang (1999), we ask what is the probability of a method to reject the null hypothesis that the price of risk is zero, in case the factor is useless. However, we deviate from the analysis by Kan and Zhang (1999) by considering sample sizes of common data sets in empirical applications instead of hypothetical data. Moreover, we do not exclusively study test designs where the intercept (a common pricing error) is estimated as part of the cross-sectional asset pricing model (as in Kan and Zhang, 1999), but also specifications where the intercept is imposed to be equal to its theoretical value of zero.

In general, we refer to the probability of a test to reject the H0 in case it is true as the "size" of the test. To compare different approaches, we argue that the size is an important but insufficient metric. A good approach should also have power and allow to reject the null hypothesis in case it is actually false. We point out that the robust asset pricing literature often does not at all investigate the question of power of a test. For that reason, we further extend the experiment by considering the probability of a method to reject the H0: $\lambda = 0$ for a *useful and priced* factor. This is important, as useful inference should come with a good balance between size and power.

We consider four types of model specification: i) an empirical factor that comes with the empirically observed pricing errors in the population, ii) a perfect factor that has zero pricing error in the population, iii) a factor that explains the average mean excess return but not the cross-sectional spread in mean returns, and iv) a factor that perfectly explains the cross-sectional spread in mean returns but not the average mean excess return. This way, we shed light on the question of how the different approaches work for correctly specified as well as misspecified asset pricing models.

Our main results are summarized as follows. First, for the empirically observed sample sizes, the classic Fama-MacBeth/Shanken approach does not over-reject useless factors, particularly when estimation imposes a zero intercept. The reason is that the useless factor problem is a large sample problem, but rarely applies to sample sizes asset pricing researchers work with. Moreover, the Fama-MacBeth/Shanken approach provides a (relatively) good power and allows to detect useful factors. Second, we find that some alternative approaches (e.g., Kan et al. (2013), Bryzgalova et al. (2020), pairwise bootstrap) are useful and further improve the trustworthiness in empirically relevant situations. Third, many other approaches suffer from poor power, in particular when the asset pricing model is misspecified and when the sample size or the number of test assets is limited (e.g., Gospodinov et al., 2014; Kleibergen and Zhan, 2020; Giglio and Xiu, 2021). We stress that these papers provide important theoretical results (which could help develop future alternative approaches) and can be more appropriate in other settings. By design, our study is limited to evaluate how useful a method is in common empirical applications.

Our contribution to the literature is twofold. First, we provide guidance for empirical researchers on which method should be used.[2] We provide clues on why different approaches might lead to identical or opposing conclusions. Second, we point out that the alternative methods are evaluated using very different specifications (or solely analytically) in the original publications. This makes it practically impossible to compare them to each other; or to the study by Kan and Zhang (1999) in particular. This also raises the concern that studies merely rely on specifications where a method happens to work well, even though this specification is of limited relevance for empirical work. For these reasons, we argue that the bar in the literature should be increased and future alternative robust asset pricing approaches should be evaluated more thoroughly based on empirically relevant cases.

Finally, the paper is accompanied by replication code that allows other researchers to easily implement all the methods we cover. We hope this improves transparency on the properties of different tests and will facilitate future research in this important area.

The paper proceeds as follows. In Section 2 we describe the design of our study. This includes a thorough discussion of the hypotheses that are tested, the data we use, and our method to generate different degrees of model misspecification and factor strength. Section 3 provides our results concerning the empirical performance of the different candidate tests. Section 5 gives a comprehensive overview of the main findings and concludes.

[2] The last comprehensive comparison can be found in Cochrane (2005), which does not include the alternative approaches developed since then.

2 Test design

In this paper, we analyze and compare the size and power of different tests in a coherent framework close to standard applications in asset pricing. For that purpose, we study a sample of actual financial market data and manipulate some of their properties when necessary. We then bootstrap from that sample to generate finite sample distributions of the tests. This section describes the general framework of linear asset pricing models, the construction of the data set, and the general design to study size and power of the tests.

2.1 The linear asset pricing model

Popular asset pricing theories imply a linear relationship between the assets' mean excess returns and their exposures (betas) to certain risk factors. For non-linear models, an approximate linear relationship can be often derived. We adopt the notation of Burnside (2011) and write the mean return-beta representation of linear asset pricing models as

$$E\left(\mathbf{R}_t^e\right) = \boldsymbol{\beta}\boldsymbol{\lambda}, \tag{1}$$

where $E\left(\mathbf{R}_t^e\right)$ is a $n \times 1$ vector of expected excess returns, $\boldsymbol{\beta} = cov\left(\mathbf{R}_t^e, \mathbf{F}_t\right)\boldsymbol{\Sigma}_{\boldsymbol{f}}^{\mathbf{-1}}$ is a $n \times k$ matrix of factor loadings (or betas), $\boldsymbol{\lambda}$ is a $k \times 1$ vector containing the price of risk of the $k \times 1$ vector of risk factors $\mathbf{F}_t$, which have the $k \times k$ covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{f}}$ and $k \times 1$ mean vector $\boldsymbol{\mu}_f = E(\mathbf{F}_t)$.

The mean return-beta representation of the model can be estimated following the Fama and MacBeth (1973) approach. First, estimate the factor loadings using time-series regressions

$$R_{t,i}^e = \alpha + \boldsymbol{F}_t'\boldsymbol{\beta}_i + \varepsilon_{i,t},\ t = 1, ..., T \text{ and } i = 1, ..., n \tag{2}$$

where $\boldsymbol{\beta}_i$ is a $k \times 1$ vector that stores the factor loadings of the test asset i. Second, estimate the vector of the price of risk using the cross-sectional regression

$$\bar{R}_i^e = \hat{\boldsymbol{\beta}}_i'\boldsymbol{\lambda} + a_i,\ i = 1, ..., n \tag{3}$$

where $\bar{R}_i^e = \frac{1}{T}\sum_{t=1}^{T} R_{t,i}^e$ is the sample average excess returns of the test assets. The OLS estimator for the price of risk is obtained as $\hat{\boldsymbol{\lambda}} = \left(\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}\right)^{-1}\hat{\boldsymbol{\beta}}'\bar{\mathbf{R}}^e$, where $\bar{\mathbf{R}}^e$ is a $k \times 1$ vector of average excess returns. In the remainder of the paper, we compare systematically different approaches that have been proposed to find the standard errors of the price of risk, $s.e.(\hat{\boldsymbol{\lambda}})$.

An alternative approach includes an intercept λ_0 in the cross-sectional regression model

$$\bar{R}_i^e = \lambda_0 + \hat{\boldsymbol{\beta}}_i'\boldsymbol{\lambda} + a_i,\ i = 1, ..., n, \tag{4}$$

and the OLS estimator for the intercept and the price of risk is obtained as $\hat{\boldsymbol{\theta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\bar{\mathbf{R}}^e$, where $\hat{\boldsymbol{\theta}} = \left[\lambda_0,\ \hat{\boldsymbol{\lambda}}'\right]'$, $\mathbf{X} = \left[\mathbf{1}_{n\times1},\ \hat{\boldsymbol{\beta}}\right]$, and $\mathbf{1}_{n\times1}$ denotes a $n \times 1$ vector of ones.

An equivalent way of phrasing linear factor models is to say that the stochastic discount factor (SDF) is affine-linear in the candidate factors. More precisely, linear factor models imply that the Euler Equation

$$E\left(\mathbf{R}_t^e M_t\right) = E\left(\mathbf{R}_t^e\left[1 - \left(\mathbf{F}_t - \boldsymbol{\mu}_f\right)'\mathbf{b}\right]\right) = 0, \tag{5}$$

holds for all test asset returns, where M_t is the SDF. The SDF is normalized to have a mean of one. The $k \times 1$ vector of SDF-loadings $\mathbf{b}$ are linked to the price of risk via $\boldsymbol{\lambda} = \boldsymbol{\Sigma_f}\mathbf{b}$. As described in Cochrane (2005) and Burnside (2011), using the identity weighting of the moment restrictions allows to set up a one-step GMM estimator that leads to the *same* point estimates as using the Fama-MacBeth approach.

2.2 Estimating the price of risk

In this paper, we investigate the small sample properties of alternative testing strategies for the hypothesis that a risk factor is "unpriced" in the cross-section of stock returns, that means H0: $\lambda = 0$. This is arguably the most important question asked in applied research in asset pricing. Without loss of generality, in the following, we consider one factor models, i.e. $k = 1$, so that $\boldsymbol{\lambda} = \lambda$. This will simplify the notation and discussions.

Model without intercept: First, we consider the hypothesis H0: $\lambda = 0$ when the model imposes a zero intercept:

$$E\left(\mathbf{R}_t^e\right) = 0 + \boldsymbol{\beta}\lambda. \tag{6}$$

This hypothesis is motivated by asset pricing theory which usually suggests that expected excess returns are proportional to factor exposures, so the intercept in the linear relation should be equal to zero. Constraining the intercept to zero has been recommended, e.g., by Cochrane (2005).

Economically, the price of risk (λ) can be interpreted as the mean return of a portfolio that has a factor loading (β) of exactly one (Fama, 1976). Intuitively, the H0 should be rejected in favor of a positive λ if assets with high betas *relative to a benchmark beta of zero* have high returns on average *relative to a benchmark return of zero*. A significant price of risk does not imply, however, that the factor model can explain well the cross-sectional dispersion of average returns. Rejecting the H0 should not be confused with a correct model specification. To achieve a good cross-sectional fit, the magnitude of factor betas would also have to line up with the magnitude of the test asset mean excess returns. This fact motivates metrics of cross-sectional fit and tests of pricing errors, which are outside the scope of this paper (see, e.g., Cochrane (2005)).[3]

Model with intercept: Second, we also consider the H0: $\lambda = 0$, where the candidate model comes with a cross-sectional intercept, λ_0:

$$E\left(\mathbf{R}_t^e\right) = \lambda_0 + \boldsymbol{\beta}\lambda, \tag{7}$$

While asset pricing theory implies that the intercept should be equal to zero, the intercept is often included to allow all test asset returns to be mispriced by a common amount (e.g. Parker and Julliard, 2005).

The sum of the two parameters λ_0 and λ can be interpreted as the mean return on a

[3] An empirical example of a risk factor that is priced but comes with poor cross-sectional fit is testing the CAPM with a zero intercept using portfolios sorted by market equity and the book-to-market ratio of equity as test assets.

portfolio that has a factor loading of exactly one. Intuitively, H0 should be rejected in favor of a positive λ if assets with high betas *relative to the other test assets* have high returns on average *relative to the other test assets*. Thus, again, rejection of H0 should not be confused with a statement about correct model specification, since asset pricing models typically also imply that the intercept λ_0 should be equal to zero.

Estimation with or without the intercept? For perfectly specified models, both test specifications should lead to the identical conclusion. However, the two test specifications allow to account for *different* forms of model misspecification. To illustrate the difference between estimation with and without an intercept, we consider in this paper two empirically relevant situations where estimation with and without the intercept will give seemingly conflicting results.

The first situation occurs when the test assets have large absolute betas and also large absolute expected returns, but there is no cross-sectional relation between them. In this case, the specification with imposing a zero intercept should indicate a significant price of risk (which is economically similar to an intercept), while the specification with estimating the intercept should indicate an insignificant price of risk, as the intercept absorbs the common mean return of the test assets. We call this situation "slope misspecified" in the following.[4]

The second situation occurs when the test asset mean excess returns perfectly line up with the factor loadings, the mean factor loadings are non-zero, but the mean of the test asset mean excess returns is zero. In this case, the specification with imposing a zero intercept should indicate an insignificant price of risk (because the mean of the mean excess returns is zero), while the specification with estimating the intercept should indicate a significant price of risk. We call this situation "intercept misspecified" in the paper.[5]

While there are arguments for estimation with as well as estimation without an intercept, we do not take a strong view of what specification should be chosen based on economic grounds. However, in this paper, we will shed light on whether the two specifications come

[4]An empirical example of such a case can be found, e.g., in Savov (2011).

[5]An empirical example of such a case can be found in studies of currency risk premia, e.g., in Burnside (2011).

with a different size or power. To the best of our knowledge, the difference in the size and power of the two specifications has not yet been analyzed in a comprehensive manner.

2.3 Test assets and factors

We are interested in whether a method for inference on the price of risk has good properties in actual applications, rather than under laboratory conditions. We thus choose to work with real data instead of simulated data whenever possible.

As test assets we use the 25 portfolios sorted by market equity and book-to-market ratio of equity, as suggested by Fama and French (1993).[6] We subtract the 1-month treasury bill rate to calculate excess returns. Our sample spans from January 1964 till December 2019, i.e. the monthly sample includes 672 return observations per asset. We also construct data sets on the quarterly and annual frequency, with sample sizes of 224 and 56 observations, respectively. This procedure is motivated by the fact that many candidate risk factors are only observable at the quarterly or annual frequency.

Our goal is to analyze the properties of asset pricing tests for different types of factors, in particular strong factors, i.e., factors that are strongly correlated with test asset returns in the time series, and weak factors for which the correlation with the test assets is on average low. Since it is easier to make factors weaker (by adding noise to the factor time series) than to make them stronger, we choose a strong factor as the starting point of our analysis. The Fama-French 3-factor model (see Fama and French, 1993) is based on factors that are essentially less granular sorts by market equity and book-to-market ratio. Therefore, they pick up a large part of the time series variation in the test asset returns.

To facilitate the subsequent analyses we construct a single factor from the three factors $\mathbf{F}_t^{\mathbf{FF3}} = [MKT_t, SMB_t, HML_t]'$. Since the factors are excess returns themselves, we can construct the tangency portfolio $F_t = c \times \mathbf{F}_t^{\mathbf{FF3}\prime}\mathbf{w}$ with weights

$$\mathbf{w} = \boldsymbol{\Sigma}_{\boldsymbol{f}}^{\boldsymbol{FF3}^{-1}} \boldsymbol{\mu}^{\mathbf{FF3}} \left(\mathbf{1}_{1\times 3} \boldsymbol{\Sigma}_{\boldsymbol{f}}^{\boldsymbol{FF3}^{-1}} \boldsymbol{\mu}^{\mathbf{FF3}} \right)^{-1}, \tag{8}$$

[6] All data are downloaded from Kenneth French's website at https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html.

where $\mathbf{1}_{1\times3}$ denotes a 1×3 vector of ones, and c is a scaling parameter.[7] By construction, the tangency portfolio has the same explanatory power as the original factors combined (Huberman and Kandel (1987), Gibbons et al. (1989)). In the monthly data set the weights are given by $\mathbf{w} = (0.33, 0.17, 0.52)'$ and those corresponding to the quarterly and annual data are very similar.

Because our candidate risk factor is traded, it must price itself. This implies that the price of risk must be equal to the factor mean return, $\lambda = \bar{F} = \frac{1}{T}\sum_{t=1}^{T} F_t$.

2.4 Properties of the data set

The estimated time series correlation of the factor F_t with the test asset returns is 0.69 on average on the monthly sample. This means that for the average test asset, about half of the time-series variation in returns can be explained by the asset's exposure to the factor. There is a pronounced spread in correlation coefficients. On the monthly sample, the correlation coefficients range from 0.41 for the big growth portfolio to 0.85 for the value portfolio in the second lowest size quintile. We see a similar cross-sectional variation in betas. While the average beta is 1.00 on the monthly sample, the highest (1.33) and lowest (0.49) betas are again those with the highest and lowest correlations. The numbers are very similar for the quarterly and annual frequency. Given these properties, we can conclude that the factor F_t is strong vis-a-vis the test asset returns, and that there are pronounced cross-sectional differences in factor exposures.[8]

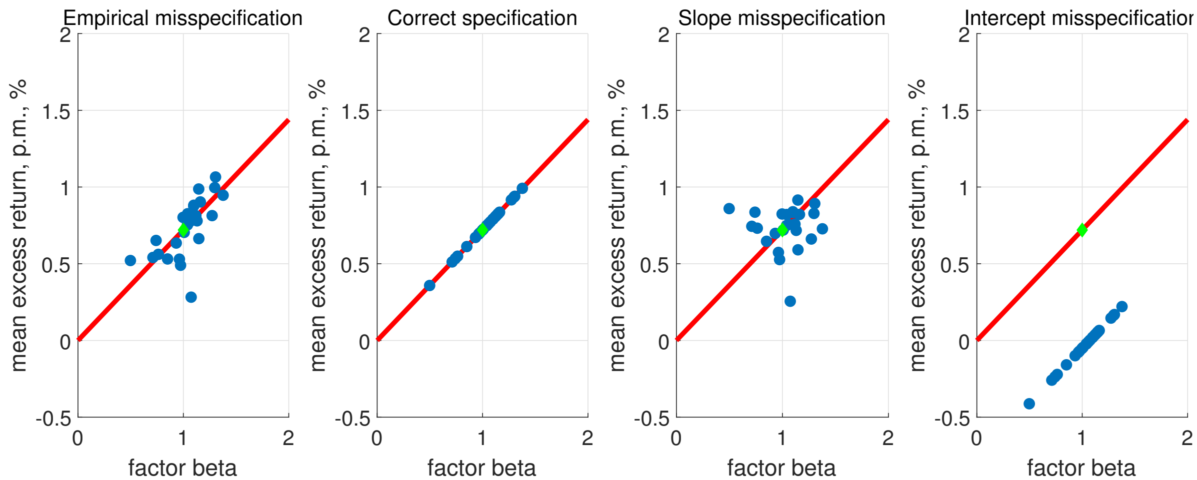
Regressing average returns on betas yields an R^2 of 0.45 on the monthly sample (0.53 and 0.55 on the quarterly and annual sample, respectively). The average excess returns $\bar{R}_i^e$ are plotted against factor betas β_i in the left graph of Figure 1.

For the monthly data, we estimate a price of risk of $\lambda = 0.71$ ($t_{Shanken} = 3.8$) when imposing a zero intercept. Estimation with an intercept gives an estimate of the price of risk

[7]Because the Fama-French factors are all excess returns, multiplying $\mathbf{w}$ with any nonzero constant leads to a factor with identical pricing properties and, ultimately, identical statistical inference. We set the scaling parameter to $c = 2$, which implies that the betas of our test assets vary around 1. This convention makes some of our figures easier to interpret.

[8]In our simulation experiment, we will impose the empirical betas as the population betas. For that reason, it is not necessary to test for the joint significance of the betas.

Figure 1: Explanatory Power of Our Risk Factors



This figure plots the test asset mean excess returns $\bar{R}_i^e$ against their factor loadings β_i. In all subfigures, the risk factor is the tangency portfolio of the Fama-French three-factor model. In the left figure, the test assets are the 25 Fama-French portfolios sorted by market equity and the book-to-market ratio. The second figure shows results when we modify the mean test asset returns such that the risk factor has perfect cross-sectional explanatory power. The third figure shows results when we modify the mean test asset returns such that the risk factor has zero cross-sectional explanatory power. The right figure shows results when we modify the mean test asset returns such that the risk factor has perfect cross-sectional explanatory power but the intercept is different from zero. The construction of the latter three cases is described in Section 2.5. The sample is from 01/1964 to 12/2019 with 672 monthly observations.

of $\lambda = 0.63$ ($t_{Shanken} = 3.1$) and a common pricing error of $\lambda_0 = 0.08$ ($t_{Shanken} = 0.3$). The estimated price of risk is close to the factor mean of $\bar{F} = 0.72$, i.e. close to the model implied price of risk.

As can be easily seen in Figure 1 (left graph), the model comes with pricing errors for some portfolios. Thus, under the assumption that our sample is representative for the population, the model is misspecified. In our analysis, we will consider several cases which differ with respect to the degree of model misspecification. One important case will be the case of “empirical misspecification”, i.e., we leave the properties of the original data set untouched. Thus, whenever we consider this case in our analysis, we explicitly impose that the model is misspecified in population, i.e., the variation in factor exposures does not perfectly explain the variation in expected returns. All model specification tests should thus reject the model. Still, the factor has some explanatory power and a useful asset pricing test should acknowledge that fact, given that all models are necessarily misspecified.

2.5 Scaling the degree of model misspecification

To allow tests of asset pricing models with various degrees of misspecification, we manipulate the pricing errors. We first write the excess returns on asset i as $R^e_{i,t} = \bar{R}^e_i + \varepsilon_{i,t}$ with a mean zero component $\varepsilon_{i,t}$. We then perform a cross-sectional regression with intercept to decompose $\bar{\mathbf{R}}^e = (\bar{R}^e_1, \ldots, \bar{R}^e_{25})'$ into model-implied returns, common pricing error, and individual pricing errors:

$$\bar{R}^e_i = \lambda_0 + \lambda\beta_i + a_i \tag{9}$$

To generate alternative populations, we set

$$R^{e,\kappa}_{i,t} = \kappa_0 + \kappa_\beta\, \lambda\beta_i + \kappa_{pe}\, a_i + \varepsilon_{i,t}. \tag{10}$$

By varying the parameters $\kappa = (\kappa_0, \kappa_\beta, \kappa_{pe})'$, we can control the cross-sectional distribution in average returns and the share of explained variation in total variation. At the same time, the time series properties (except the mean) of returns stay untouched. Most importantly, the betas are identical across different choices of κ, because they are only determined by the covariance of ε_i with the factor F and the variance of F, which we leave unaltered.

We consider four different choices of κ. In the first, we set $\kappa_0 = \lambda_0$, $\kappa_\beta = 1$, and $\kappa_{pe} = 1$ and, by that, recover the properties of the original sample. In this case, the model is misspecified but the factor exposures explain about half of the variation in expected returns (see Section 2.4). We refer to this calibration as the case of *empirical misspecification*.

In the second specification, we set $\kappa_0 = 0$, $\kappa_\beta = 1$, and $\kappa_{pe} = 0$. In this case, common and individual pricing errors are equal to zero, which means that in population, expected returns perfectly line up with betas. We refer to this calibration as the case of *correct specification*. The second graph in Figure 1 shows expected returns plotted against population betas for this specification, using the monthly sample.

Third, we set $\kappa_0 = \frac{1}{n}\sum_{i=1}^{n} \bar{R}_i$, $\kappa_\beta = 0$ and $\kappa_{pe} = 1$. This specification corresponds to the case where the factor is perfectly unpriced, in the sense that the population betas do not explain any of the cross-sectional variations in expected returns. Importantly, expected returns and population betas are positive throughout, so the factor exposures can explain the level of the expected returns. Still, in a setting with intercept, the slope λ cannot be identified, so we refer to this calibration as the case of *slope misspecification*. The third graph in Figure 1 shows expected returns plotted against population betas for this specification.

Finally, we set $\kappa_0 = -\frac{\beta'\beta}{\mathbf{1}_{1\times n}\beta}\lambda$, $\kappa_\beta = 1$, and $\kappa_{pe} = 0$. Since $\kappa_\beta = 1$ and $\kappa_{pe} = 0$ the expected returns are affine linear in betas. However, since $\kappa_0 \neq 0$, the model is still misspecified, since the intercept in the cross-sectional regression is different from zero. The choice of κ_0 implies that the slope coefficient λ in a cross-sectional regression without intercept is equal to zero in population. Intuitively, all betas are positive but some of the expected returns are negative, so the average level of returns cannot be explained by the level of the factor exposures. Alternatively, one could consider the case where expected returns are positive but the betas vary around zero.[9] We refer to this calibration as the case of *intercept misspecification*. The fourth graph in Figure 1 shows expected returns plotted against population betas for this specification.

2.6 Finite sample distributions and useless factors

To study the small sample properties of the different model tests, we bootstrap B samples from the original samples. More precisely, for each $b = 1, \dots, B$, we draw a set of time indexes $\mathcal{T}_b = \{t_1^b, \dots, t_T^b\}$ with replacement, where T is the sample size of the original sample. Moreover, for each b, we independently draw a second set of time indexes $\widetilde{\mathcal{T}}_b$. Then, the b^{th} bootstrapped sample consists of the risk factor

$$\mathbf{F}^b = \rho F_{t\in\mathcal{T}_b} + \sqrt{1-\rho^2} F_{t\in\widetilde{\mathcal{T}}_b}, \tag{11}$$

where $\rho \in [0, 1]$ is a meta parameter, which determines the correlation coefficient between the "observed" risk factor and the "true" risk factor. We re-sample the excess returns ($\mathbf{R}^e_{t\in\mathcal{T}_b}$) using only the first set of time indexes $\mathcal{T}_b$. Thus, setting $\rho = 1$ preserves the joint time series properties of returns and factors. In the extreme case where $\rho = 0$, the factor is "useless", in the sense that it is uncorrelated with returns in population.

[9] An empirical example of such a case can be found when considering currency returns, see e.g. Burnside (2011).

Note that in the "useless" factor case, the factor has no pricing ability by construction. Because the useless factor is uncorrelated with all possible asset returns and, thus, the population betas of all test assets are equal to zero, it is not possible to identify the market price of risk λ.[10] Accordingly, the estimated standard error for λ should be so large (it is infinite in population) that H0: $\lambda = 0$ should not be rejected by asset pricing tests. It has been shown in the literature that some asset pricing tests still reject H0 too often, which is referred to as the useless factor problem.[11]

2.7 Power curves for testing "priced" risk factors

For each of the bootstrapped samples, we perform the candidate asset pricing tests and count, how often they reject H0: $\lambda = 0$ on the 5% level.[12] Intuitively, the tests ask whether a risk factor is "priced" in the cross-section of stock returns. This is arguably the first and most common question asked in applied research.

We start with the candidate risk factor that is perfectly uncorrelated with all assets ($\rho = 0$). This case reprises the "useless factor" scenario considered by Kan and Zhang (1999). A reliable test should have the correct *size*, meaning that it rejects 5% of the time (type I error) in this scenario. A test is unreliable if it rejects the H0: $\lambda = 0$ much more often than in 5% of the cases. A researcher would conclude too often that a risk factor is "priced" when it is actually unrelated to the test asset returns.

We then increase the levels of factor strength, with ρ varying between 0 and 1. The case of $\rho = 1$ recovers the "true" risk factors with the properties described in Section 2.5. A reliable test should have a high *power* to detect "priced" risk factors and, optimally, reject the H0: $\lambda = 0$ in 100% of the samples (small type II error), if λ is really different from zero. A good test should be well balanced between a correct size and high power. For example, a size correct test with low power is arguably less useful in applied research than as a test

[10]Because the useless factor is uncorrelated with all risky assets by construction (not just the tested assets), we know that our useless factor is also a zero-beta asset, which implies a price of risk of $\lambda = 0$. However, this argument cannot be applied to empirically observed "useless" factors.

[11]See Kan and Zhang (1999).

[12]For most tests, we bootstrap 10,000 samples. We bootstrap 1,000 times when the method is computationally demanding, as discussed below.

that has a slightly incorrect size but high power to detect useful factors.

To report our findings, we show *power curves*, in which we plot the relative rejection frequency against ρ. For each test we show four lines, one for each type of model misspecification as explained in Section 2.5. Moreover, we show three graphs for each test and consider three different frequencies and sample sizes (monthly data with T=672, quarterly data with T=224, and annual data with T=56).

To sum up, a reliable test should produce power curves with the following properties:

- The power curve should always start on the left at 5% for the useless factor ($\rho = 0$). Note that all models are misspecified by definition in this "corner" case.
- In the cases of empirical misspecification and correct specification, the power curve should increase when moving from the left ($\rho = 0$) to the right ($\rho = 1$). The power curves should be close to 1 when ρ is reasonably large.
- In the case of slope misspecification, the power curve should increase when the estimation is performed without an intercept. For large values of ρ, the factor is correlated with the test asset returns in the time series, but there is no spread in the factor exposures that lines up with mean returns. The estimate of λ might be interpreted as an intercept (a common mean return component explained) in this specific scenario.
- In the case of slope misspecification, the power curve should, however, remain at 5% when an intercept is included. Because there is no spread in factor loadings that lines up with mean returns, the estimate of λ is expected to be zero, so the H0 should not be rejected.
- In the case of intercept misspecification, the power curve should remain at 5% when the estimation is performed without an intercept. By construction of the common pricing error, the population λ is equal to zero in the framework that imposes a zero intercept, so the H0 should never be rejected.
- In the case of intercept misspecification, the power curve should, however, increase when an intercept is included. In this specification, the true expected returns are affine linear

in the betas and the nonzero common pricing error that makes this model misspecified should be taken out by the intercept in the cross-sectional regression.

The problem of identification: It is important to stress that in the scenario of a useless factor ($\rho = 0$) the price of risk λ (or the SDF loading b) is NOT identified. This means that it is not possible to conduct inference about the location of the true parameter (λ). For some of the tests considered in Section 3, the confidence bands constructed from the standard errors would lose their usual interpretation.

We believe that this aspect is not an issue in applications, as long as the hypothesis H0: $\lambda = 0$ is rejected with the correct size. If this is the case, it would still be possible to conclude with the correct likelihood that the factor is (not) "priced". A factor that is not "priced" is economically uninteresting, no matter if it is useless or strong but unpriced, and no further inference on the parameter is of interest anyway.

Alternative power curves: In our power curves, we fix the properties of the test assets and vary the risk factors, i.e. the factor loadings. We make transparent how the probability to reject the the H0: $\lambda = 0$ changes when moving from a "true" risk factor to a "useless" risk factor. In addition, we consider different forms of model misspecification as well as "true" prices of risk.

An alternative way to draw power curves is a data generating process where the factor loadings are fixed and simply the price of risk is varied.[13] We believe that our framework is more appropriate. First, with fixed factor loadings, it is impossible to analyze useless factors. Using a design that is silent on the useless factor problem is insufficient when some of the analyzed methods are explicitly motivated to be (more) "robust" to the useless factor problem. Second, in the case of traded risk factors, varying the price of risk (factor mean returns) but fixing the factor loadings and the test asset mean excess returns would lead to an arbitrage opportunity. Covariation *is* the risk that is compensated by a factor premium (Cochrane (2011)). Such a data generating process can be inconsistent with any asset pricing theory that rests on a no-arbitrage condition. On the other hand, if the test asset mean excess

[13] See, for example, Kleibergen (2009) and Kleibergen and Zhan (2020).

returns are not fixed, the properties of the left-hand side variable can change arbitrarily between specifications, rendering meaningful comparisons difficult.

3 Alternative standard error estimators

This section analyses the size and power of different testing strategies that researchers employ when asking whether a risk factor is "priced" in the cross-section of stock returns. We delegate most of the formulas and technical details to the appendix of the paper.

3.1 Fama-MacBeth

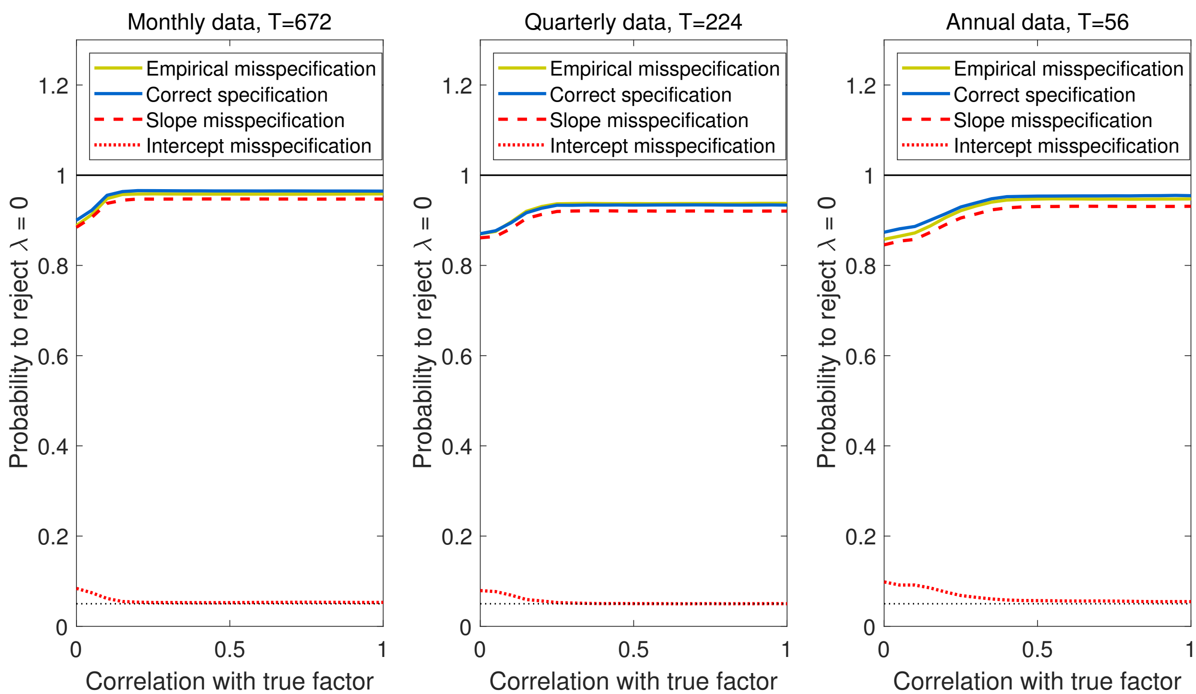
We start with the classic Fama and MacBeth (1973)-approach to compute standard errors. Because the regressors are constant over time, these standard errors are equivalent to OLS standard errors from the cross-sectional regression in Equations (6) and (7), see, e.g., Cochrane (2005) and Burnside (2011). Fama-MacBeth/OLS standard errors are robust to cross-sectional correlation, but they do not account for the fact that the betas are estimated. As shown in Shanken (1992), this implies that standard errors will be too small when betas are estimated imprecisely.

Indeed, as can be seen in Figure 2, the classic standard errors heavily over-reject the useless factors in the first three scenarios of model misspecification.[14] For example, when testing at the 5% significance level, we reject the H0: $\lambda = 0$ about 90% of the time when the risk factor is actually useless. The results are very similar when the estimation is performed with an intercept. For that reason, we do not report the respective results.

We conclude that the classic FMB/OLS standard errors should not be used in asset pricing applications. It is now interesting to investigate whether the over-rejection is driven by the "useless" factor problem (Kan and Zhang, 1999), or by the fact that the betas are estimated in the first-pass regression (Shanken, 1992). To disentangle the two sources of bias, we next turn to Shanken-corrected standard errors.

[14] In the case of intercept misspecification and $\rho = 0$, the population betas and the population expected returns are close to zero. In this case, the test does not overreject. In the other three cases, population betas are equal to zero but expected returns are positive.

Figure 2: Power Curves: Fama-MacBeth with Shanken Correction, Fama-MacBeth/OLS , Imposing a Zero Intercept



Method: This figure presents the probability to reject the null hypothesis $\lambda = 0$ (5% significance level) of the cross-sectional regression model:

$$\bar{R}_i^e = 0 + \hat{\beta}_i \lambda + a_i$$

where $\bar{R}_i^e$ is the mean excess return of test asset i, $\hat{\beta}_i$ is the factor loading to the risk factor estimated in the first-pass regression, λ is the price of risk, and a_i is a cross-sectional pricing error. The probability to reject is based on 10,000 pairwise bootstrap re-samples.

- Standard errors for λ are based on the classic Fama-MacBeth approach and do not account for the fact that betas are estimated.
- The intercept (a common pricing error) is imposed to be zero.

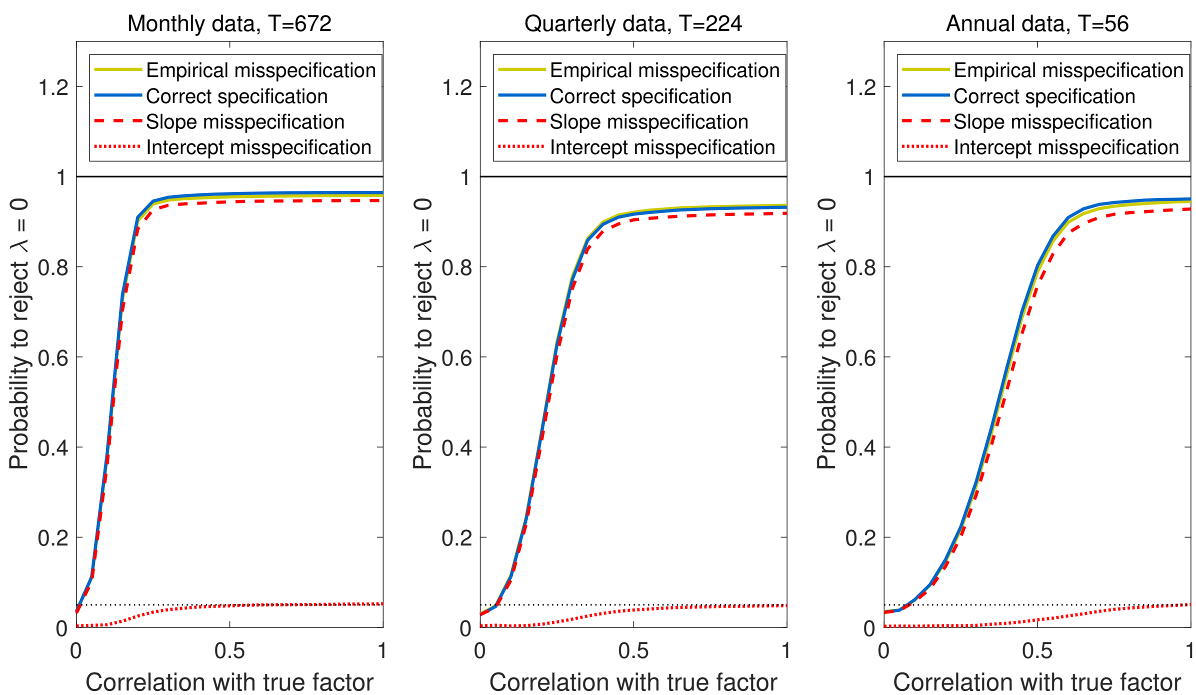
Test assets and risk factors: The true risk factor is the tangency portfolio of the three Fama-French factors. The factor loadings of the test assets are computed with respect to the true factor ($\rho = 1$), or a useless factor that has population correlation of zero to all test assets ($\rho = 0$), as well as intermediate cases when the factor is a combination of the true and the useless factor. The yellow lines (empirical misspecification) correspond to the original test assets; the 25 Fama-French portfolios sorted by market equity and the book-to-market ratio. The blue line modifies the test asset mean returns such that the population pricing error is equal to zero (correct specification). The red lines show variations where the test assets are modified such that the true cross-sectional slope is zero (slope misspecification), or the true intercept is non-zero and the average mean return of the test assets is zero (intercept misspecification). Section 2 and Figure 1 describe the properties of the test assets and the risk factors in detail.

3.2 Fama-MacBeth with Shanken correction

It is well understood that the FMB/OLS standard errors are problematic. For that reason, popular textbooks in asset pricing (e.g., Cochrane, 2005; Campbell, 2017; Ferson, 2019) recommend to correct standard errors for the fact that betas are estimated.

Figure 3 reports the probability to reject the H0: $\lambda = 0$, when inference is based on Fama-MacBeth standard errors, accounting for the fact that the betas are estimated. These standard errors are often referred to as Shanken (1992)-corrected standard errors in the literature. The intercept is imposed to be zero.

Figure 3: Power Curves: Fama-MacBeth with Shanken Correction, Imposing a Zero Intercept



Method: This figure presents the probability to reject the null hypothesis $\lambda = 0$ (5% significance level) of the cross-sectional regression model $\bar{R}_i^e = 0 + \hat{\beta}_i \lambda + a_i$. The probability to reject is based on 10,000 pairwise bootstrap re-samples.

- Standard errors for λ account for that the betas are estimated (Shanken (1992)-corrected).
- The intercept (a common pricing error) is imposed to be zero.

Test assets and risk factors: The risk factors and the test assets are the same as described in the caption to Figure 2.

First, we find that Fama-MacBeth/Shanken standard errors do not over-reject the useless factors in any scenario, as the rejection frequency is always close to 0.05 when $\rho = 0$. This is noteworthy, given that some authors have criticized the textbook approach for the estimation of models with weak factors (for example Kleibergen and Zhan (2020) argue that the approach is not "trustworthy" and suffered from a "malfunction", even in small samples). Other authors (e.g. Kan and Zhang, 1999) have noted that the useless factor problem is a large sample rather than a small sample problem. Sample sizes of the empirically relevant cases are usually small, such that there is no (or only modest) over-rejection as shown in Figure 3.

Second, the Fama-MacBeth/Shanken-approach is powerful. Reasonably strong risk factors ($\rho = 50\%$ and higher) are concluded to be priced in the cross-section of stock returns with a probability of 90% or more, even in the annual data sample that comes with small sample size.

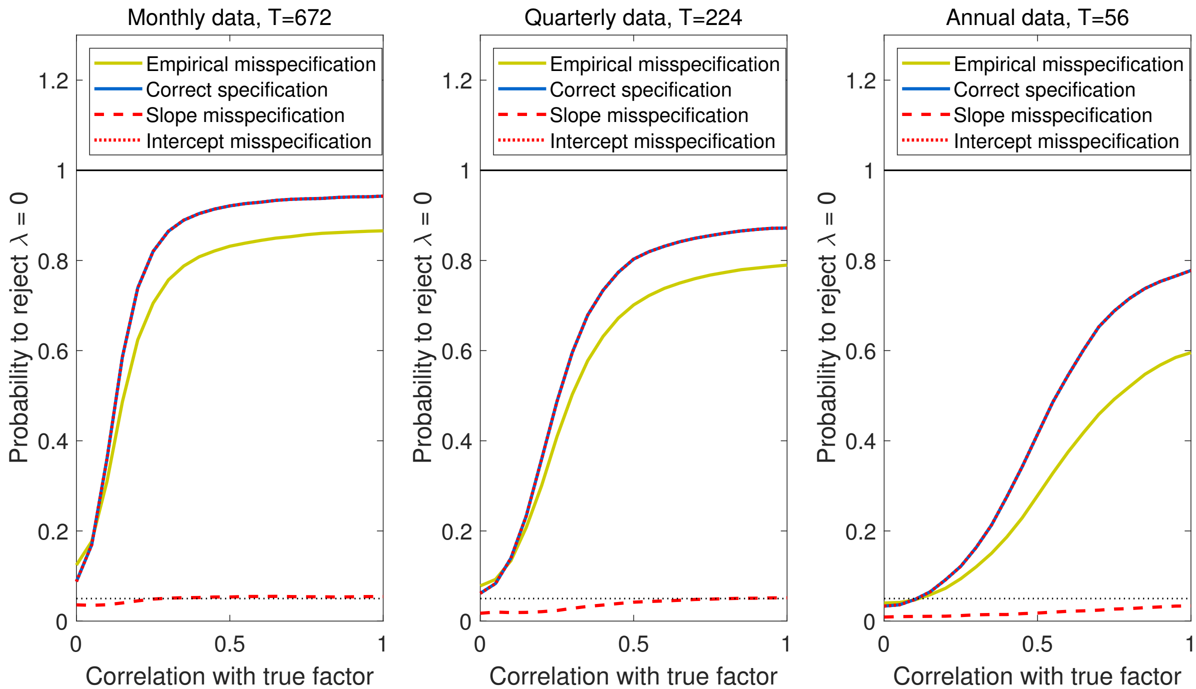
Third, there is no noticeable difference in the power curves between (mildly) misspecified models and the perfectly specified model. As discussed in Section 2.2, in case of slope misspecification, H0 should also be rejected, because estimation is conducted without an intercept.[15] In the case of intercept misspecification, however, H0 is actually true in population and we find that the test has the correct size in this case.

In Figure 4, we estimate a cross-sectional model that features an intercept. Two things are noteworthy. In the case of a useless factor, we find that the Fama-MacBeth/Shanken approach rejects H0 with a probability of around 10% when using monthly data. The inclusion of an intercept thus tends to increase the useless factor problem. Again, the issue is less of a problem for smaller sample sizes (quarterly or annual data).

Furthermore, we find that the power deteriorates in all scenarios, in particular for mildly misspecified models (yellow line). Because the intercept is estimated, the slope is solely identified by the spread in factor loadings. This spread is naturally more noisy for misspecified models. Accordingly, we now find that H0 is usually not rejected in case of the slope misspecification, because the betas are not related with the average returns in the cross-

[15] A rejection of H0 indicates that the risk factor is significantly positively correlated with the returns on the test assets such that the factor exposures can explain the positive cross-sectional average mean return. However, there is no spread in these correlations such that the risk factor cannot explain the cross-sectional spread in mean returns.

Figure 4: Power Curves: Fama-MacBeth with Shanken Correction, with Estimation of the Intercept



Method: This figure presents the probability to reject the null hypothesis $\lambda = 0$ (5% significance level) of the cross-sectional regression model $\bar{R}_i^e = \lambda_0 + \hat{\beta}_i \lambda + a_i$. The probability to reject is based on 10,000 pairwise bootstrap re-samples.

- Standard errors for λ account for that the betas are estimated (Shanken (1992)-corrected).
- The intercept (a common pricing error) λ_0 is estimated.

Test assets and risk factors: The risk factors and the test assets are the same as described in the caption to Figure 2.

section, i.e., the population slope is equal to zero. In the case of intercept misspecification, the rejection frequencies are perfectly identical to the case of correct specification (making the power curves appear purple).

Taken together, our results show that using the Fama-MacBeth/Shanken approach is a very good choice in small samples, as the test is size-correct and reasonably powerful for useful factors. Moreover, the test which imposes a zero intercept considerably reduces the useless factor problem and increases power to detect useful and priced factors, relative to the approach with intercept. To the best of our knowledge, this fact has been unnoticed in the literature thus far.[16]

3.3 Fama-MacBeth, with Shanken correction and HAC-robust

Figures 5 and 6 provide results when the estimation approach for the standard errors does not only account for the fact that the betas are estimated (as do the Fama-MacBeth/Shanken-standard errors) but also for possible heteroskedasticity and autocorrelation (HAC). Cochrane (2005) and Burnside (2011) show how to derive these standard errors from a more general GMM framework. Although covered in one of the major textbooks, these GMM formulas still do not seem to be well known.

Instead, it is popular to apply the Newey-West approach to the time-series of the Fama-MacBeth estimates to account for autocorrelation. This approach, however, does not account for the fact that the betas are estimated, which is often more relevant for the standard errors than autocorrelation.

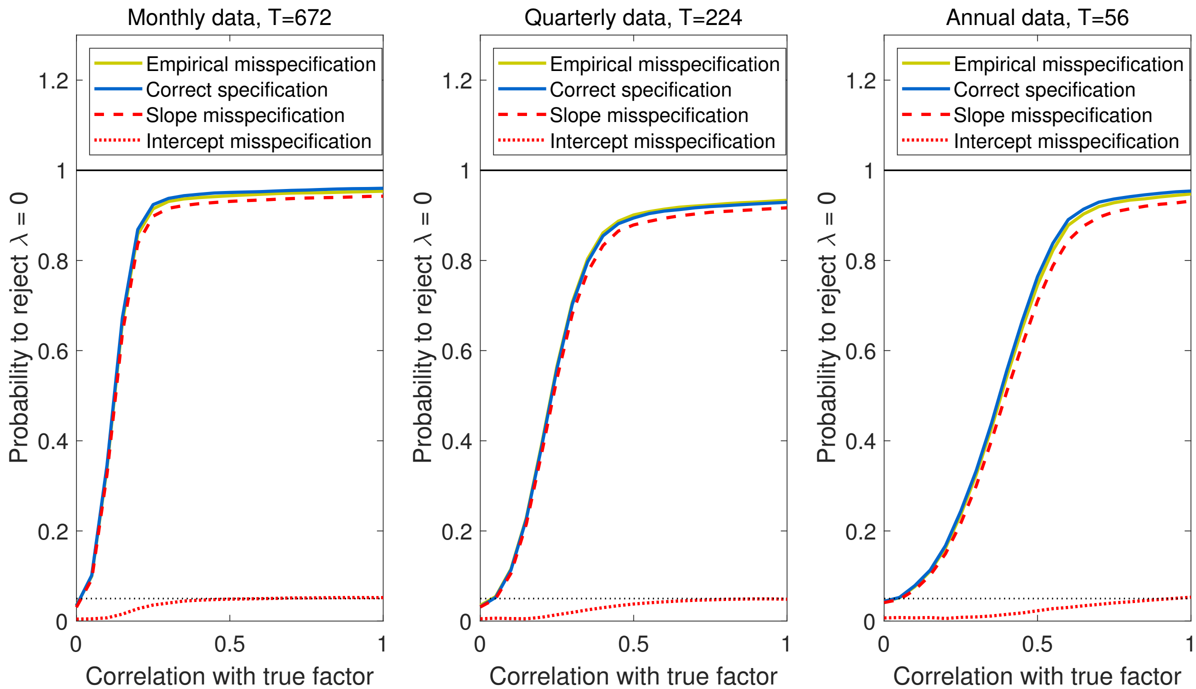
We find that the HAC-robust modifications lead to similar power curves compared to the standard errors with the Fama-MacBeth/Shanken-approach. This was expected, because there is no autocorrelation in the test asset returns and factors of our bootstrapped samples. Still, if a researcher suspects that heteroskedasticity and autocorrelation could play an important role in their data, this approach is preferable to the Fama-MacBeth/Shanken-approach.

3.4 Misspecification Robust Standard Errors

The classic Fama and MacBeth (1973) standard errors (and their refinements discussed so far) assume that the asset pricing model is correctly specified. Kan et al. (2013) derive standard errors for cross-sectional regressions when the model is misspecified. This means that the researcher expects that the factor model has at least some non-zero pricing errors

[16] One reason for this circumstance may be that Kan and Zhang (1999) only provide results when the estimation is conducted with an intercept. In addition, most of the presented tables do not incorporate the Shanken correction. Many of their results therefore mix several problems, i.e. the useless factor problem, the estimated betas problem, and estimation with intercept.

Figure 5: Power Curves: Fama-MacBeth, HAC-robust with Shanken Correction, Imposing a Zero Intercept



Method: This figure presents the probability to reject the null hypothesis $\lambda = 0$ (5% significance level) of the cross-sectional regression model $\bar{R}_i^e = 0 + \hat{\beta}_i \lambda + a_i$. The probability to reject is based on 10,000 pairwise bootstrap re-samples.

- Standard errors for λ account for that the betas are estimated (Shanken (1992)-corrected) and are HAC-robust.
- The intercept (a common pricing error) is imposed to be zero.

Test assets and risk factors: The risk factors and the test assets are the same as described in the caption to Figure 2.

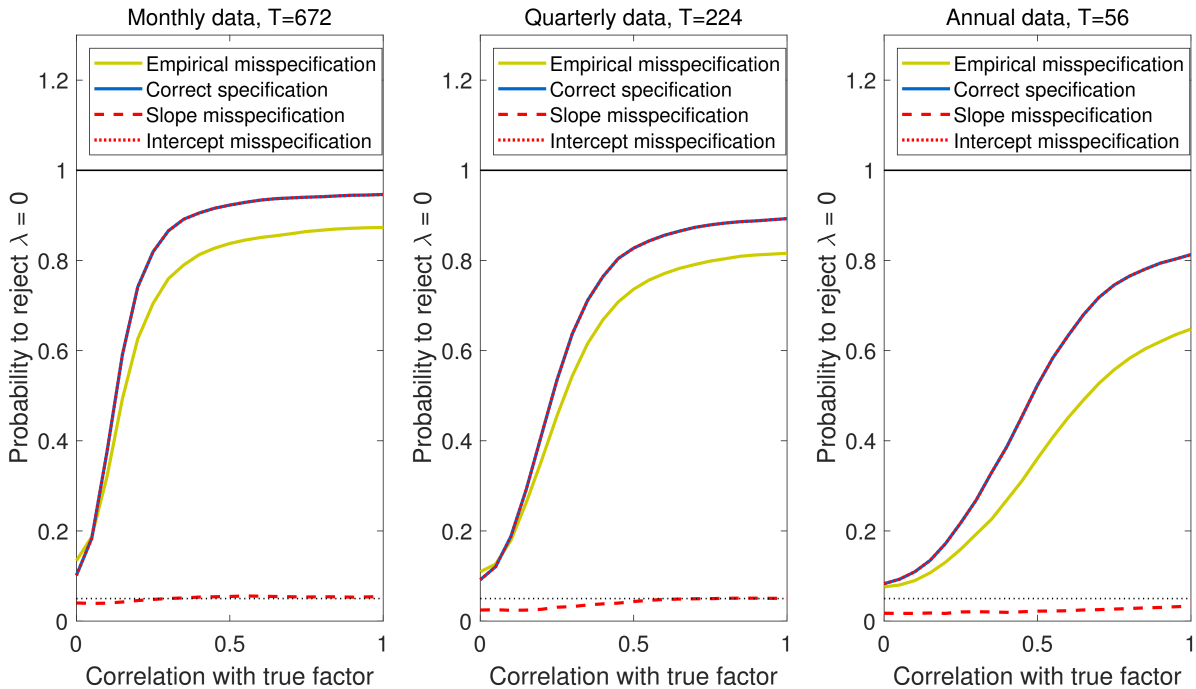
in population. Indeed, several researchers have concluded that factor models are expected to be imperfect (e.g. Fama and French, 2015; Hou et al., 2015).

However, the misspecification robust standard errors are still derived under the assumption that the risk factor is useful (i.e., correlated with the test assets). Thus, they are not necessarily robust to the useless factor problem and their respective small sample performance is unknown to the best of our knowledge.[17]

Figure 7 sheds light on this open question and provides the probability to reject the

[17] Kan et al. (2013) provide analytical results (i.e., asymptotic, or large sample properties) and an empirical application.

Figure 6: Power Curves: Fama-MacBeth, HAC-robust with Shanken Correction, with Estimation of the Intercept



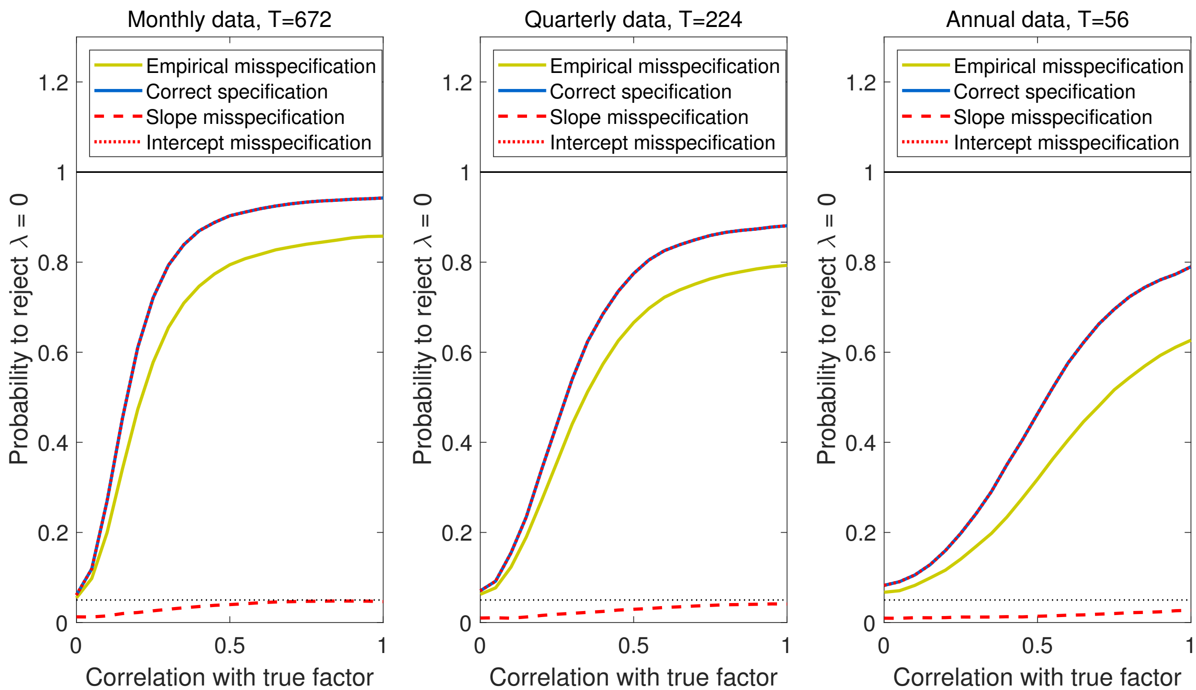
Method: This figure presents the probability to reject the null hypothesis $\lambda = 0$ (5% significance level) of the cross-sectional regression model $\bar{R}_i^e = \lambda_0 + \hat{\beta}_i \lambda + a_i$. The probability to reject is based on 10,000 pairwise bootstrap re-samples.

- Standard errors for λ account for that the betas are estimated (Shanken (1992)-corrected) and are HAC-robust.
- The intercept (a common pricing error) λ_0 is estimated.

Test assets and risk factors: The risk factors and the test assets are the same as described in the caption to Figure 2.

H0: $\lambda = 0$ based on the t-test using the standard errors suggested by Kan et al. (2013). We include an intercept in the cross-sectional regression so that the results are best compared to those in Figure 4. In the case of slope misspecification, i.e. the factor exposures are orthogonal to the expected test asset returns, the probability of rejecting the hypothesis $\lambda = 0$ is always close to 5%, so the test has the correct size for useful but unpriced factors. Moreover, even though the method by Kan et al. (2013) do not explicitly account for the useless factor problem, they seem to reduce the problem: The rejection frequencies with $\rho = 0$ are now lower compared to Figure 4. Based on their empirical analysis, Kan et al. (2013) conclude that the misspecification adjustment tends to be large for risk factors with low correlation

Figure 7: Power Curves: Misspecification Robust Standard Errors, with Estimation of the Intercept



Method: This figure presents the probability to reject the null hypothesis $\lambda = 0$ (5% significance level) of the cross-sectional regression model $\bar{R}_i^e = \lambda_0 + \hat{\beta}_i \lambda + a_i$. The probability to reject is based on 10,000 pairwise bootstrap re-samples.

- Standard errors for λ account for that the betas are estimated (Shanken (1992)-corrected) and are misspecification robust as proposed by Kan et al. (2013).
- The intercept (a common pricing error) λ_0 is estimated.

Test assets and risk factors: The risk factors and the test assets are the same as described in the caption to Figure 2.

to the test assets. Therefore, the correction is likely to be large for "useless" factors. The adjustment is negligible for factors with a high correlation to the test assets.

We conclude that the Kan et al. (2013) standard errors do not resolve, but they mitigate the useless factor problem in settings where the intercept is estimated. Importantly, they come with good power to detect useful factors and strike a good balance between type I and type II error. Unfortunately, Kan et al. (2013) do not provide misspecification robust standard errors when the intercept is imposed to be zero, and we cannot provide the respective results.

3.5 Misspecification and Useless Factor Robust Standard Errors for SDF loadings

Gospodinov et al. (2014) derive standard errors that are misspecification *and* useless factor robust. Their approach is based on the SDF-representation, as outlined in Section 2. Accordingly, we test the H0: $b = 0$.[18] Moreover, the Gospodinov et al. (2014) estimator is set up such that the moments are weighted by the inverse of the sample covariance matrix of the test asset returns.

Figure 8 shows that the Gospodinov et al. (2014)-approach does not over-reject the H0 in the monthly and quarterly data in case the factors are useless ($\rho = 0$). This approach is powerful in detecting useful factors in relatively large samples. However, in small samples, we tend to over-reject the H0 for useless factors. Intuitively, in small samples, the inverse of the covariance matrix of the test assets tends to be estimated imprecisely, which renders inference with this weighting matrix unreliable (see, e.g., the discussion in Cochrane (2005)).

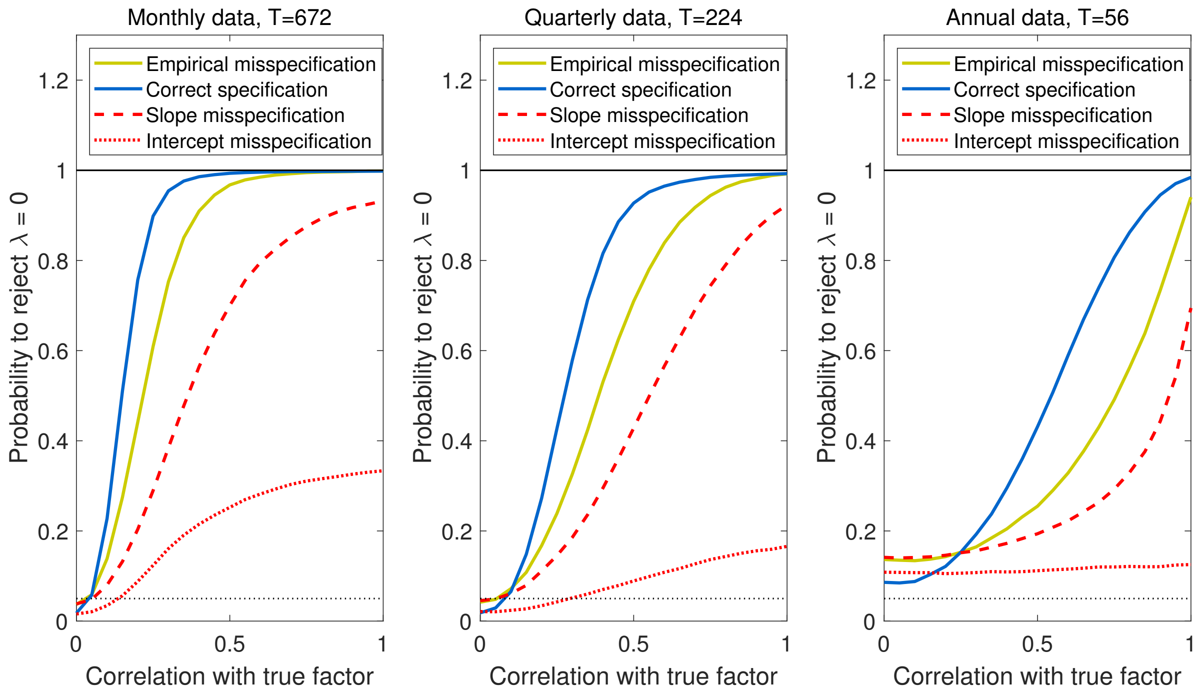
Importantly, the method heavily overrejects the H0 in the case of slope misspecification. Since expected returns are orthogonal to the covariances of returns with the factor in population, one would expect to find a constant pricing kernel ($b = 0$) and a common pricing error close to the mean expected return. In small samples, however, random sample correlations often seem to lead to a rejection of H0. At the same time, the method has very little power to reject the H0 in the case of intercept misspecification. Note that in this case, there actually is a value b such that $E(R_i^e [1 - (F_t - \mu_f)b])$ is constant across test assets in population, so the H0 should be rejected.

In Figure 9, we show results using the classic textbook formulas that are neither misspecification *nor* useless factor robust. Serious over-rejection is eminent and the benefit of applying the Gospodinov et al. (2014) standard errors is clearly given.

Against this backdrop, it is important to notice that there is a priori no economic reason to chose the inverse of the covariance matrix of the test assets as the weighting matrix.

[18] For single factor models, H0: $b = 0$ and H0: $\lambda = 0$ are perfectly equivalent. For multi-factor models, the SDF-loadings (b) allow to test whether a risk-factor has marginal explanatory power given the other risk factors, while the price of risk (λ) allows to test whether its factor-mimicking portfolio earns a significant risk premium (Cochrane, 2005).

Figure 8: Power Curves: SDF-loadings, Misspecification Robust Standard Errors, with Estimation of the Intercept



Method: This figure presents the probability to reject the null hypothesis $b = 0$ (5% significance level). GMM estimates of the SDF-loading b are based on the moment restriction $E(R_i^e\,[1 - (F_t - \mu_f)b] - b_0) = 0$, where R_i^e is the excess return of a test asset, and F_t is the risk factor. The probability to reject is based on 10,000 pairwise bootstrap re-samples.

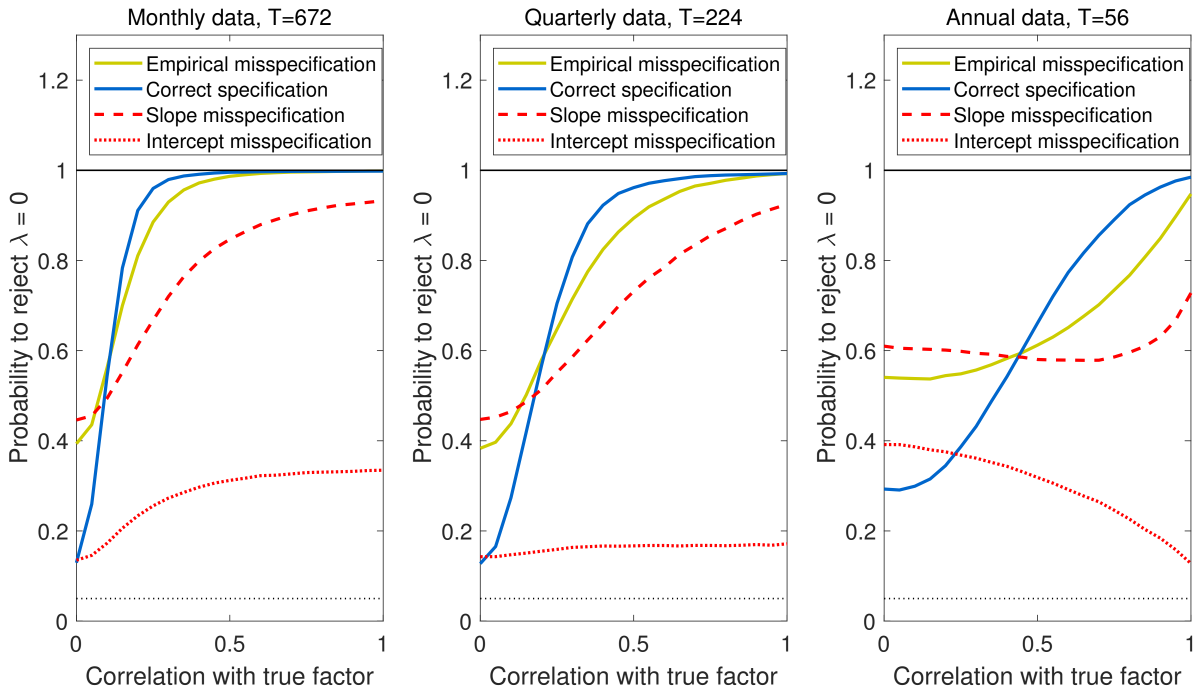
- The moments are weighted by the inverse of the covariance matrix of the asset returns.
- Standard errors are misspecification and useless factor robust as proposed by Gospodinov et al. (2014).
- The intercept (a common pricing error) b_0 is estimated.

Test assets and risk factors: The risk factors and the test assets are the same as described in the caption to Figure 2.

Intuitively, this choice implies that (combinations of) test assets with low volatility get a higher weight in the estimation, in the sense that special attention is devoted to minimizing their pricing errors. If a researcher beliefs that the same priority should be assigned to each of the chosen test assets, the identity weighting matrix is the appropriate choice, independent of any statistical arguments that might apply.

Figure 10 reports the results using the classic GMM formulas (Cochrane, 2005; Burnside, 2011), using the identity matrix for weighting and including a common pricing error. We find power curves that are very similar to those for Fama-MacBeth/Shanken estimates of the

Figure 9: Power Curves: SDF-loadings, classic GMM with $W = S^{-1}$, with Estimation of the Intercept



Method: This figure presents the probability to reject the null hypothesis $b = 0$ (5% significance level). GMM estimates of the SDF-loading b are based on the moment restriction $E(R_i^e\left[1-(F_t-\mu_f)b\right]-b_0)=0$. The probability to reject is based on 10,000 pairwise bootstrap re-samples.

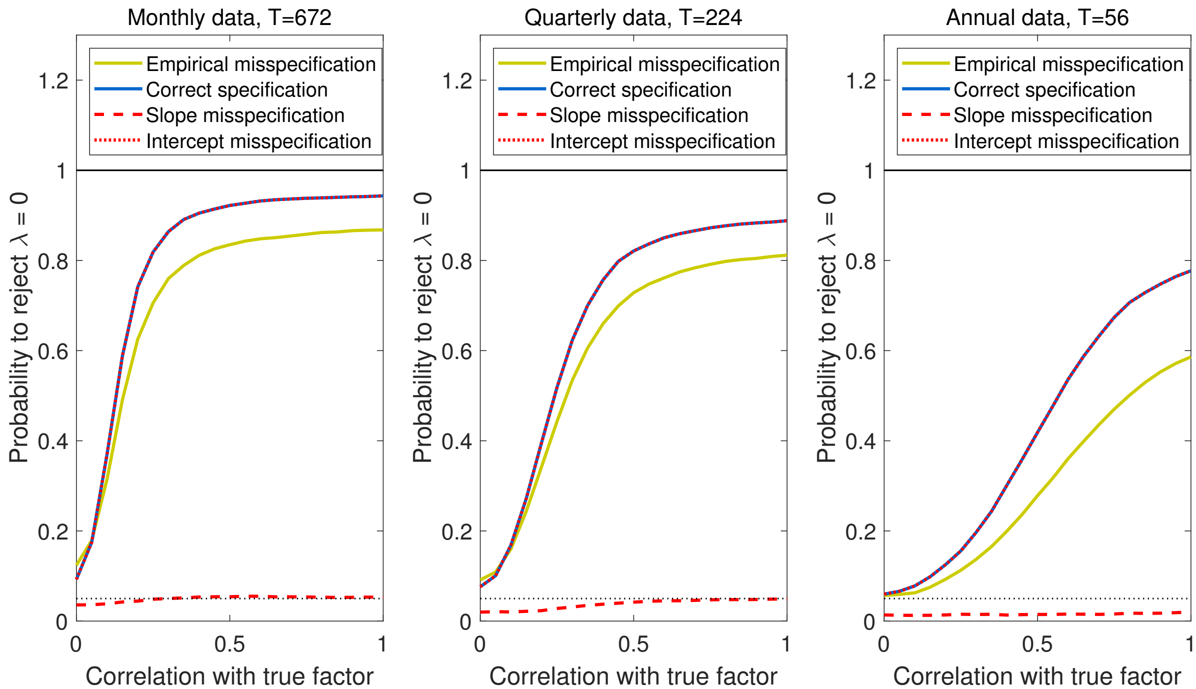
- The moments are weighted by the inverse of the covariance matrix of the asset returns.
- Standard errors are based on the classic GMM formulas and assume that the model is correctly specified (Cochrane (2005), Burnside (2011)).
- The intercept (a common pricing error) b_0 is estimated.

Test assets and risk factors: The risk factors and the test assets are the same as described in the caption to Figure 2.

Beta-representation (Figure 6). Again, in the case of useless factors ($\rho = 0$), the test rejects the H0 too often for the monthly and quarterly data, but not for the annual data. Moreover, as before, the test has reasonable power which increases in sample size, but there is moderate overrejection when the factor is useful but slope misspecified (at least with monthly data).

We also consider the case where the common pricing error b_0 is constraint to be zero, as implied by standard asset pricing theories. This specification has been suggested by Cochrane (2005). The previous results in the Fama-MacBeth/Shanken setting suggest that constraining the common pricing error to zero could improve the properties of the estimator. Figure 11

Figure 10: Power Curves: SDF-loadings, classic GMM with $W = I$, with Estimation of the Intercept



Method: This figure presents the probability to reject the null hypothesis $b = 0$ (5% significance level). GMM estimates of the SDF-loading b are based on the moment restriction $E(R_i^e[1 - (F_t - \mu_f)b] - b_0) = 0$. The probability to reject is based on 10,000 pairwise bootstrap re-samples.

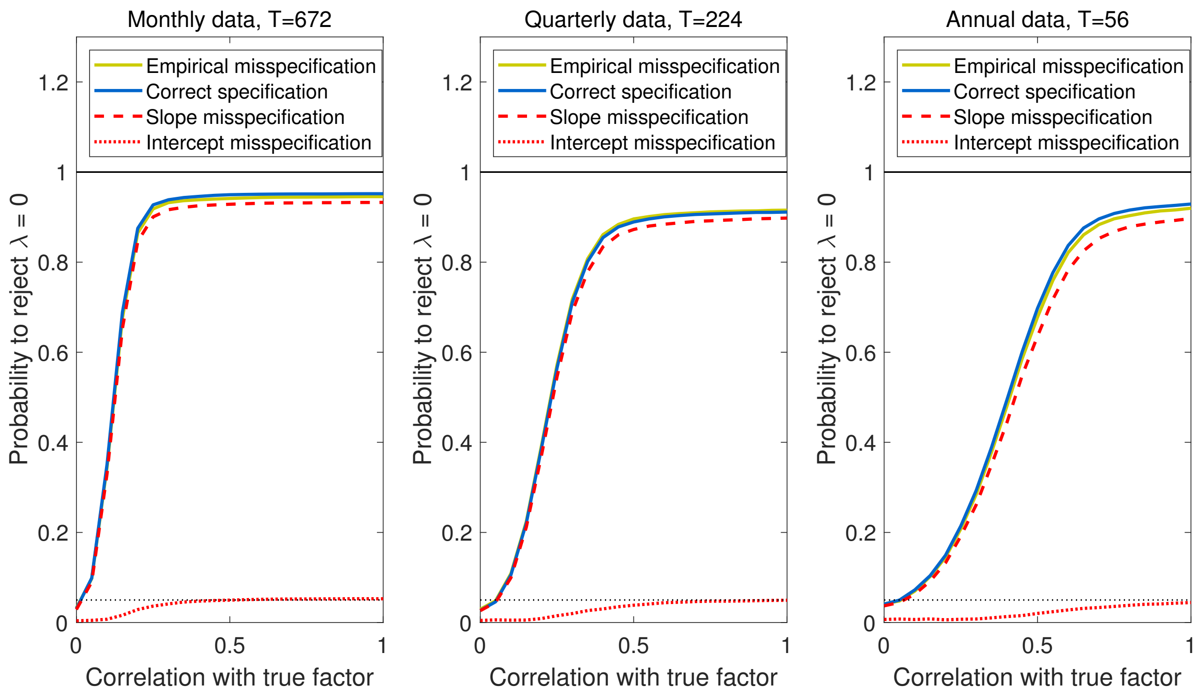
- All moments are equally weighted.
- Standard errors are based on the classic GMM formulas and assume that the model is correctly specified (Cochrane (2005), Burnside (2011)).
- The intercept (a common pricing error) b_0 is estimated.

Test assets and risk factors: The risk factors and the test assets are the same as described in the caption to Figure 2.

shows the results for this specification. We again find power curves that are very similar to those of the Fama-MacBeth/Shanken specification (Figure 5). The classic GMM tests reject H0 with a probability of 5% in case the factors are useless ($\rho = 0$) and are, thus, size correct. Importantly, the classic formulas also come with a steep power curve and allow to distinguish between priced factor and unpriced ones (by intercept misspecification).

To sum up, we confirm that the Gospodinov et al. (2014)-standard errors are important when the pricing equations are weighted by the inverse of the return covariance matrix. However, what is less well acknowledged in the literature so far is the fact that traditional

Figure 11: Power Curves: SDF-loadings, classic GMM with $W = I$, Imposing a Zero Intercept



Method: This figure presents the probability to reject the null hypothesis $b = 0$ (5% significance level). GMM estimates of the SDF-loading b are based on the moment restriction $E(R_i^e\,[1 - (F_t - \mu_f)b] - 0) = 0$. The probability to reject is based on 10,000 pairwise bootstrap re-samples.

- All moments are equally weighted.
- Standard errors are based on the classic GMM formulas and assume that the model is correctly specified (Cochrane (2005), Burnside (2011)).
- The intercept (a common pricing error) b_0 is imposed to be zero.

Test assets and risk factors: The risk factors and the test assets are the same as described in the caption to Figure 2.

inference on SDF-loadings is actually robust to the useless factor problem in empirically relevant sample sizes if the moment restrictions are equally weighted, and when the common pricing error is imposed to its theoretical value of zero. Apart from the useless factor problem, it turns out that the test using the unit weighting matrix is superior when it comes to inference about useful factors.

3.6 Using Pricing Errors to Conduct Inference on the Price of Risk

Kleibergen and Zhan (2020) are worried about the useless factor problem when inference is based on the Fama-MacBeth/Shanken approach. They propose the so-called GRS-FAR test[19] to conduct inference on the price of risk (λ) based on the pricing errors (α_i) from time-series regressions. The GRS-FAR follows the following three steps:

1. Run a time-series regression for each of the N test assets:

$$R_{it} = \alpha_i + \beta_i \left(\tilde{F}_t + \lambda_{f,0} \right) + \varepsilon_{it}, \quad \text{for } \ i = 1, ..., N \text{ and } \ t = 1, ..., T$$

   Here $\tilde{F}_t = F_t - \hat{\mu}_f$ denotes the de-meaned factor with variance $\hat{\sigma}_f^2$ and $\lambda_{f,0}$ is the hypothesized price of risk.

2. Compute the well known Gibbons et al. (1989)-test on the joint significance of Jensen's alphas α_i:

$$GRS = \frac{T - N - 1}{N} \left(1 + \left(\frac{\hat{\mu}_f}{\hat{\sigma}_f^2} \right)^2 \right)^{-1} \hat{\boldsymbol{\alpha}}' \Sigma_{\varepsilon}^{-1} \hat{\boldsymbol{\alpha}} \sim F_{N, T-N-1},$$

   where $\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_1, ...\hat{\alpha}_N)'$ is a vector collecting the N alphas, and Σ_{ε} is the covariance matrix of the N time-series regression residuals. Write down the p-value of the GRS-test under the assumption that the data are normal-i.i.d.

3. Repeat step 1 and 2 for a wide array of values $\lambda_{f,0}$ for the candidate market price of risk. The GRS-FAR $100 \times (1 - \alpha)\%$ confidence set is the region of $\lambda_{f,0}$, for which the GRS test does not reject at the $\alpha\%$ significance level.

This test has four possible outcomes, characterized by the structure of the confidence sets:

[19] GRS-FAR is an abbreviation of Gibbons-Ross-Shanken-Factor Anderson-Rubin, referring to Gibbons et al. (1989) and Anderson et al. (1949).

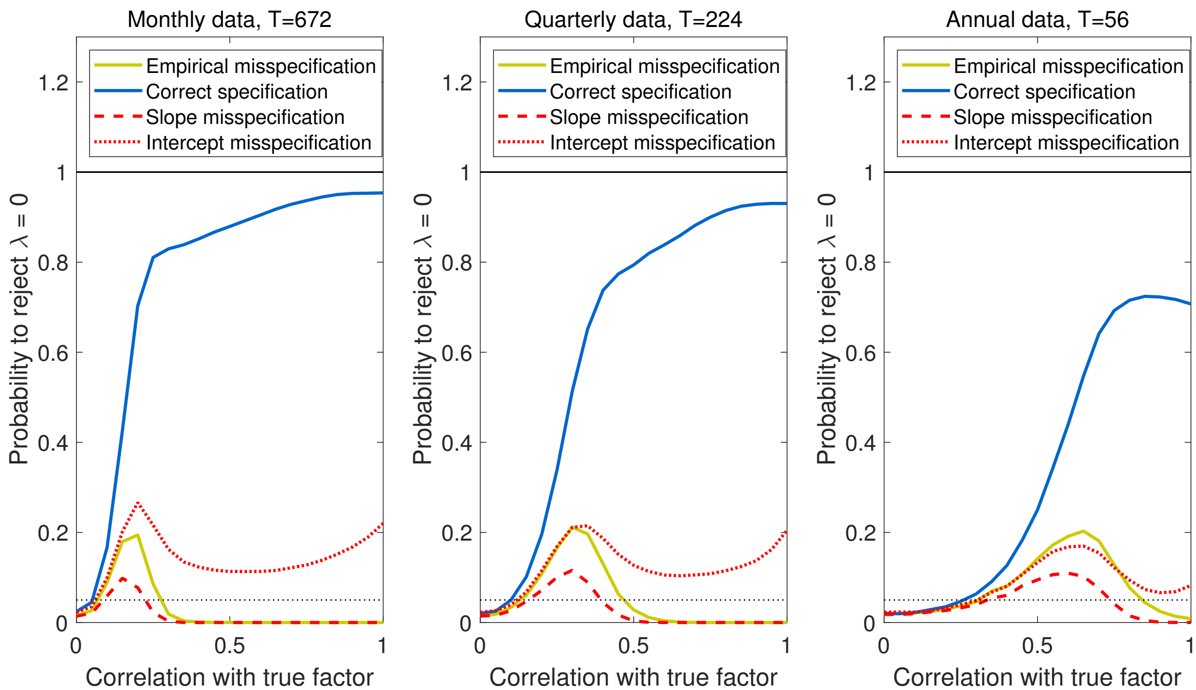
- The confidence set is bounded from two-sides, indicating that the model is not rejected for all values $\lambda_{f,0}$ that are greater than the lower and smaller than the upper bound. At the same time, the model is rejected for all $\lambda_{f,0}$ that are outside of that interval.
- The confidence set is dis-joint, indicating that the confidence set is bounded from one side but unbounded from the other side.
- The confidence set is unbounded, indicating that the model is never rejected.
- The confidence set is empty, indicating that the model is rejected for any $\lambda_{f,0}$.

This method is by construction robust to the useless factor problem, because the price of risk $\lambda_{f,0}$ is imposed and not estimated. However, only in the first case of a bounded confidence set can a researcher conduct inference on the price of risk. In this case, the test rejects H0: $\lambda = 0$ whenever the interval does not contain 0. In all other cases, the GRS-FAR test is inconclusive. As pointed out by Kroencke (2020), a shortcoming of this method is that it requires the model to be correctly specified, i.e. all pricing errors need to be equal to zero in population. Even in case of only slight model misspecification, the GRS test should reject for the true value λ and also for all other possible values $\lambda_{f,0}$. Thus, the GRS-FAR confidence set is expected to be empty so that inference on the price of risk is not possible.

The assumption of a correctly specified asset pricing model is in contrast to a large literature in empirical asset pricing (e.g. Fama and French, 2015; Hou et al., 2015), arguing that all asset pricing models should be regarded as imperfect (misspecified). It is also incompatible with the other "robust" tests discussed earlier, which stress the importance of regarding asset pricing models as misspecified when conducting inference (Kan and Zhang, 1999; Kan et al., 2013; Gospodinov et al., 2014).

Figures 10 and 11 report the probability to reject the H0: $\lambda = 0$ when applying the GRS-FAR test as proposed by Kleibergen and Zhan (2020). We confirm their finding that the GRS-FAR test is robust to the useless factor problem. However, the GRS-FAR test can only detect useful factors if the model is correctly specified, as indicated by the blue line, and when it is intercept misspecified and an intercept is estimated (red dotted line in Figure 10). In these cases, the population pricing errors are exactly equal to zero (or to a

Figure 12: Power Curves: GRS-FAR statistic, Imposing a Zero Intercept



Method: This figure presents the probability to reject the null hypothesis $\lambda = 0$ (5% significance level) of the cross-sectional model $E(R_i^e) = 0 + \lambda\beta_i$. Inference is based on the GRS-FAR test proposed by Kleibergen and Zhan (2020). The probability to reject is based on 10,000 pairwise bootstrap re-samples.

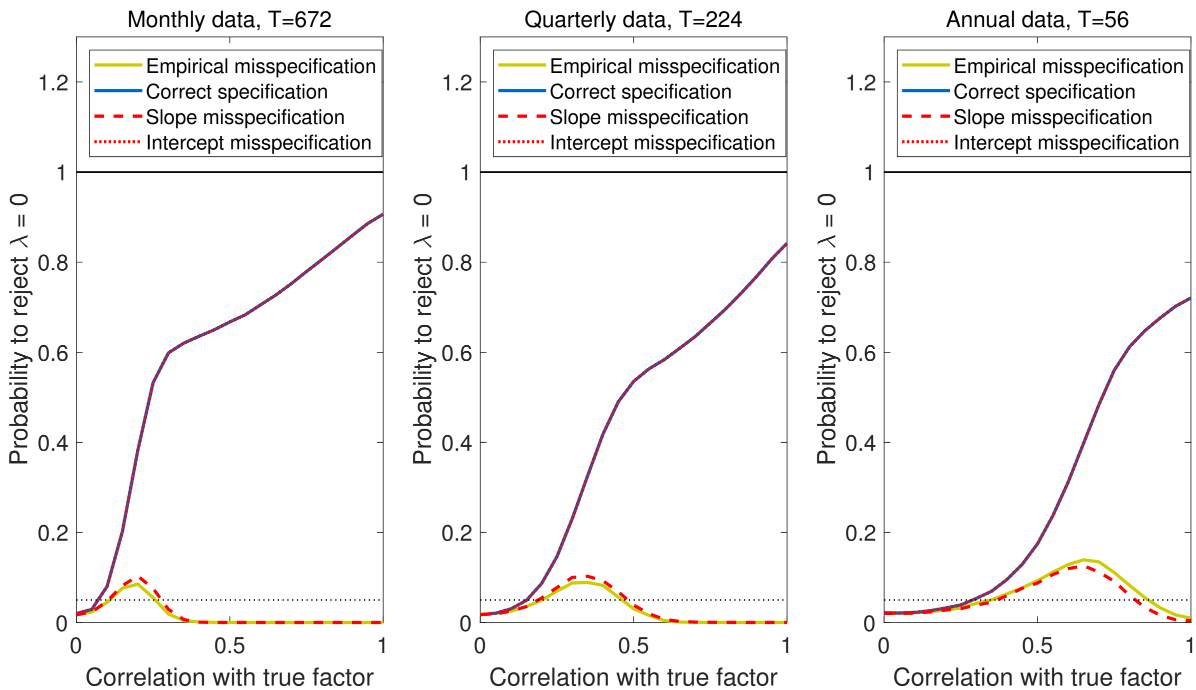
- The method assumes that the asset pricing model is correctly specified (as is the case for the blue line, but not for the red and yellow lines).
- The intercept (a common pricing error) is imposed to be zero.

Test assets and risk factors: The risk factors and the test assets are the same as described in the caption to Figure 2.

constant). As soon as there is mild mispricing in population (yellow lines), the power curve of the GRS-FAR test is virtually flat. An inconclusive outcome of the test is, thus, almost pre-defined.

We conclude that the GRS-FAR test is not useful in empirical research when the model is likely to be even slightly misspecified as it lacks power. The GRS-FAR test should only be used if a researcher hypothesizes that the model is correctly specified. In this case, the test still has less power compared to other size correct approaches discussed earlier.

Figure 13: Power Curves: GRS-FAR statistic, with Estimation of the Intercept



Method: This figure presents the probability to reject the null hypothesis $\lambda = 0$ (5% significance level) of the cross-sectional model $E(R_i^e) = \lambda_0 + \lambda\beta_i$. Inference is based on the GRS-FAR test proposed by Kleibergen and Zhan (2020). The probability to reject is based on 10,000 pairwise bootstrap re-samples.

- The method assumes that the asset pricing model is correctly specified (as is the case for the blue line, but not for the red and yellow lines).
- The intercept (a common pricing error), λ_0, is assumed to be equal to the Fama-MacBeth estimate.

Test assets and risk factors: The risk factors and the test assets are the same as described in the caption to Figure 2.

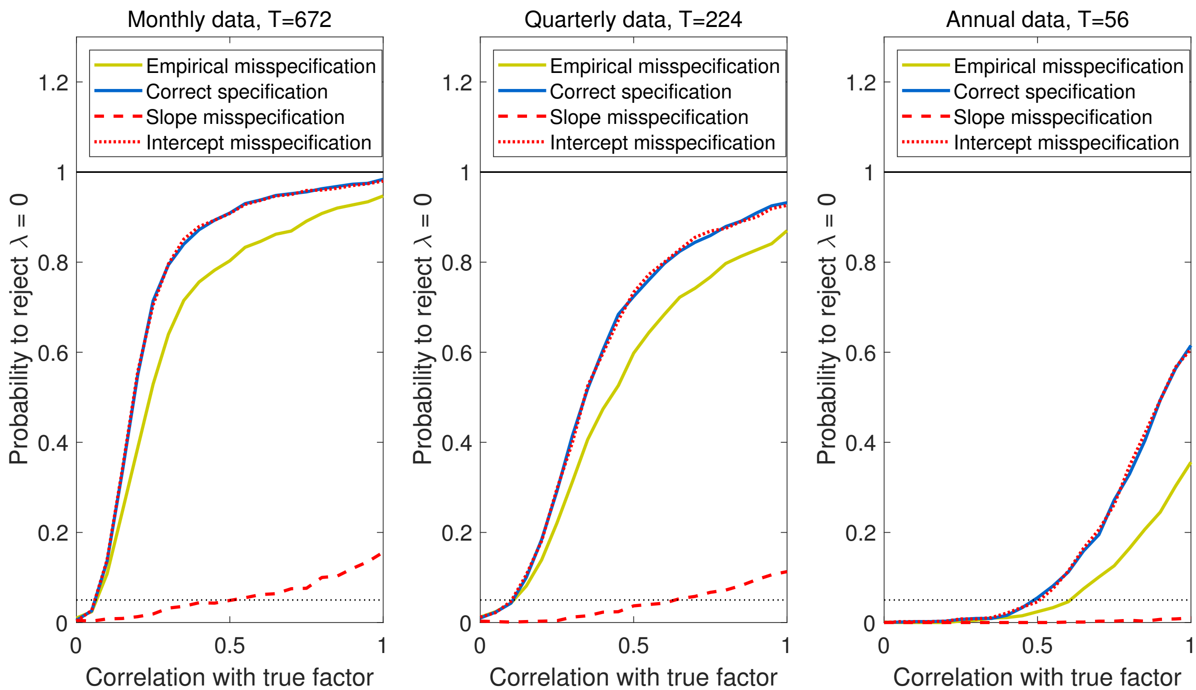
3.7 Bayesian Fama-MacBeth

Bryzgalova et al. (2020) propose Bayesian methods for cross-sectional regressions and show that this approach is robust against useless factors. Their estimator is based on a simulation-based posterior distribution where a non-informative prior on the data is imposed. In practical terms, their prior imposes that mean returns and betas are uncorrelated.[20]

Figure 14 reports the probability to reject the H0: $\lambda = 0$ using the Bayesian Fama-MacBeth approach. Estimation is with an intercept. Our results are in line with Bryzgalova

[20]Importantly, their procedure can not only be used for inference about λ of single candidate models, but it is particularly suited for model comparison. We do not discuss this aspect here.

Figure 14: Power Curves: Bayesian Fama-MacBeth, with Estimation of the Intercept



Method: This figure presents the probability to reject the null hypothesis $\lambda = 0$ (5% significance level) of the cross-sectional regression model $\bar{R}_i^e = \lambda_0 + \hat{\boldsymbol{\beta}}_i'\lambda + a_i$. The probability to reject is based on 1,000 pairwise bootstrap re-samples.

- Inference is based on Bayesian Fama-MacBeth as proposed by Bryzgalova et al. (2020). We use 500 Monte Carlo simulations to obtain Bayesian confidence intervals.
- The intercept (a common pricing error) λ_0 is estimated.

Test assets and risk factors: The risk factors and the test assets are the same as described in the caption to Figure 2.

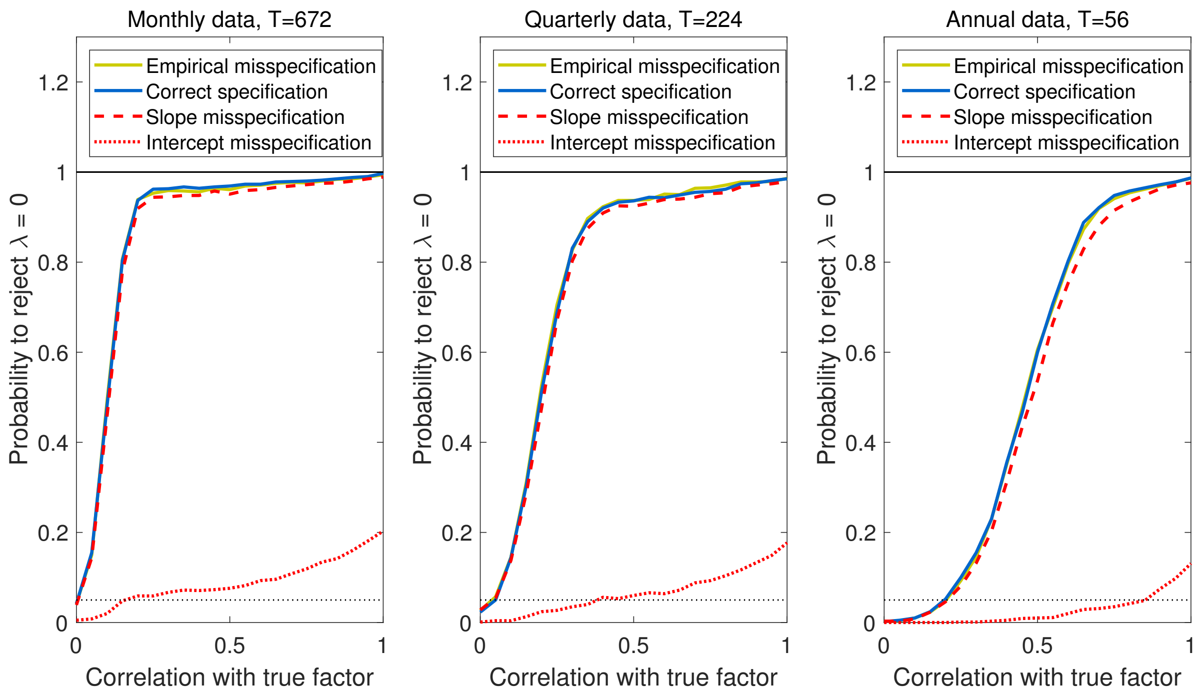
et al. (2020) in the sense that the estimator is robust to the useless factor problem. Importantly, the power is high for useful factors (correlation with the true factor of at least 50%), except when the sample size is small (T=65, annual data).[21]

Our previous results led us to the conjecture that imposing a zero intercept increases power. For that reason, we extend the results provided by Bryzgalova et al. (2020),[22] and show in Figure 15 the according rejection probabilities when the intercept is imposed to be zero. We find that the power is considerably increased and very similar to the frequen-

[21] However, we find that the estimator tends to over-reject useful but unpriced factors (slope misspecification) when the sample size is larger (T=672, monthly data).

[22] As of the writing of this paper.

Figure 15: Power Curves: Bayesian Fama-MacBeth, Imposing a Zero Intercept



Method: This figure presents the probability to reject the null hypothesis $\lambda = 0$ (5% significance level) of the cross-sectional regression model $\bar{R}_i^e = 0 + \hat{\boldsymbol{\beta}}_i' \lambda + a_i$. The probability to reject is based on 1,000 pairwise bootstrap re-samples.

- Inference is based on Bayesian Fama-MacBeth as proposed by Bryzgalova et al. (2020). We use 500 Monte Carlo simulations to obtain Bayesian confidence intervals.
- The intercept (a common pricing error) is imposed to be zero.

Test assets and risk factors: The risk factors and the test assets are the same as described in the caption to Figure 2.

tist inference based on Fama-MacBeth/Shanken regressions. In both test designs (with and without intercept), we find moderate overrejection of useful but unpriced factors (slope misspecification in Figure 14 and intercept misspecification in Figure 15), at least in the monthly sample.

To sum up, the Bayesian approach is a reliable and powerful alternative to the traditional Fama-MacBeth/Shanken method, except for the case where the sample size is small and the intercept is estimated. This happens to be the case where the useless factor problem has the least relevance for traditional inference.

3.8 The Three-Pass Method

Giglio and Xiu (2021) propose a three-pass method to estimate the price of risk that is misspecification and useless factor robust. The method can be described as follows:

1. Extract the first p principal components (PCs) of the test asset returns.

2. Run cross-sectional regressions using the PCs as factors (instead of the candidate factors) to find the risk premia of the PCs.

3. Use time-series regressions of the risk factor (F_t) on the PCs and use the risk premia estimates of the PCs from step 2 to find risk premia estimates of the risk factor.

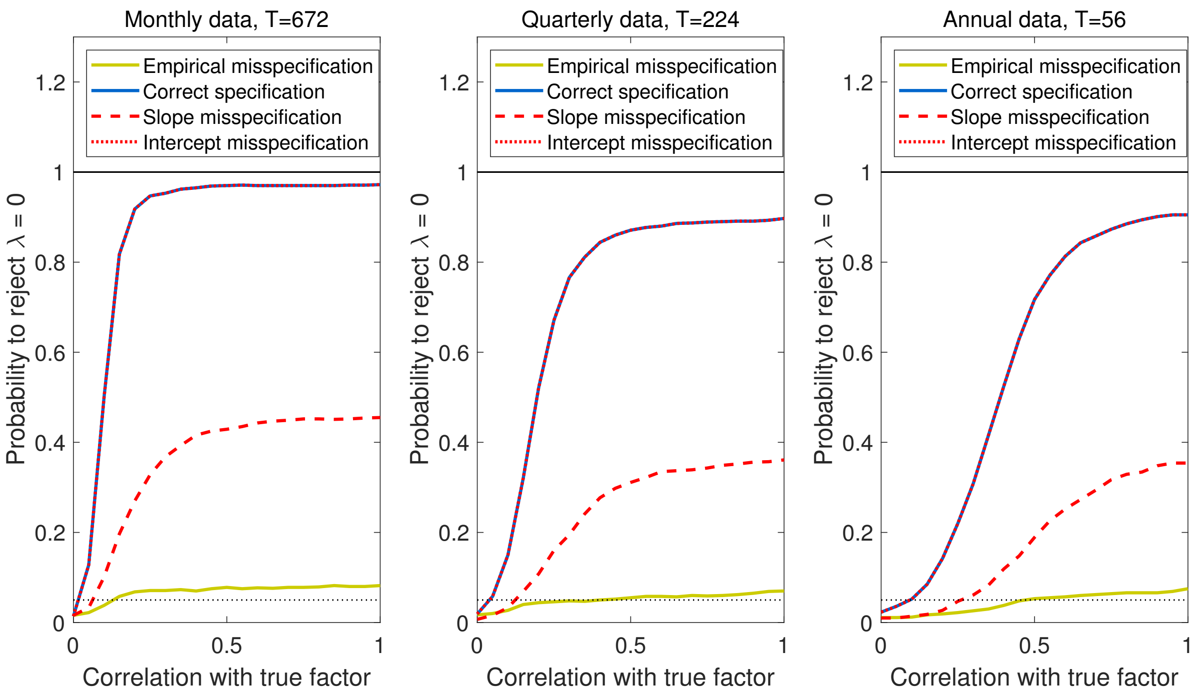
This estimator can be viewed as a regulated version of the factor mimicking portfolio approach, or, alternatively, as a cross-sectional regression that adds PCs as control factors. By that step, the method allows to incorporate omitted factors in the model. Giglio and Xiu (2021) point out that for this idea to work well, the set of test assets should be large enough such that all priced risks in the economy are spanned by the test asset returns. Accordingly, they choose a rather large set of test assets (including stock, bond, and currency portfolios) in their empirical analysis. We complement their analysis by stress testing the three-pass method when T and N are small, as it is often the case in the literature.

Figure 16 first of all confirms that the three-pass method is robust to the useless factor problem in small samples. However, power to detect useful factors is only reasonably high when the model is correctly specified in population (blue lines) or intercept misspecified (dotted red lines). For the empirically misspecified model (yellow lines), we find that the rejection rates are very low. Thus, when applied in settings with a small number of test assets the test lacks power to reject the H0 in case of mild model misspecification.

To conclude, we point out that the three-pass method should not be applied in situations where the number of test assets is small, as it lacks power when the asset pricing model is misspecified. We believe that our conclusion is in line with Giglio and Xiu (2021)'s recommendation to apply the three-pass method to large sets of test assets.[23]

[23] Nevertheless, it is interesting to see that the three-pass method works well for correctly specified models even in the usual cross-section of stock returns.

Figure 16: Power Curves: Three-Pass Method, with Estimation of the Intercept



Method: This figure presents the probability to reject the null hypothesis $\lambda = 0$ (5% significance level) of the cross-sectional regression model $E(R_i^e) = \lambda_0 + \lambda\beta_i + a_i$. The probability to reject is based on 10,000 pairwise bootstrap re-samples.

- Inference is based on the three-pass methods proposed by Giglio and Xiu (2021).
- The intercept (a common pricing error), λ_0, is estimated.

Test assets and risk factors: The risk factors and the test assets are the same as described in the caption to Figure 2.

3.9 The Pairwise-Bootstrap Approach

Finally, we study a computationally convenient alternative that is robust to misspecification and the useless factor problem. The pairwise bootstrap for inference on the price of risk, discussed by Cochrane (2005) and Burnside (2011), works as follows:

1. First, from the original data $(F_t, R_{1,t}, ..., R_{N,t})$, sample with replacement a new set of observations $(F_t^b, R_{1,t}^b, ..., R_{N,t}^b)$.

2. Estimate the factor betas for the new sample, $\hat{\beta}_i^{b\prime} = \left(R_{i,t}^{b\prime} R_{i,t}^b\right)^{-1} \left(R_{i,t}^{b\prime} F_t^b\right)$.

3. Compute the price of risk, $\hat{\lambda}^b = \left(\hat{\beta}_i^{b\prime} \hat{\beta}_i^b\right)^{-1} \left(\hat{\beta}_i^{b\prime} R_{i,t}^b\right)$.

4. Repeat steps 1, 2, and 3 N times.

5. Use the $\alpha/2$ and $1-\alpha/2$ percentiles of the resulting bootstrapped distribution of $(\hat{\lambda}^b)_{b=1,\ldots,N}$ to find the $100 \times (1-\alpha)\%$ bootstrapped confidence set of λ.

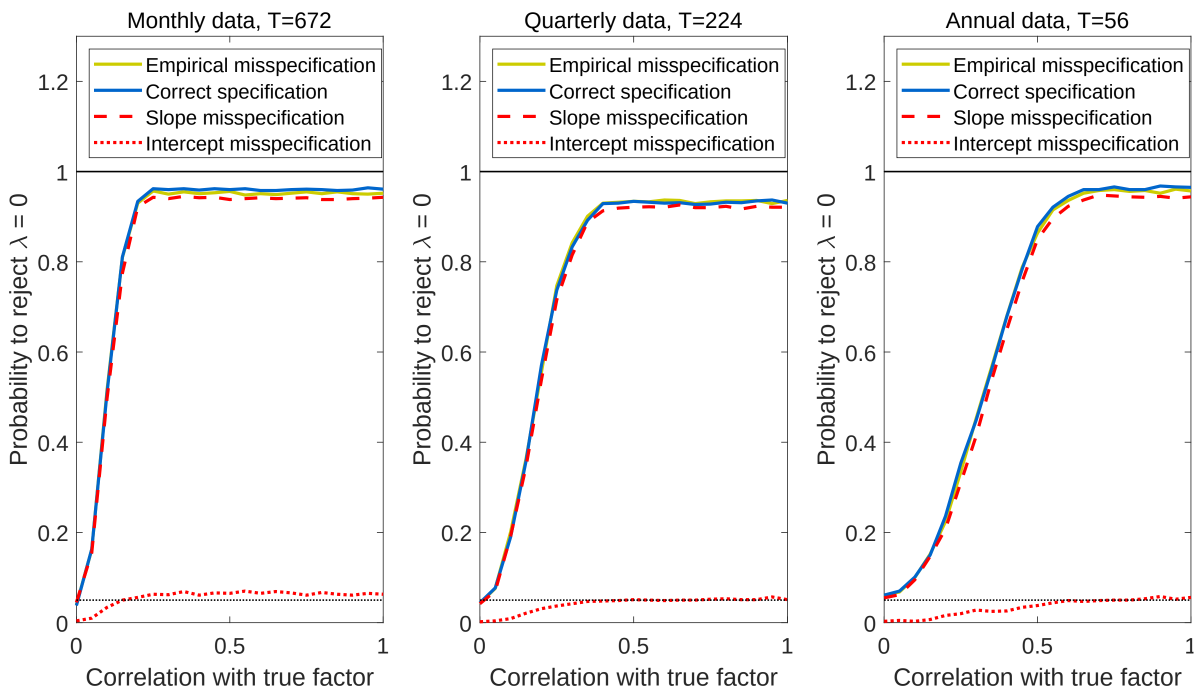
These confidence sets will account for the fact that the model might be misspecified, as well as that betas are estimated. Mild autocorrelation can be accounted for by using a block bootstrap, which should be sufficient for returns on financial assets and most risk factors. Importantly, the bootstrapped confidence intervals are robust against the useless factor problem, in the sense that one does not over-reject the hypothesis H0: $\lambda = 0$.[24]

Figures 17 and 18 illustrate the performance of bootstrap confidence intervals. Useless factors are never over-rejected. Estimation with imposing a zero intercept leads to rejection of intercept misspecified models with the expected probability of 5%. The same applies to slope misspecified models when the intercept is estimated. The power is high even in relatively small samples when risk factors are reasonably accurately measured ($\rho > 0.50$). These are ideal properties. In Section 4, we illustrate that bootstrap confidence intervals are also robust to the useless factor problem in large samples ($T = 6,000$).

We point out an important caveat. In the case of useless factors, the price of risk is not identified by any method (as discussed earlier). This means that the bootstrapped confidence intervals are not informative about the true location of the price of risk, i.e. they lose their usual interpretation (see also, e.g., Burnside, 2011). In practical terms, this means that when the bootstrap confidence interval does include $\lambda = 0$, the correct conclusion is that "the price of risk is not different from zero OR the price of risk is not identified". We believe that the distinction between the two possibilities is not important in practice, because inference on the distribution of an "unpriced" risk factor is of limited interest in research (as far as we are aware). If the difference *is* of importance to the researcher, she can still conduct a test on the significance of the betas/correlations of the risk factors to test for the identification condition.

[24] We prefer bootstrap confidence intervals over bootstrap standard errors because the price of risk estimates are not necessarily normally distributed, especially in the presence of a useless (or weak) factor.

Figure 17: Power Curves: Pairwise Bootstrap, Imposing a Zero Intercept

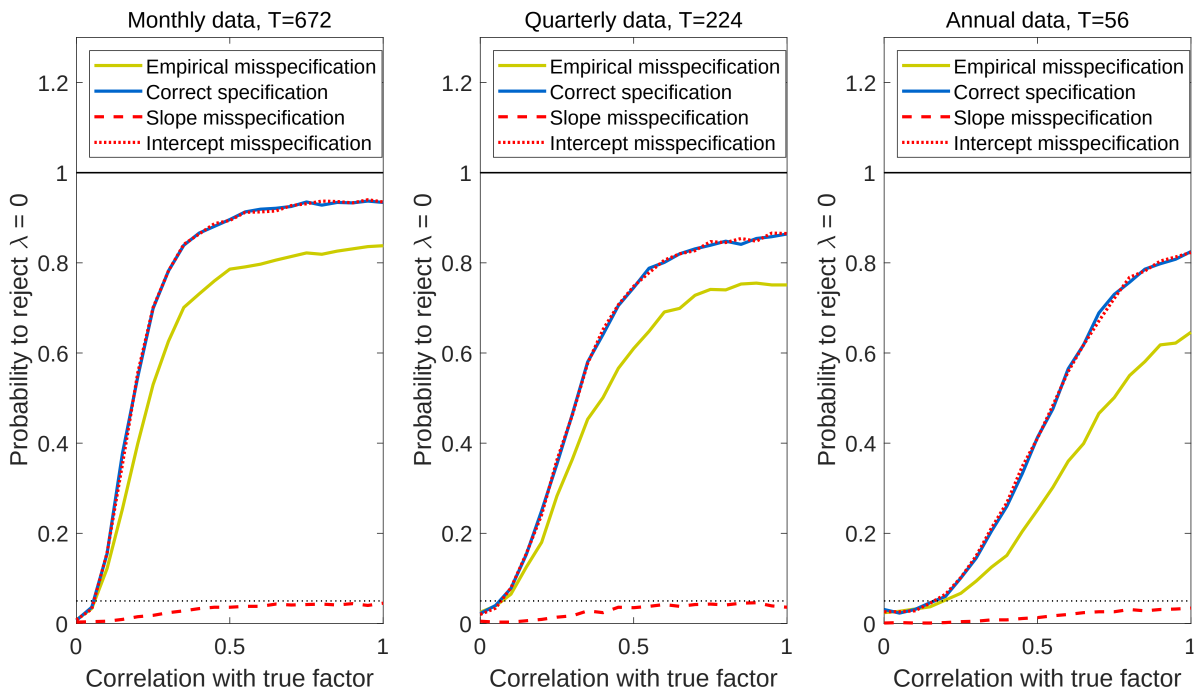


Method: This figure presents the probability to reject the null hypothesis $\lambda = 0$ (5% significance level) of the cross-sectional regression model $E(R_i^e) = \lambda_0 + \lambda\beta_i + a_i$. The probability to reject is based on 1,000 pairwise bootstrap re-samples.

- Inference is based on bootstrapped confidence intervals constructed for each re-sample.
- The intercept (a common pricing error), λ_0, is estimated.

Test assets and risk factors: The risk factors and the test assets are the same as described in the caption to Figure 2.

Figure 18: Power Curves: Pairwise Bootstrap, with Estimation of the Intercept



Method: This figure presents the probability to reject the null hypothesis $\lambda = 0$ (5% significance level) of the cross-sectional regression model $E(R_i^e) = \lambda_0 + \lambda\beta_i + a_i$. The probability to reject is based on 1,000 pairwise bootstrap re-samples.

- Inference is based on bootstrapped confidence intervals constructed for each re-sample.
- The intercept (a common pricing error), λ_0, is estimated.

Test assets and risk factors: The risk factors and the test assets are the same as described in the caption to Figure 2.

4 Large Sample Properties

We repeat parts of our analysis using Monte Carlo simulations. These allow us to increase the sample size and to study the different estimators under ideal conditions, i.e., multivariate normality of the factors and return data. We simulate artificial data as follows:

1. For given set of test assets and a candidate risk factor, collect the monthly data in the $T \times (K + 1)$ matrix $\mathbf{X}_t = [\mathbf{R}_t^{e\prime}, F_t]$.
2. Estimate the sample mean $\hat{\boldsymbol{\mu}}_{\boldsymbol{X}}$ and variance-covariance matrix $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{X}}$.
3. Simulate $m = 1, ..., M$ samples of the data assuming $\mathbf{X}_t^m \sim MVN\left(\hat{\boldsymbol{\mu}}_{\boldsymbol{X}}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\right)$. For each sample m, estimate asymptotic standard errors, apply simulation methods (e.g., Bayesian-FMB), or bootstrap $\mathbf{X}_t^m$, as outlined in Section 3.

Size and Power: Table 2 shows rejection frequencies for the H0: $\lambda = 0$ given $T = 6{,}000$ monthly observations. Overall, the results confirm our conclusions based on small samples. As shown earlier in the literature, the over-rejection problem of the traditional FMB-method is a concern when the intercept is estimated and the sample is large ($T = 6{,}000$). When the intercept is imposed to be equal to zero, an important case not reported by Kan and Zhang (1999) and mainly omitted in the literature so far, we reject a zero price of risk about 9% of the time when the significance level is 5%. The over-rejection problem tends to be of limited relevance in this specification, even in very large samples.

The most accurate results (size and high power) are provided by the pairwise bootstrap confidence intervals. The bootstrap does not over-reject useless factors even in large samples, as the confidence intervals rely on the entire distribution of bootstrapped estimates. As stressed by Burnside (2011), it is important to use bootstrap confidence intervals and not bootstrap standard errors. The latter also rely on distributional assumptions (e.g. normality of the estimates), which are not satisfied for useless factors.

Table 2: Monte Carlo Simulation, T=6,000

This table shows the probability to reject the hypothesis (5% significance level) of a zero price of risk, $\bar{R}_i^e = (\lambda_0 +)\beta_i\lambda + a_i$, H0: $\lambda = 0$, as well as a zero SDF loading, $SDF_t = 1 - (F_t - \mu_f)b$, H0: $b = 0$. The probability to reject is based on 10,000 (1,000 for Bayesian inference and the Bootstrap confidence intervals) Monte Carlo simulations imposing a multivariate normal distribution. The Bayesian inference and the bootstrap confidence intervals are based on 500 resamples for each of the 1,000 simulated samples. Each Monte Carlos sample has T=6,000 time-series observations.

Methods: The methods are the same as described in in Section 3.

Test assets and risk factors: The risk factors and the test assets are the same as described in the caption to Figure 2, monthly data.

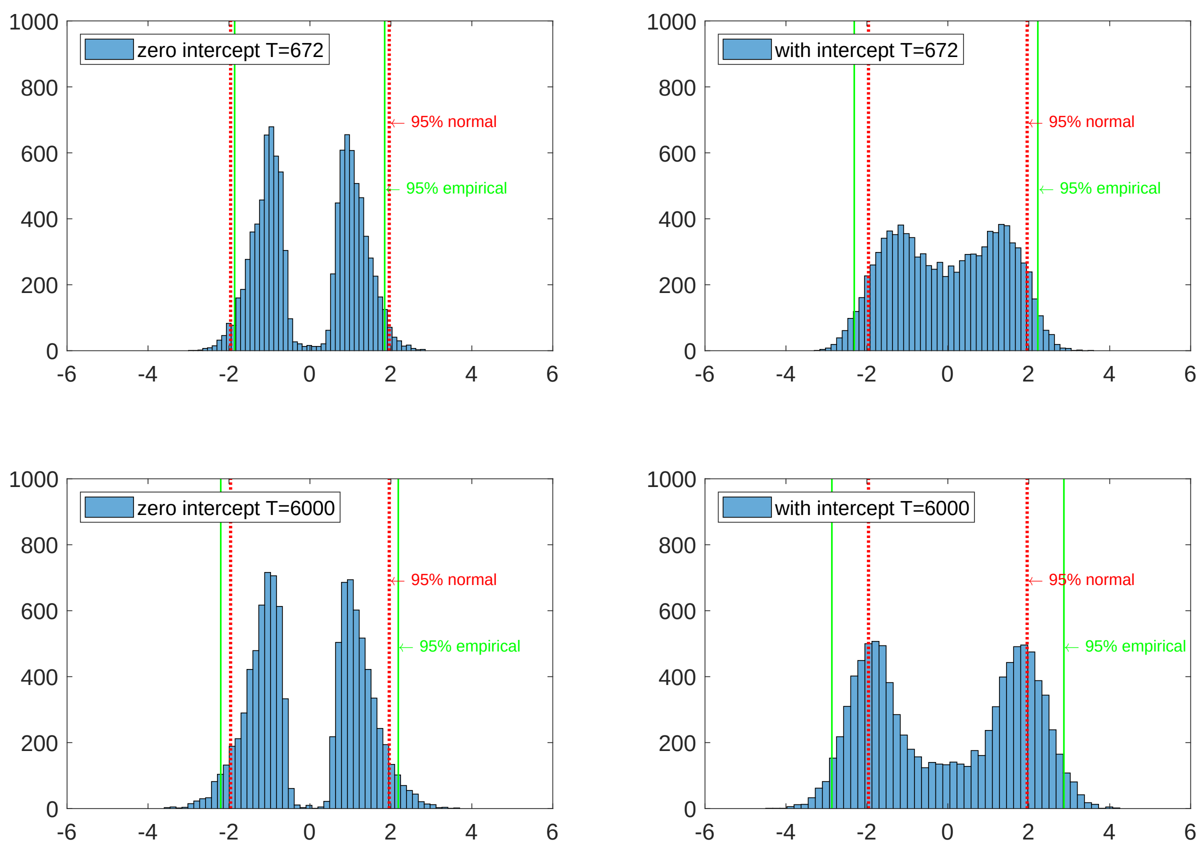
Monte Carlo Simulation, T=6,000					
	useless factor	empirical misspec.	correct spec.	slope misspec.	intercept misspec.
Price of risk: $reject\ H0(5\%) : \lambda = 0?\ zero\ intercept$					
$FMB - OLS$	0.99	1.00	1.00	1.00	0.05
$FMB - Shanken$	0.09	1.00	1.00	1.00	0.05
$Bayesian - FMB$	0.07	1.00	1.00	1.00	0.23
$Robust\ GRS - FAR$	0.01	0.00	0.97	0.00	0.00
$Bootstrap\ CIs$	0.06	1.00	1.00	1.00	0.06
Price of risk: $reject\ H0(5\%) : \lambda = 0?\ with\ intercept$					
$FMB - OLS$	0.73	1.00	1.00	0.05	1.00
$FMB - Shanken$	0.40	1.00	1.00	0.05	1.00
$Robust\ FMB$	0.08	1.00	1.00	0.05	1.00
$Bayesian - FMB$	0.04	1.00	1.00	0.17	1.00
$Three\ pass\ method$	0.03	0.00	1.00	0.94	1.00
$Robust\ GRS - FAR$	0.01	0.00	0.96	0.00	0.97
$Bootstrap\ CIs$	0.05	1.00	1.00	0.06	1.00
SDF loading: $reject\ H0(5\%)\ b = 0?$					
$SDF - W = I,\ zero\ int.$	0.09	1.00	1.00	1.00	0.05
$SDF - W = I,\ with\ int.$	0.40	1.00	1.00	0.05	1.00
$SDF - W = S^{-1},\ with\ int.$	0.76	1.00	1.00	1.00	1.00
$Robust\ SDF - W = S^{-1},\ with\ int.$	0.05	1.00	1.00	1.00	1.00

In the Appendix, we report Monte Carlo Simulation results for sample sizes of $T = 672$ (Table 3), and $T = 120$ (4). These small sample results confirm our conclusions based on the bootstrap analysis provided in Section 3.

Distribution of FMB/Shanken t-statistics: In this section, we provide plots of the classic FMB/Shanken t-statistics from the Monte Carlo simulation. This allows us to illustrate the distributional properties in the presence of a useless and useful factor.

Figure 19 considers the case of a useless factor ($\rho = 0.00$). It is easy to see that i) the distribution of t-statistics is not normal, ii) the usual critical values for testing at the 5%-level (-1.96, 1.96) are too small when estimation is performed with an intercept or when the sample size is large.[25]

Figure 19: Distribution of Fama-MacBeth/Shanken t-statistics: Useless Factor

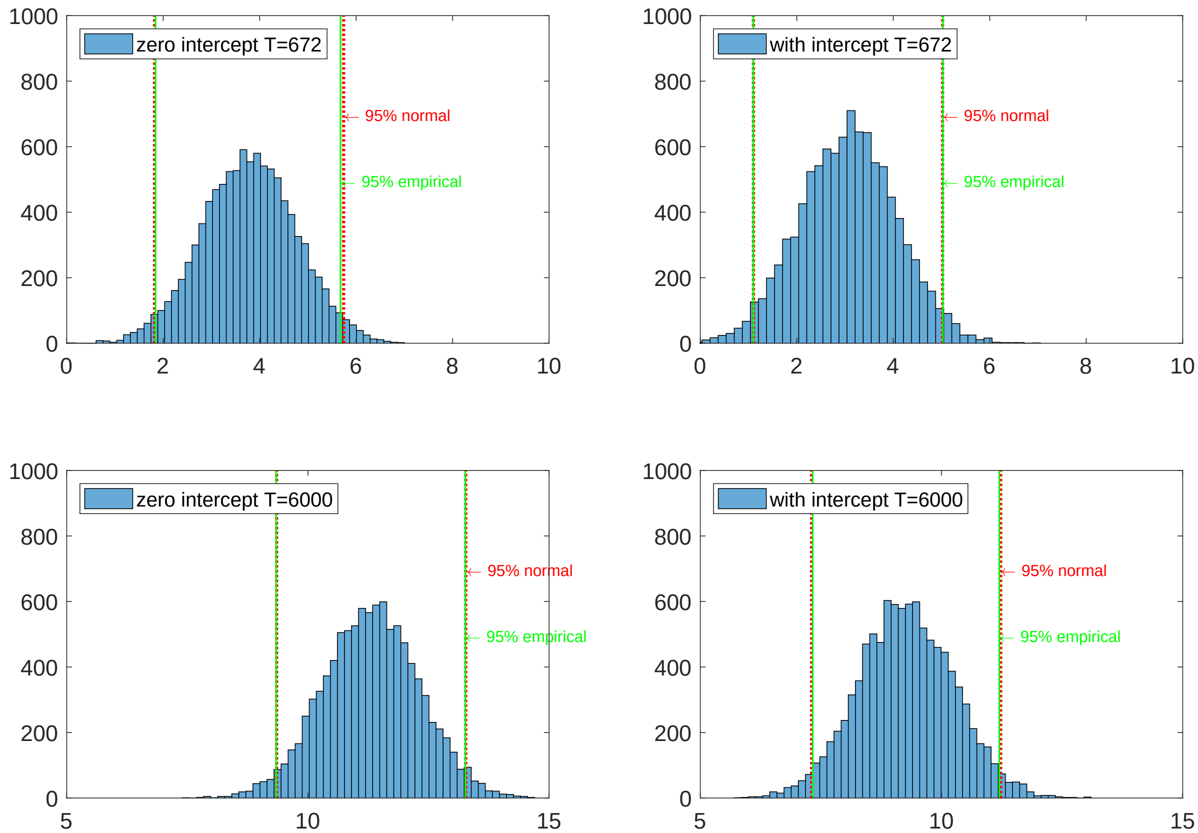


This figure presents the distribution of the $t(\lambda = 0)$ when the factor is useless using Fama-MacBeth/Shanken t-statistics using 10,000 Monte Carlo simulations. Estimation is with or without the intercept and the number of observations in each simulation is T=672 or T=6,000. Red lines indicate the 95% confidence set under the assumption of a normal distributed t-statistics while the green lines indicate the 95% confidence set based on the 10,000 simulations.

[25]Intuitively, the distribution is bi-modal because the estimated prices of risk of a useless factor can be positive or negative with equal probability when betas are around zero. There is only little mass around zero, as the price of risk estimates tend to be large when estimated betas are small.

Figure 20 provides the distribution of the t-statistic for a useful factor ($\rho = 1.00$) for comparisons. These t-statistics are normally distributed and asymptotic inference is reliable.

Figure 20: Distribution of Fama-MacBeth/Shanken t-statistics: Useful Factors



This figure presents the distribution of the $t(\lambda = 0)$ when the factor is useless using Fama-MacBeth/Shanken t-statistics using 10,000 Monte Carlo simulations. Estimation is with or without the intercept and the number of observations in each simulation is T=672 or T=6,000. Red lines indicate the 95% confidence set under the assumption of a normal distributed t-statistics while the green lines indicate the 95% confidence set based on the 10,000 simulations.

When Does FMB/Shanken Under/over-reject Useless Factors When T is Fixed?

As pointed out by Kan and Zhang (1999), the classic Fama-MacBeth/Shanken approach does not over-reject in small samples, as the estimation error in betas is sufficient to ensure that a researcher is likely to find a price of risk close the zero. Intuitively, it is unlikely in small samples to run into a "division-by-zero-problem", because estimation error in betas is sufficiently large. By increasing the sample size, estimation error in betas gets smaller while test asset mean returns remain unchanged.

In Figure 21, we illustrate that the useless factor problem can be more/less problematic even when keeping the estimation error in betas is fixed. To this end, we vary the dispersion in mean returns (upper figure), by varying the scale parameter s in

$$\bar{R}_i' = \frac{1}{N}\sum_{i=1}^{N}\bar{R}_i + \left(\bar{R}_i - \frac{1}{N}\sum_{i=1}^{N}\bar{R}_i\right) \times s.$$

Alternatively, we vary the overall mean returns (lower figure), i.e., we vary s in

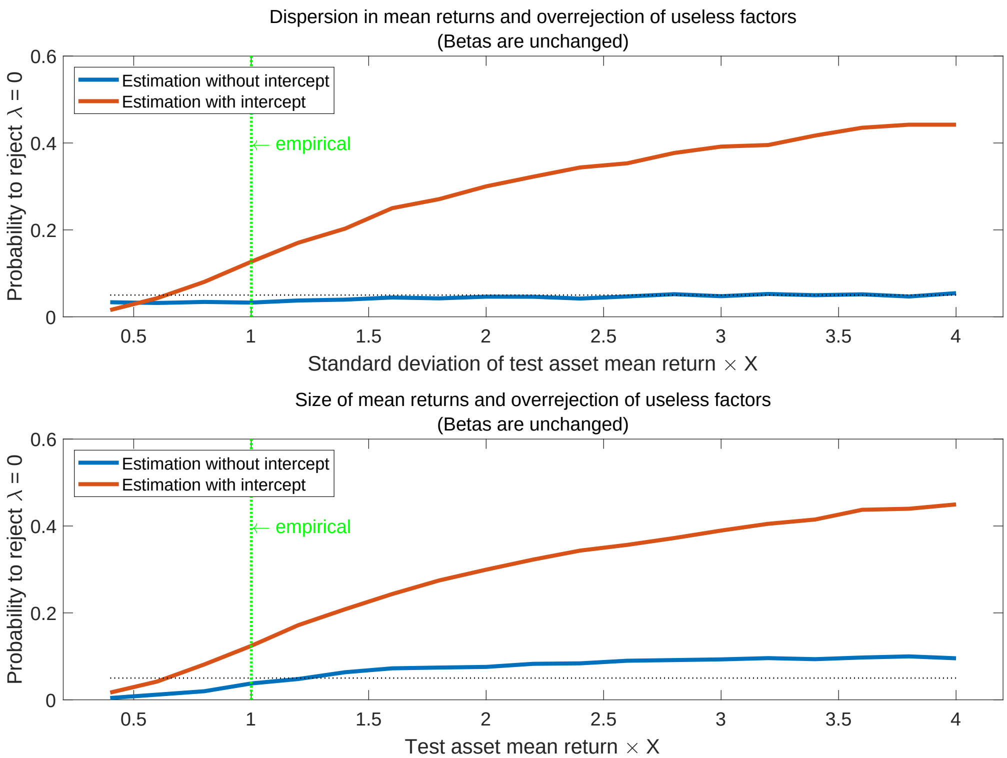
$$\bar{R}_i' = \bar{R}_i \times s.$$

In both cases, we keep the factor covariance matrix unchanged.

We find that the over-rejection problem is more severe when there is larger cross-sectional dispersion in mean returns, in particular when estimation is conducted with an intercept. Estimation without an intercept is robust against a larger spread in mean returns. However, as the lower figure shows, the specification with a zero intercept is more prone to over-rejection when the mean returns are overall increased relative to the betas.

These results span rather extreme scenarios and are mainly intended for illustrative purposes. The test assets (25 Fama-French portfolios) are popular because of their already large spread in mean returns; and stock returns have relatively large overall mean returns among financial assets.

Figure 21: Varying Mean Returns Holding T Fixed



This figure presents the probability to reject that a useless factor is priced based on 10,000 Monte Carlo simulations. Inference is based on Fama-MacBeth/Shanken t-statistics. Estimation is with or without the intercept and the number of observations in each simulation is T=672. In the upper figure, the dispersion of mean returns are decreased/increased as displayed on the x-axis; the mean of the mean returns is the same as in the empirical data. In the lower figure, the mean returns are decreased/increased as displayed on the x-axis. The factor betas (covariance matrix) is unchanged in all specifications.

5 Conclusion

We summarize our results that are specific to certain estimators as follows:

- Inference based on the Fama-MacBeth methodology that comes with the usual "textbook" refinements is reasonably trustworthy. This is especially true when the intercept is imposed to be zero (as suggested by Cochrane, 2005). For empirically relevant cross-sections ($N = 25$) and sample sizes ($T \leq 672$), we find that it is unlikely that the hypothesis of factor pricing is over-rejected for a useless factor. Importantly, the traditional approach has a high power to detect useful factors. In case of doubt, we advise reporting bootstrapped confidence intervals, which are robust to misspecification and cannot lead to over-rejection of useless factors.

- The misspecification robust standard errors, as proposed by Kan et al. (2013) are helpful. They tend to reduce the useless factor problem without sacrificing power. The same applies to the Bayesian Fama-MacBeth approach proposed by Bryzgalova et al. (2020), except for small sample sizes (e.g., annual data, $T \leq 56$), where we are concerned about low power.

- The SDF-based approach proposed by Gospodinov et al. (2014) comes with an advantage only when the moment restrictions are weighted by the covariance matrix of the test asset returns. However, we find that traditional model estimation with identical weighting and imposing a zero common pricing error is similarly robust to the useless factor problem in empirical sample sizes but comes with a higher power to detect useful factors. The GRS-FAR test (Kleibergen and Zhan, 2020) has no power to detect useful factors for misspecified models. Similarly, the three-pass method by Giglio and Xiu (2021) has low power for misspecified models. However, these methods might be useful in the analyses of alternative hypotheses which are outside the scope of our study. Moreover, these papers provide important theoretical results that might be fruitful for developing future alternative estimators; this is not recognized by our research design.

We make some more general observations when reviewing and replicating the literature that proposes alternative approaches:

- We detect a common practice in the literature to put a strong focus on the size of a test, but to omit a discussion of the power to detect a useful factor. However, we argue that a good balance between type I and type II error should be an important criterion when recommending one method over another.

- Several papers limit the analysis to correctly specified asset pricing models. This is in contrast with a literature that largely agrees that asset pricing models should be viewed as imperfectly specified. We find that the performance of "robust" asset pricing tests can be surprisingly "unrobust" even for only mildly misspecified models.

- Several papers focus (or even limit) the analysis to analytical results, describing only the large sample properties of an estimator. We find that the small sample properties can be remarkably different, echoing Kan and Zhang (1999). To study these, one has to rely on simulation or bootstrapping techniques.

- Surprisingly, almost all papers on "robust" asset pricing do not consider the case of imposing a zero intercept, even though this is a common approach in the literature and even the "textbook" baseline approach (Cochrane, 2005).

Taken together, we are concerned that many papers on robust asset pricing tend to report cases where the new approach works well. However, empirically relevant cases tend to be under-represented in the analysis (small samples, misspecified models, imposing a zero intercept). We argue that the bar in the literature should be increased and future alternative "robust" asset pricing approaches should be more thoroughly evaluated based on empirically relevant cases.

Appendix A Monte Carlo simulation

Table 3: Monte Carlo Simulation, T=672

This table shows the probability to reject the hypothesis (5% significance level) of a zero price of risk, $\bar{R}_i^e = (\lambda_0+)\beta_i\lambda + a_i$, H0: $\lambda = 0$, as well as a zero SDF loading, $SDF_t = 1 - (F_t - \mu_f)b$, H0: $b = 0$. The probability to reject is based on 10,000 (1,000 for Bayesian inference and the Bootstrap confidence intervals) Monte Carlo simulations imposing a multivariate normal distribution. The Bayesian inference and the bootstrap confidence intervals are based on 500 resamples for each of the 1,000 simulated samples. Each Monte Carlos sample has T=672 time-series observations.

Methods: The methods are the same as described in the main paper.

Test assets and risk factors: The risk factors and the test assets are the same as described in the caption to Figure 2, monthly data.

Monte Carlo Simulation, T=672					
	useless factor	empirical misspec.	correct spec.	slope misspec.	intercept misspec.
Price of risk: $reject\ H0(5\%) : \lambda = 0?\ zero\ intercept$					
$FMB - OLS$	0.89	0.97	0.97	0.96	0.05
$FMB - Shanken$	0.04	0.97	0.97	0.95	0.05
$Bayesian - FMB$	0.04	1.00	1.00	0.99	0.22
$Robust\ GRS - FAR$	0.01	0.00	0.96	0.00	0.22
$Bootstrap\ CIs$	0.05	0.98	0.97	0.96	0.05
Price of risk: $reject\ H0(5\%) : \lambda = 0?\ with\ intercept$					
$FMB - OLS$	0.37	0.87	0.94	0.05	0.94
$FMB - Shanken$	0.13	0.87	0.94	0.05	0.94
$Robust\ FMB$	0.05	0.86	0.94	0.05	0.94
$Bayesian - FMB$	0.01	0.94	0.98	0.15	0.97
$Three\ pass\ method$	0.01	0.07	0.97	0.50	0.97
$Robust\ GRS - FAR$	0.02	0.00	0.92	0.00	0.92
$Bootstrap\ CIs$	0.01	0.93	0.86	0.06	0.93
SDF loading: $reject\ H0(5\%)\ b = 0?$					
$SDF - W = I,\ zero\ int.$	0.04	0.97	0.97	0.96	0.05
$SDF - W = I,\ with\ int.$	0.13	0.87	0.94	0.05	0.94
$SDF - W = S^{-1},\ with\ int.$	0.40	1.00	1.00	0.95	0.35
$Robust\ SDF - W = S^{-1},\ with\ int.$	0.04	1.00	1.00	0.94	0.36

Table 4: Monte Carlo Simulation, T=120

This table shows the probability to reject the hypothesis (5% significance level) of a zero price of risk, $\bar{R}_i^e = (\lambda_0+)\beta_i\lambda + a_i$, H0: $\lambda = 0$, as well as a zero SDF loading, $SDF_t = 1 - (F_t - \mu_f)b$, H0: $b = 0$. The probability to reject is based on 10,000 (1,000 for Bayesian inference and the Bootstrap confidence intervals) Monte Carlo simulations imposing a multivariate normal distribution. The Bayesian inference and the bootstrap confidence intervals are based on 500 resamples for each of the 1,000 simulated samples. Each Monte Carlos sample has T=120 time-series observations.

Methods: The methods are the same as described in the main paper.

Test assets and risk factors: The risk factors and the test assets are the same as described in the caption to Figure 2, monthly data.

Monte Carlo Simulation, T=120					
	useless factor	empirical misspec.	correct spec.	slope misspec.	intercept misspec.
Price of risk: $reject\ H0(5\%) : \lambda = 0?\ zero\ intercept$					
$FMB - OLS$	0.31	0.37	0.38	0.34	0.05
$FMB - Shanken$	0.00	0.35	0.36	0.33	0.05
$Bayesian - FMB$	0.00	0.55	0.57	0.55	0.12
$Robust\ GRS - FAR$	0.02	0.50	0.92	0.23	0.65
$Bootstrap\ CIs$	0.01	0.36	0.37	0.35	0.04
Price of risk: $reject\ H0(5\%) : \lambda = 0?\ with\ intercept$					
$FMB - OLS$	0.12	0.23	0.29	0.05	0.27
$FMB - Shanken$	0.02	0.22	0.27	0.04	0.27
$Robust\ FMB$	0.02	0.19	0.24	0.04	0.25
$Bayesian - FMB$	0.00	0.24	0.30	0.06	0.30
$Three\ pass\ method$	0.00	0.10	0.37	0.18	0.38
$Robust\ GRS - FAR$	0.01	0.16	0.28	0.19	0.29
$Bootstrap\ CIs$	0.00	0.23	0.18	0.03	0.25
SDF loading: $reject\ H0(5\%)\ b = 0?$					
$SDF - W = I,\ zero\ int.$	0.01	0.34	0.35	0.33	0.05
$SDF - W = I,\ with\ int.$	0.02	0.20	0.25	0.03	0.24
$SDF - W = S^{-1},\ with\ int.$	0.16	0.54	0.54	0.30	0.09
$Robust\ SDF - W = S^{-1},\ with\ int.$	0.02	0.54	0.53	0.31	0.09

Appendix B Fama-MacBeth Standard Errors and Refinements

In this section, we summarize the classic Fama-MacBeth formulas and their refinements from the literature. The derivations of the formulas are provided by, e.g., Cochrane (2005) and Burnside (2011).

Variables: A time-series regression of n test asset excess returns on k factors can be written as:

$$\mathbf{R}_t^e = \boldsymbol{\alpha} + \boldsymbol{\beta} \boldsymbol{F}_t + \boldsymbol{\varepsilon}_t, \tag{12}$$

where $\mathbf{R}_t^e$ is an $n \times 1$ vector of excess returns, $\boldsymbol{\alpha}$ is an $n \times 1$ vector of alphas, $\boldsymbol{F}_t$ is an $k \times 1$ vector of risk factors, $\boldsymbol{\beta}$ is an $n \times k$ matrix of factor betas, and $\boldsymbol{\varepsilon}_t$ is $n \times 1$ vector of regression residuals. In addition, we define the $k \times k$ covariance matrix of the factors, $Var\left(\boldsymbol{F}_t\right) = \Sigma_f$, the $n \times n$ covariance matrix of the regression residuals, $\mathbf{Var}\left(\boldsymbol{\varepsilon}_t\right) = \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$, and the averages over time as $\bar{\mathbf{R}}^e$, $\bar{\mathbf{F}}$, and $\bar{\boldsymbol{\varepsilon}}$, and the $n \times k$ matrix of factor betas, $\boldsymbol{\beta} = cov\left(\mathbf{R}_t^e, \mathbf{F}_t\right) \boldsymbol{\Sigma}_{\boldsymbol{f}}^{\mathbf{-1}}$.

OLS standard errors without a constant: These standard errors are not HAC-robust and also do not account for the fact that the betas are estimated.

The $k \times 1$ vector of slope coefficient lambda estimates is $\hat{\boldsymbol{\lambda}} = \left(\hat{\boldsymbol{\beta}}' \hat{\boldsymbol{\beta}}\right)^{-1} \hat{\boldsymbol{\beta}}' \bar{\mathbf{R}}^e$. Defining $\hat{\mathbf{A}} = \left(\hat{\boldsymbol{\beta}}' \hat{\boldsymbol{\beta}}\right)^{-1} \hat{\boldsymbol{\beta}}'$, with $\mathbf{A} = plim \hat{\mathbf{A}}$, allows to find the variance of $\hat{\boldsymbol{\lambda}}$ as:

$$V_{OLS}\left[\hat{\boldsymbol{\lambda}}\right] = \frac{1}{T}\left[\mathbf{A} \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} \mathbf{A}' + \boldsymbol{\Sigma}_f\right].$$

The according standard errors are the square root of the diagonal of this matrix and account for cross-sectional correlation. They are also equivalent to the classic Fama-MacBeth standard errors when $\hat{\boldsymbol{\beta}}$ does not change over time.

OLS standard errors with a constant: When a constant is included in the second stage, the $(k+1) \times 1$ cross-sectional estimates are

$$\hat{\boldsymbol{\theta}} = \left(\hat{\mathbf{X}}'\hat{\mathbf{X}}\right)^{-1} \hat{\mathbf{X}}'\bar{\mathbf{R}}^{e},$$

where $\hat{\boldsymbol{\theta}} = \left(\hat{\lambda}_0 \; \hat{\boldsymbol{\lambda}}'\right)'$ and $\hat{\mathbf{X}} = \begin{pmatrix} \mathbf{1}_{n\times 1} & \hat{\boldsymbol{\beta}} \end{pmatrix}$. Defining $\hat{\mathbf{B}} = \left(\hat{\mathbf{X}}'\hat{\mathbf{X}}\right)^{-1} \hat{\mathbf{X}}'$, with $\mathbf{B} = plim\hat{\mathbf{B}}$, $Var\left(\mathbf{F}_t\right) = \boldsymbol{\Sigma}_f$, $\mathbf{Var}\left(\varepsilon_t\right) = \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$, and

$$\bar{\boldsymbol{\Sigma}}_{\boldsymbol{f}} = \begin{pmatrix} 0 & \mathbf{0}_{1\times n} \\ \mathbf{0}_{n\times 1} & \boldsymbol{\Sigma}_f \end{pmatrix},$$

the standard errors with Shanken correction can be derived from the covariance matrix:

$$V_{OLS}\left[\hat{\boldsymbol{\theta}}\right] = \frac{1}{T}\left[\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\mathbf{B}' + \bar{\boldsymbol{\Sigma}}_{\boldsymbol{f}}\right].$$

Shanken standard errors without a constant: The Shanken standard errors are not HAC-robust but they account for the fact that the betas are estimated.

They are based on modified formula:

$$V_{Shanken}\left[\hat{\boldsymbol{\lambda}}\right] = \frac{1}{T}\left[\left(1 + \boldsymbol{\lambda}\boldsymbol{\Sigma}_f^{-1}\boldsymbol{\lambda}\right)\mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\mathbf{A}' + \boldsymbol{\Sigma}_f\right].$$

For a single traded factor, the adjustment term, $\left(1 + \boldsymbol{\lambda}\boldsymbol{\Sigma}_f^{-1}\boldsymbol{\lambda}\right)$, is equivalent to one plus the squared factor Sharpe ratio.

Shanken standard errors with a constant: When a constant is included in the second stage, the standard errors with Shanken correction can be derived from the covariance matrix:

$$V_{Shanken}\left[\hat{\boldsymbol{\theta}}\right] = \frac{1}{T}\left[\left(1 + \boldsymbol{\lambda}\boldsymbol{\Sigma}_f^{-1}\boldsymbol{\lambda}\right)\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\mathbf{B}' + \bar{\boldsymbol{\Sigma}}_{\boldsymbol{f}}\right].$$

***GMM standard errors without a constant*:** These standard errors are HAC-robust <u>and</u> account for the fact the betas are estimated ("Shanken correction"). Start with writing down the following moment restrictions for n test assets:

$$E\left[R_{i,t}^{e} - a_i - \boldsymbol{\beta}_i \mathbf{F}_t\right] = 0, \ i = 1...., n,$$

$$E\left[R_{i,t}^{e} - a_i - \boldsymbol{\beta}_i \mathbf{F}_t\right] F_t = 0, \ i = 1...., n,$$

$$E\left[R_{i,t}^{e} - \boldsymbol{\beta}_i \boldsymbol{\lambda}\right] = 0, \ i = 1...., n.$$

Define $\mathbf{f}_t = \begin{pmatrix} 1 & \mathbf{F}_t' \end{pmatrix}'$, $\mathbf{b}_i = \begin{pmatrix} a_i & \boldsymbol{\beta}_i' \end{pmatrix}'$, and $\boldsymbol{\delta} = \begin{pmatrix} \mathbf{b}_1 & \mathbf{b}_2 & ... & \mathbf{b}_n & \boldsymbol{\lambda} \end{pmatrix}'$. A $n\left((k+1)+1\right) \times 1$ vector collecting the errors at time t is

$$\mathbf{u}_t\left(\boldsymbol{\delta}\right) = \begin{pmatrix} \mathbf{f}_t\left(R_{1,t}^{e} - \mathbf{f}_t'\mathbf{b}_1\right) \\ \mathbf{f}_t\left(R_{2,t}^{e} - \mathbf{f}_t'\mathbf{b}_2\right) \\ ... \\ \mathbf{f}_t\left(R_{n,t}^{e} - \mathbf{f}_t'\mathbf{b}_n\right) \\ \mathbf{R}_t^{e} - \boldsymbol{\beta}\boldsymbol{\lambda} \end{pmatrix}.$$

Accordingly, the $n\left(2+1\right) \times 1$ vector of average errors is $\mathbf{g}_T\left(\boldsymbol{\delta}\right) = \frac{1}{T}\Sigma_{t=1}^{T}\mathbf{u}_t\left(\boldsymbol{\delta}\right)$. When the $[n(k+1)+1] \times [n\left((k+1)+1\right)]$ matrix $\mathbf{a}_T$ is set to

$$\mathbf{a}_T = \begin{pmatrix} \mathbf{I}_{n(k+1)} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\beta}' \end{pmatrix},$$

the GMM estimator $\mathbf{a}_T\mathbf{g}_T\left(\hat{\boldsymbol{\delta}}\right) = 0$ results in the same point estimates of $\boldsymbol{a}$, $\boldsymbol{\beta}$, and $\boldsymbol{\lambda}$ as the traditional Fama-MacBeth two-pass estimate *without* a constant.

For standard errors of the estimates, define $\mathbf{M} = \frac{1}{T}\Sigma_{t=1}^{T}\mathbf{f}_t\mathbf{f}_t'$,

$$\mathbf{d}_T = \frac{\partial \mathbf{g}_T\left(\hat{\boldsymbol{\delta}}\right)}{\partial \hat{\boldsymbol{\delta}}} = \begin{pmatrix} -\mathbf{I}_n \otimes \mathbf{M} & \mathbf{0}_{n(k+1)\times 1} \\ -\mathbf{I}_n \otimes \begin{pmatrix} 0 & \hat{\boldsymbol{\lambda}}' \end{pmatrix} & -\boldsymbol{\beta} \end{pmatrix},$$

$\mathbf{a} = plim\ \mathbf{a}_T$, $\mathbf{d} = plim\ \mathbf{d}_T$, and $\mathbf{S} = \sum_{j=-\infty}^{\infty} E\left[\mathbf{u}_t \mathbf{u}_{t-j}'\right]$. Then the GMM formulas provide the covariance matrix of $\hat{\boldsymbol{\delta}}$,

$$\mathbf{V}_{GMM}\left[\hat{\boldsymbol{\delta}}\right] = \frac{1}{T}\left(\mathbf{ad}\right)^{-1}\mathbf{aSa}'\left[\left(\mathbf{ad}\right)^{-1}\right]',$$

where the standard errors are the square root of the diagonal of this matrix.

GMM standard errors with a constant: The last set of moment conditions changes to $E\left[R_{i,t}^e - \lambda_0 - \boldsymbol{\beta}_i \boldsymbol{\lambda}\right] = 0,\ i = 1....,n$. Accordingly, define $\boldsymbol{\vartheta} = \begin{pmatrix} \mathbf{b}_1 & \mathbf{b}_2 & ... & \mathbf{b}_n & \lambda_0 & \boldsymbol{\lambda} \end{pmatrix}'$ and $\hat{\mathbf{X}} = \begin{pmatrix} \mathbf{1}_{n\times 1} & \hat{\boldsymbol{\beta}} \end{pmatrix}$. A $n\left((k+1)+1\right)\times 1$ vector collecting the errors at time t is

$$\mathbf{u}_t\left(\boldsymbol{\vartheta}\right) = \begin{pmatrix} \mathbf{f}_t\left(R_{1,t}^e - \mathbf{f}_t'\mathbf{b}_1\right) \\ \mathbf{f}_t\left(R_{2,t}^e - \mathbf{f}_t'\mathbf{b}_2\right) \\ \cdots \\ \mathbf{f}_t\left(R_{n,t}^e - \mathbf{f}_t'\mathbf{b}_n\right) \\ \mathbf{R}_t^e - \lambda_0 - \boldsymbol{\beta}\lambda \end{pmatrix}.$$

The matrix $\mathbf{a}_T$ is becomes

$$\mathbf{a}_T = \begin{pmatrix} \mathbf{I}_{n(k+1)} & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{X}}' \end{pmatrix},$$

the GMM estimator $\mathbf{a}_T\mathbf{g}_T\left(\hat{\boldsymbol{\vartheta}}\right) = 0$ results in the same point estimates of $\boldsymbol{a}$, $\boldsymbol{\beta}$, λ_0, and $\boldsymbol{\lambda}$ as the traditional Fama-MacBeth two-pass estimate *with* a constant.

For standard errors of the estimates, define

$$\mathbf{d}_T = \frac{\partial \mathbf{g}_T\left(\hat{\boldsymbol{\vartheta}}\right)}{\partial \hat{\boldsymbol{\vartheta}}} = \begin{pmatrix} -\mathbf{I}_n \otimes \mathbf{M} & \mathbf{0}_{n(k+1)\times(k+1)} \\ -\mathbf{I}_n \otimes \begin{pmatrix} 0 & \hat{\boldsymbol{\lambda}}' \end{pmatrix} & -\hat{\mathbf{X}} \end{pmatrix},$$

$\mathbf{a} = plim\, \mathbf{a}_T$, $\mathbf{d} = plim\, \mathbf{d}_T$, and $\mathbf{S} = \sum_{j=-\infty}^{\infty} E\left[\mathbf{u}_t \mathbf{u}_{t-j}'\right]$. Then the GMM formulas provide the covariance matrix of $\hat{\boldsymbol{\vartheta}}$ as

$$\mathbf{V}_{GMM}\left[\hat{\boldsymbol{\vartheta}}\right] = \frac{1}{T}\left(\mathbf{ad}\right)^{-1} \mathbf{aSa}' \left[\left(\mathbf{ad}\right)^{-1}\right]'.$$

References

Adrian, T., Etula, E., Muir, T., 2014. Financial intermediaries and the cross-section of asset returns. Journal of Finance 69, 2557–2596.

Ait-Sahalia, Y., Parker, J.A., Yogo, M., 2004. Luxury goods and the equity premium. Journal of Finance 59, 2959–3004.

Anderson, T.W., Rubin, H., et al., 1949. Estimation of the parameters of a single equation in a complete system of stochastic equations. Annals of Mathematical statistics 20, 46–63.

Barroso, P., Boons, M., Karehnke, P., 2021. Time-varying state variable risk premia in the icapm. Journal of Financial Economics 139, 428–451.

Boguth, O., Kuehn, L.A., 2013. Consumption volatility risk. Journal of Finance 68, 2589–2615.

Breeden, D.T., Gibbons, M., Litzenberger, R., 1989. Empirical tests of the consumption-oriented CAPM. Journal of Finance 44, 231–262.

Bryzgalova, S., Huang, J., Julliard, C., 2020. Bayesian solutions for the factor zoo: We just ran two quadrillion models. Working paper .

Burnside, C., 2011. The forward rate premium is still a puzzle, a comment on "The cross-section of foreign currency risk premia and consumption growth risk". American Economic Review 101, 3456–3476.

Campbell, J., Vuolteenaho, T., 2004. Bad beta, good beta. American Economic Review 94, 1249–1275.

Campbell, J.Y., 2017. Financial Decisions and Markets: A Course in Asset Pricing. Princeton Univers. Press.

Campbell, J.Y., Giglio, S., Polk, C., Turley, R., 2018. An intertemporal capm with stochastic volatiltiy. Journal of Financial Economics 128, 207–233.

Cochrane, J.H., 2005. Asset Pricing. Princeton University Press, New Jersey.

Cochrane, J.H., 2011. Presidential address: Discount rates. Journal of Finance 66, 1046–1108.

Fama, E.F., 1976. Foundations of Finance. Basic Books, New York.

Fama, E.F., French, K.R., 1993. Common risk factors in the returns on stocks and bonds. Journal of Financial Economics 33, 3–53.

Fama, E.F., French, K.R., 2015. A five-factor asset pricing model. Journal of Financial Economics 116, 1–22.

Fama, E.F., MacBeth, J.D., 1973. Risk, return, and equilibrium: Emprical tests. Journal of Political Economy 81, 607–636.

Ferson, W., 2019. Empirical Asset Pricing: Models and Methods. The MIT Press.

Gibbons, M.R., Ross, S.A., Shanken, J., 1989. A test of the efficiency of a given portfolio. Econometrica 57, 1121–1152.

Giglio, S., Xiu, D., 2021. Asset pricing with omitted factors. Journal of Political Economy Forthomcing.

Gospodinov, N., Kan, R., Robotti, C., 2014. Misspecification-robust inference in linear asset-pricing models with irrelevant risk factors. Review of Financial Studies 27, 2139–2170.

He, Z., Kelly, B., Manela, A., 2017. Intermediary asset pricing: new evidence from many asset classes. Journal of Financial Economics 126, 1–35.

Hou, K., Xue, C., Zhang, L., 2015. Digesting anomalies: An investment approach. Review of Financial Studies 28, 650–705.

Huberman, G., Kandel, S., 1987. Mean-variance spanning. Journal of Finance 42, 873–888.

Jagannathan, R., Wang, Y., 2007. Lazy investors, discretionary consumption, and the cross-section of stock returns. Journal of Finance 62, 1623–1661.

Jagannathan, R., Wang, Z., 1996. The conditional capm and the cross-section of expected returns. Journal of Finance 51, 3.53.

Kan, R., Robotti, C., Shanken, J., 2013. Pricing model performance and the two pass cross-sectional regression methodology. Journal of Finance 68, 2617–2649.

Kan, R., Zhang, C., 1999. Two pass tests of asset pricing models with useless factors. Journal of Finance 54, 203–235.

Kleibergen, F., 2009. Tests of risk premia in linear factor models. Journal of Econometrics 149, 149–173.

Kleibergen, F., Zhan, Z., 2020. Robust inference for consumption-based asset pricing. Journal of Finance forthcoming.

Kroencke, T.A., 2017. Asset pricing without garbage. Journal of Finance 72, 47–98.

Kroencke, T.A., 2020. On robust inference for consumption-based asset pricing. Working Paper.

Lettau, M., Ludvigson, S., 2001. Resurrecting the (c)capm: A cross-sectional test when risk premia are time-varying. Journal of Political Economy 109, 1238–1287.

Lettau, M., Maggiori, M., Weber, M., 2014. Conditional risk premia in currency markets and other asset classes. Journal of Financial Economics 114, 197–225.

Lewellen, J., Nagel, S., Shanken, J., 2010. A skeptical appraisal of asset pricing tests. Journal of Financial Economics 96, 175–194.

Lustig, H., Verdelhan, A., 2007. The cross section of foreign currency risk premia and consumption growth risk. American Economic Review 97, 89–117.

Menkhoff, L., Sarno, L., Schmeling, M., Schrimpf, A., 2012. Carry trades and global foreign exchange volatility. Journal of Finance 67, 681–718.

Parker, J.A., Julliard, C., 2005. Consumption risk and the cross section of expected returns. Journal of Political Economy 113, 185–222.

Santos, T., Veronesi, P., 2006. Labor income and predictable stock returns. Review of Financial Studies 19, 1–44.

Savov, A., 2011. Asset pricing with garbage. Journal of Finance 66, 177–201.

Shanken, J., 1992. On the estimation of beta-pricing models. Review of Financial Studies 5, 1–33.