Journal of Financial Economics 132 (2019) 451–471

Contents lists available at ScienceDirect

Journal of Financial Economics

journal homepage: www.elsevier.com/locate/jfec

Too good to be true? Fallacies in evaluating risk factor models☆

Nikolay Gospodinov[a,*], Raymond Kan[b], Cesare Robotti[c]

[a] Research Department, Federal Reserve Bank of Atlanta, 1000 Peachtree St NE, Atlanta, GA 30309, United States
[b] Rotman School of Management, University of Toronto, 105 St. George Street, Toronto, ON, M5S 3E6, Canada
[c] Warwick Business School, The University of Warwick, Coventry CV4 7AL, United Kingdom

ARTICLE INFO

Article history:
Received 10 October 2017
Revised 25 January 2018
Accepted 16 February 2018
Available online 27 October 2018

JEL classification:
G12
C12
C13

Keywords:
Asset pricing
Spurious risk factors
Unidentified models
Model misspecification
Maximum likelihood
Goodness of fit
Rank test

ABSTRACT

This paper is concerned with statistical inference and model evaluation in possibly misspecified and unidentified linear asset pricing models estimated by maximum likelihood. Strikingly, when spurious factors (that is, factors that are uncorrelated with the returns on the test assets) are present, the model exhibits perfect fit, as measured by the squared correlation between the model's fitted expected returns and the average realized returns. Furthermore, factors that are spurious are selected with high probability, and factors that are useful are driven out of the model. While ignoring potential misspecification and lack of identification can be very problematic for models with macroeconomic factors, empirical specifications with traded factors (e.g., Fama and French, 1993; Hou et al., 2015) do not suffer from the identification problems shown in this study.

1. Introduction and motivation

The search for theoretically justified or empirically motivated risk factors that improve the pricing performance of various asset pricing models has generated a large, and constantly growing, literature in financial economics. A typical empirical strategy involves the development of a structural asset pricing model and the evaluation of the pricing ability of the proposed factors in the linearized version of the model using actual data. The resulting linear asset pricing model can be estimated and tested using a beta representation. Given the appealing efficiency and invariance properties of the maximum likelihood (ML) estimator, opting for this or other invariant estimators seems natural when conducting statistical inference (estimation, testing, and model evaluation) in these linear asset pricing models (see, for example, Shanken and Zhou, 2007; Almeida and Garcia, 2012; Almeida and Garcia, 2017; Peñaranda and Sentana, 2015; Manresa et al., 2017; Barillas and Shanken, 2017; Barillas and Shanken, 2018; Ghosh et al., 2017). Often, a high correlation between the realized and fitted expected returns or statistically small model pricing errors appears to be sufficient for the applied researcher to conclude that the model is well specified and proceed with testing for statistical significance of the risk premium

☆ We are grateful to the editor, an anonymous referee, and numerous seminar and conference participants for insightful comments and suggestions. The views expressed here are ours and not necessarily those of the Federal Reserve Bank of Atlanta or the Federal Reserve System.

* Corresponding author.
E-mail addresses: Nikolay.Gospodinov@atl.frb.org (N. Gospodinov), kan@chass.utoronto.ca (R. Kan), Cesare.Robotti@wbs.ac.uk (C. Robotti).

https://doi.org/10.1016/j.jfineco.2018.10.012
0304-405X/

parameters using the standard tools for inference. Many asset pricing studies have followed this empirical strategy and, collectively, have identified a large set of macroeconomic and financial factors (see Harvey et al., 2016; Feng et al., 2017) that are believed to explain the cross-sectional variation in various portfolio expected returns, such as the expected returns on the 25 Fama–French size and book-to-market ranked portfolios.

Despite these advances in the asset pricing literature, two observations that consistently emerge in empirical work could call for a more cautious approach to statistical validation and economic interpretation of asset pricing models. First, all asset pricing models should be viewed only as approximations to reality and, hence, potentially misspecified. Overwhelming empirical evidence, mainly based on non-invariant estimators, suggests that the asset pricing models used in practice are misspecified. This raises the concern of using standard errors, derived under the assumption of correct model specification, that tend to underestimate the degree of uncertainty that the researcher faces. Second, the macroeconomic factors in several asset pricing specifications appear to be only weakly correlated with the portfolio returns. As a result, many of these macroeconomic factors could be irrelevant for pricing and for explaining the cross-sectional variation in expected equity returns. The inclusion of spurious factors, defined as factors that are uncorrelated with the returns on the test assets, leads to serious identification issues regarding the parameters associated with all risk factors and gives rise to a nonstandard statistical inference (see, for instance, Gospodinov et al., 2014a).

Under standard regularity conditions (that include global and local identification as well as correct model specification), the ML estimator considered here, which is invariant to data scaling, reparameterizations, and normalizations, is asymptotically well behaved and efficient. We show that, in the presence of spurious factors, the tests and goodness-of-fit measures based on this estimator could be highly misleading. In summary, we argue that the standard inference procedures based on the ML estimator lead to spurious results suggesting that the model is correctly specified and the risk premium parameters are highly significant (i.e., the risk factors are priced) when, in fact, the model is misspecified and the factors are irrelevant.

To illustrate the seriousness of the problem, we start with some numerical evidence on the widely studied static capital asset pricing model (CAPM) with the market excess return (the return on the value-weighted NYSE, Amex, and Nasdaq stock market index in excess of the one-month T-bill rate, *vw*) as a risk factor. The test asset returns are the monthly returns on the popular value-weighted 25 Fama–French size and book-to-market ranked portfolios from January 1967 until December 2012.

The first column of Table 1 reports some conventional statistics for evaluating the performance of the CAPM in the beta-pricing framework estimated by ML. The statistics include the test of correct model specification $\mathcal{S}$ (Shanken, 1985), the t-statistics of statistical significance constructed using standard errors that assume correct model specification, and the (pseudo) R^2 computed as the squared correlation between the fitted expected returns and average returns. In line with the results reported elsewhere in the literature, the market factor appears to be characterized by a statistically significant risk premium. Also, consistent with the existing studies, the CAPM is rejected by the data. This requires the use of misspecification-robust standard errors in constructing the t-statistics (see Gospodinov et al., 2018). Finally, the R^2 points to some, but not particularly strong, explanatory power.

Table 1
Test statistics for capital asset pricing model (CAPM) and CAPM augmented with the *sp* factor.

The table reports test statistics for the capital asset pricing model (CAPM), the CAPM augmented with the *sp* factor, and a model with the *sp* factor only. The test asset returns are the monthly returns on the value-weighted 25 Fama–French size and book-to-market ranked portfolios from January 1967 until December 2012. $\mathcal{S}$ denotes the Shanken (1985) test of correct model specification based on the maximum likelihood estimator. t_x denotes the t-test of statistical significance for the parameter associated with factor x, with standard errors computed under the assumption of correct model specification. R^2 denotes the squared correlation coefficient between the fitted expected returns and the average realized returns.

	CAPM	CAPM + *sp* factor	*sp* factor
t_{vw}	−3.65	0.52	
(p-value)	(0.0003)	(0.6027)	
t_{sp}		−4.62	−4.64
(p-value)		(0.0000)	(0.0000)
$\mathcal{S}$	68.79	21.32	21.56
(p-value)	(0.0000)	(0.5011)	(0.5471)
R^2	0.1447	0.9999	1.0000

We now add a factor, *sp*, to the CAPM and, for the time being, we do not reveal its informational content and construction method. The test assets, the sample period, and the market factor remain unchanged; the only change is the addition of the *sp* factor to the model. The results from this specification of the model are presented in the second column (CAPM + *sp* factor) of Table 1. The specification test now suggests that the model is correctly specified. Even more surprisingly, the R^2 jumps from 14.47% to 99.99%. The *sp* factor is highly statistically significant while the market factor becomes insignificant. An applied researcher who is interested in selecting a parsimonious statistical model could be willing to remove the market factor and reestimate the model with the *sp* factor only.

The results from this third specification are reported in the last column of Table 1. The results are striking. This one-factor model exhibits a perfect fit. Based on the specification test, the model appears to be correctly specified. Furthermore, the *sp* factor is highly statistically significant and is deemed to be priced. Given this exceptional performance of the model, we now ask: “What is this *sp* factor?” It turns out that this factor is generated as a standard normal random variable that is independent of returns. The results of this numerical exercise are completely spurious because the *sp* factor does not contribute, by construction, to pricing. In summary, a misspecified model with a spurious factor is concluded to be a correctly specified model with a spectacular fit and pricing ability. Even worse, the priced factors that are highly correlated with the test asset returns are driven out (become statistically insignificant) when a spurious factor is included in the model.

The results in Table 1 are based on one draw from the standard normal distribution. Our conclusions are qualitatively similar when the analysis in the table is based on the average of 100,000 replications. Starting from the CAPM + *sp* factor specification, the average $\mathcal{S}$ is 22.50 (p-value = 0.4806) and the average t-statistic for *vw* is −0.63 (p-value = 0.4096). As for the spurious *sp* factor, the average absolute value of the t-ratio is 4.76 (p-value = 0.0001). Finally, the average R^2 is 0.9946. Turning to the *sp* factor specification, the average $\mathcal{S}$ is 23.69 (p-value = 0.4738), the average absolute value of the t-ratio for the *sp* factor is 4.90 (p -value = 0.0000), and the average R^2 is 0.9948. The results are also largely unchanged when we augment the 25 Fama–French portfolio returns with additional test asset returns (for example, the 17 Fama–French industry portfolio returns) as recommended by Lewellen et al. (2010).

This type of behavior is not specific to artificial setups and also arises in well known empirical asset pricing models. To substantiate this claim, we consider three other popular asset pricing models. The first model is the three-factor model (FF3) of Fama and French (1993) with the market excess return (*vw*), the return difference between portfolios of stocks with small and large market capitalizations (*smb*), and the return difference between portfolios of stocks with high and low book-to-market ratios (*hml*) as risk factors. All of these risk factors are either portfolio excess returns or return spreads and exhibit a relatively high correlation with the 25 Fama–French portfolio returns. The other two models have traded and non-traded factors: the model (C-LAB) proposed by Jagannathan and Wang (1996), which, in addition to the market excess return, includes the growth rate in per capita labor income (*labor*) and the lagged default premium (*prem*, the yield spread between Baa- and Aaa-rated corporate bonds) as risk factors, and the model (CC-CAY) proposed by Lettau and Ludvigson (2001) with risk factors that include the growth rate in real per capita nondurable consumption (*cg*), the lagged consumption-aggregate wealth ratio (*cay*), and an interaction term between *cg* and *cay* ($cg \cdot cay$).

Table 2 reports results for these three models. For ease of comparison, we also present results for the CAPM. We include a version of the rank test of Cragg and Donald (1997) to determine whether the asset pricing models are properly identified (for details, see Section 3, and the widely used specification test based on the non-invariant, generalized least squares (GLS) estimator: the heteroskedasticity-consistent version of the cross-sectional regression (CSR) test of Shanken (1985) denoted by $\mathcal{Q}$.

Fig. 1 depicts the cross-sectional goodness of fit of the models by plotting average realized returns versus (fitted by ML) expected returns for each model.

The results of this empirical illustration confirm the evidence from the models with artificial data above. Models that contain factors that are only weakly correlated with the test asset returns (C-LAB and CC-CAY), as reflected in the non-rejection of the null hypothesis of a reduced rank in Table 2, exhibit an almost perfect fit. The specification test based on the ML estimator cannot reject the null of correct specification, which suggests that the models are well specified and one could proceed with constructing significance tests based on standard errors derived under the correct model specification.[1] These t-tests indicate that the proposed non-traded factors (default premium in C-LAB and consumption growth and the *cay* interaction term in CC-CAY, for example) are highly statistically significant. A benchmark model such as FF3 does not perform nearly as well according to these statistical measures. Similarly to CAPM in Table 1, the test for correct model specification suggests that FF3 is rejected by the data even with an R^2 of 73.37%.

Table 2
Test statistics for various asset pricing models.

The table reports test statistics for four asset pricing models: capital asset pricing model (CAPM), Fama and French three-factor model (FF3), conditional labor model (C-LAB), and conditional version of the consumption CAPM (CC-CAY). The test asset returns are the monthly returns on the value-weighted 25 Fama–French size and book-to-market ranked portfolios from January 1967 until December 2012. $\mathcal{CD}_B$ denotes the Cragg and Donald (1997) test for the null of a reduced rank in the beta-pricing setup. $\mathcal{Q}$ and $\mathcal{S}$ denote the Shanken (1985) tests of correct model specification based on the generalized least squares and maximum likelihood (ML) estimators, respectively. The rows for the different factors report the t-tests of statistical significance with standard errors computed under the assumption of correct model specification. R^2 denotes the squared correlation coefficient between the fitted expected returns and the average realized returns.

	CAPM	FF3	C-LAB	CC-CAY
Panel A: Rank and CSR tests				
$\mathcal{CD}_B$	465.03	321.18	20.87	14.10
(p-value)	(0.0000)	(0.0000)	(0.5290)	(0.8978)
$\mathcal{Q}$	71.96	55.61	69.68	71.77
(p-value)	(0.0000)	(0.0004)	(0.0000)	(0.0009)
Panel B: ML				
$\mathcal{S}$	68.79	51.05	20.87	13.85
(p-value)	(0.0000)	(0.0003)	(0.4672)	(0.8758)
vw	−3.65	−3.80	1.42	
smb		1.73		
hml		3.04		
labor			−3.14	
prem			−4.07	
cg				−2.23
cay				−0.77
$cg \cdot cay$				3.63
R^2	0.1447	0.7337	1.0000	0.9995

For comparison, Fig. 2 plots the average realized returns versus the fitted expected returns based on the non-invariant (GLS) estimator for each model. In sharp contrast with the results for the ML estimator in Fig. 1, the models that contain factors that are only weakly correlated with the test asset returns (C-LAB and CC-CAY) no longer exhibit a perfect fit. As a result, the non-invariant GLS estimator appears to be more robust to lack of identification and can detect model misspecification with a higher probability than its invariant counterpart.

In this paper, we show that, due to the combined effect of identification failure and model misspecification, the results for C-LAB and CC-CAY are likely to be spurious. While some warning signs of these problems are already present

[1] Gospodinov et al. (2014b, 2017a) show that the specification test, based on an invariant estimator, lacks power under the alternative of misspecified models when spurious factors are present. They demonstrate that the specification test has power equal to (or below) its size in reduced-rank asset pricing models.

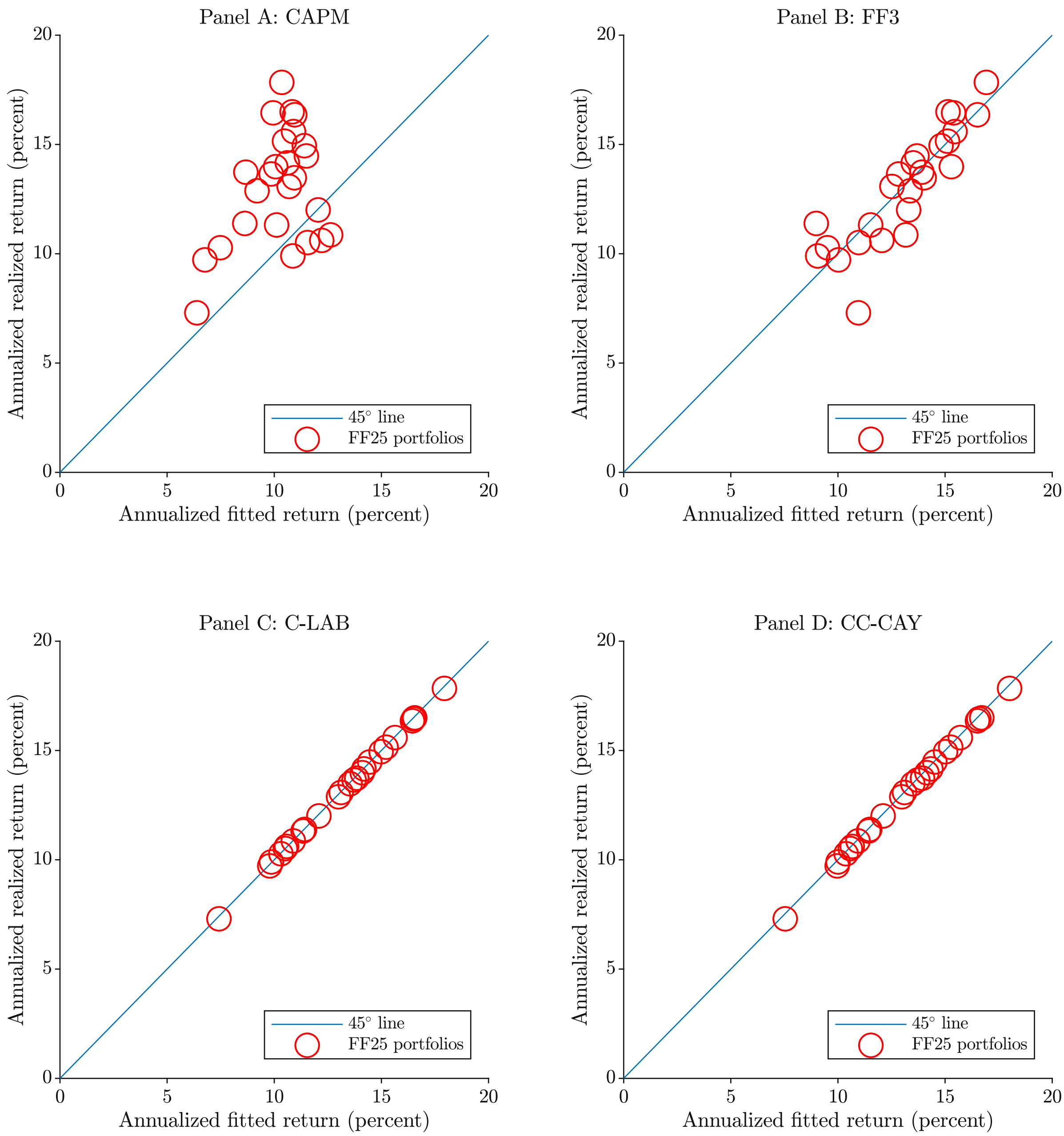

Fig. 1. Realized versus fitted [by maximum likelihood (ML)] returns: 25 Fama–French portfolios. The figure shows the average realized returns versus fitted expected returns (by ML) for each of the 25 Fama–French portfolios for capital asset pricing model (CAPM), Fama and French three-factor model (FF3), conditional labor model (C-LAB), and conditional version of the consumption CAPM (CC-CAY).

in Table 2, they are often ignored by applied researchers. For example, the rank tests provide strong evidence that C-LAB and CC-CAY are not identified, which violates the regularity conditions for consistency and asymptotic normality of the ML estimator. Furthermore, the $\mathcal{Q}$ test points to severe misspecification of all the considered asset pricing models.

Another interesting observation that emerges from these results is that the factors with low correlations with the returns tend to drive out the factors that are highly correlated with the returns. For example, the highly significant market factor in CAPM turns insignificant with the inclusion of labor growth and default premium in the C-LAB model. To further examine this point, we simulate data for the returns on the test assets and the market factor from a misspecified model that is calibrated to the CAPM as estimated in Table 1 (for more details on the simulation design, see Section 3). With a sample size of six hundred time series observations, the rejection rate (at the 5% significance level) of the t-test of whether the market factor is priced is 93.4%, and the mean R^2 is 18.6%. In sharp contrast, when a spurious factor (generated as an inde-

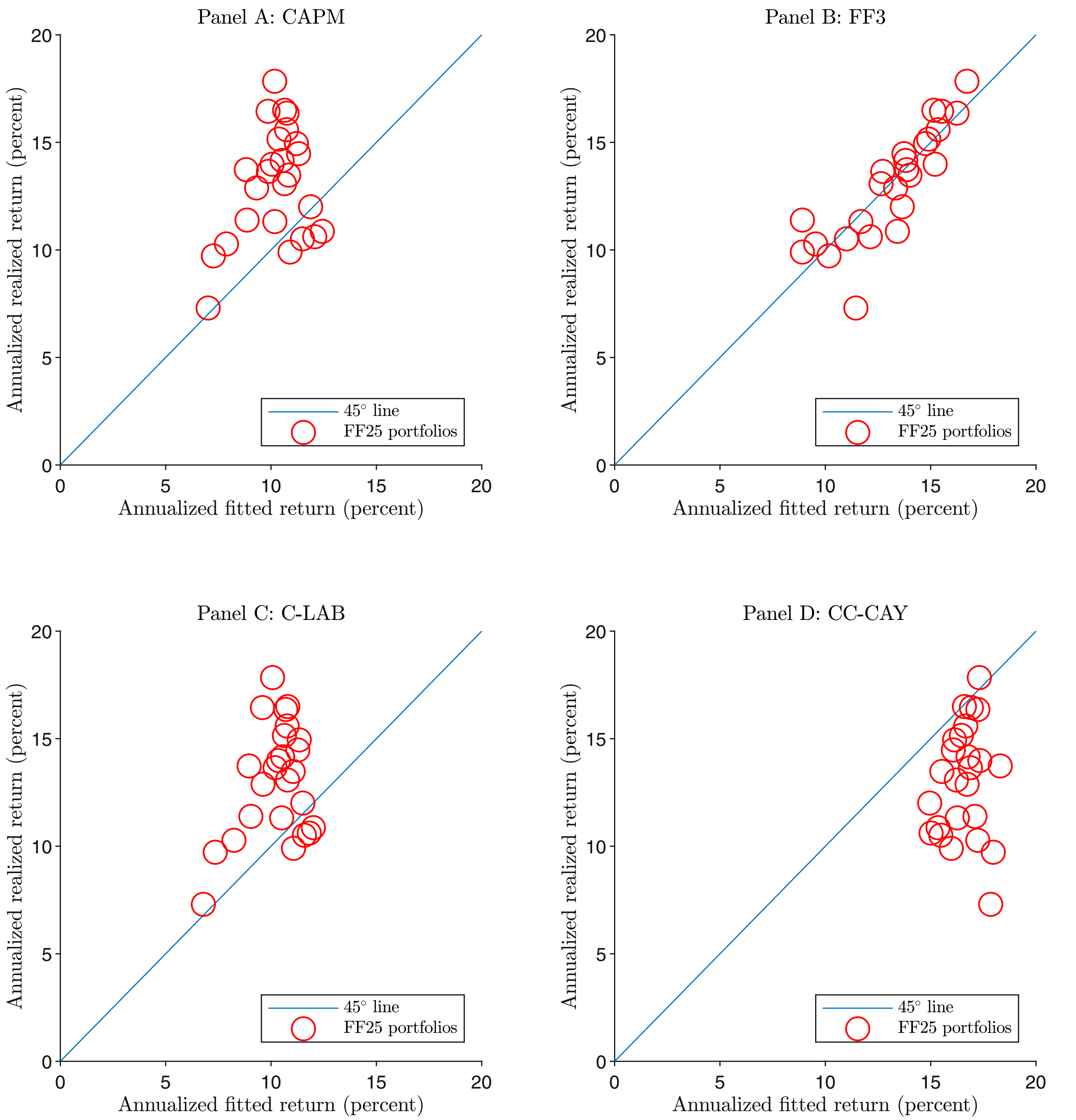

Fig. 2. Realized versus fitted [by generalized least squares (GLS)] returns: 25 Fama–French portfolios. The figure shows the average realized returns versus fitted expected returns (by GLS) for each of the 25 Fama–French portfolios for capital asset pricing model (CAPM), Fama and French three-factor model (FF3), conditional labor model (C-LAB), and conditional version of the consumption CAPM (CC-CAY).

pendent standard normal random variable) is added to the model, the rejection rate of the t-test for the market factor drops to 9.9% and the mean R^2 jumps to 99.7%. Strikingly, the rejection rate of the t-test for the spurious factor is 100%. This example clearly illustrates the severity of the problem and the perils for inference based on invariant estimators in unidentified models.[2]

In addition to identifying a serious problem with invariant estimators of asset pricing models, we characterize the limiting behavior of the ML estimator and the t-statistics under model misspecification and identification failure. We show that the ML estimator is inconsistent and the t-tests have a bimodal and heavy-tailed distribution. The estimates on the spurious factors exhibit an explo-

[2] An earlier version of the paper (Gospodinov et al., 2017b) contains the corresponding limiting results for the stochastic discount factor representation estimated by the continuously-updated generalized method of moments (CU-GMM) estimator. The results are qualitatively very similar to the ones for the ML estimator.

sive behavior that forces the goodness-of-fit statistic to approach one.

Some recent asset pricing studies have also expressed concerns about the appropriateness of the R^2 as a reliable goodness-of-fit measure. In models with excess returns, Burnside (2016) derives a similar behavior of the goodness-of-fit statistic for non-invariant generalized method of moments (GMM) estimators. This result, however, is normalization- and setup-specific, and alternative normalizations or models based on gross returns render the non-invariant estimators immune to the perfect fit problem. Furthermore, Kleibergen and Zhan (2015) show that a sizable unexplained factor structure (generated by a low correlation between the observed proxy factors and the true unobserved factors) in a two-pass CSR framework can also produce spuriously large values of the ordinary least squares (OLS) R^2 coefficient. Their results complement the findings of Lewellen et al. (2010), who criticize the use of the OLS R^2 coefficient by showing that it provides an overly positive assessment of the performance of the asset pricing model. Despite the suggestive nature of these findings, model evaluation tests based on non-invariant estimators, which are the focus of the analysis in these studies, tend to be more robust to lack of identification. For invariant estimators in under-identified asset pricing models, the spurious perfect fit is pervasive regardless of the model structure (gross or excess returns), estimation framework, and chosen normalization.

The rest of the paper is organized as follows. Section 2 studies the limiting behavior of the parameter estimates, t-statistics, and goodness-of-fit measures in the beta-pricing setup. Section 3 reports Monte Carlo simulation results. Section 4 presents our empirical findings. Section 5 summarizes our main conclusions and provides some practical recommendations. The technical proofs are in the Appendix.

2. Beta-pricing model and maximum likelihood

This section presents the main theoretical results for ML-based inference in a misspecified beta-pricing model with a spurious factor.

2.1. Model and notation

Let f_t be a $(K-1)$-vector of systematic risk factors and R_t denote the returns on N $(N>K)$ test assets. We define $Y_t=[f_t',\ R_t']'$ and its population mean and covariance matrix as

$$\mu=E[Y_t]\equiv\begin{bmatrix}\mu_f\\ \mu_R\end{bmatrix} \tag{1}$$

and

$$V=\mathrm{Var}[Y_t]\equiv\begin{bmatrix}V_f & V_{fR}\\ V_{Rf} & V_R\end{bmatrix}, \tag{2}$$

where V is assumed to be a positive-definite matrix. Furthermore, let $\gamma=[\gamma_0,\ \gamma_1']'$ be a K-vector of zero-beta rate and risk premium parameters associated with the factors. When the asset pricing model is correctly specified and well identified, there exists a unique $\gamma^*=[\gamma_0^*,\ \gamma_1^{*\prime}]'$ such that

$$\mu_R=1_N\gamma_0^*+\beta\gamma_1^*, \tag{3}$$

where $\beta=[\beta_1,\dots,\beta_{K-1}]=V_{Rf}V_f^{-1}$ is an $N\times(K-1)$ matrix of the betas of the N assets. Also, define

$$\alpha=\mu_R-\beta\mu_f, \tag{4}$$

and $\Sigma=V_R-V_{Rf}V_f^{-1}V_{fR}$. Combining Eqs. (3) and (4), we arrive at the restriction

$$\alpha=1_N\gamma_0^*+\beta(\gamma_1^*-\mu_f). \tag{5}$$

The primary focus of our analysis lies in characterizing the limiting behavior of the t-tests for statistical significance of the γ_1 estimates and the goodness-of-fit statistic defined as the squared correlation between the realized and model-implied expected returns.[3] The asymptotic approximations of these statistics are crucially affected by the rank of the matrix $G\equiv[1_N,\ \mathcal{B}]$, where $\mathcal{B}=[\alpha,\ \beta]$. The reduced rank of G can result either from validity of the asset pricing model restriction $\alpha=1_N\gamma_0+\beta(\gamma_1-\mu_f)$ or from a rank deficiency in the matrix $B=[1_N,\ \beta]$, which is caused by the presence of spurious factors.

2.2. ML-based inference and main results

We consider the ML estimation of the beta-pricing model that imposes the joint normality assumption on Y_t. The joint normality of Y_t is assumed for convenience, and the results continue to hold under weaker conditions. For example, this assumption could be relaxed by assuming conditional normality on the regression errors or adopting a quasi-maximum likelihood framework as in White (1994). The main reason for making this assumption here is to interpret $\hat{\mu}_f$ in $\gamma_1-\hat{\mu}_f$ as the ML estimator of μ_f. Otherwise, we need to replace $\hat{\mu}_f$ below with its appropriate ML estimator. The CU-GMM results in an earlier version of the paper (see Gospodinov et al., 2017b) do not hinge on any distributional assumptions. This comes at the cost of losing the closed-form solution for the estimator and some of the sharpness of the results. Thus, for expositional clarity, the focus of this paper is on the ML estimator under the normality assumption.

Then, the ML estimator of γ^* is defined as (see Shanken, 1992; Shanken and Zhou, 2007)

$$\hat{\gamma}^{ML}=\underset{\gamma}{\operatorname{argmin}}\ \frac{[\hat{\alpha}-1_N\gamma_0-\hat{\beta}(\gamma_1-\hat{\mu}_f)]'\hat{\Sigma}^{-1}[\hat{\alpha}-1_N\gamma_0-\hat{\beta}(\gamma_1-\hat{\mu}_f)]}{1+\gamma_1'\hat{V}_f^{-1}\gamma_1}, \tag{6}$$

where $\hat{\alpha}$, $\hat{\beta}$, $\hat{\mu}_f$, $\hat{V}_f$, and $\hat{\Sigma}$ are the sample estimators of α, β, μ_f, V_f, and Σ, respectively. By rewriting Eq. (6) as $\hat{\gamma}^{ML}=\underset{\gamma}{\operatorname{argmin}}(\hat{\mu}_R-\hat{B}\gamma)'[(1+\gamma_1'\hat{V}_f^{-1}\gamma_1)\hat{\Sigma}]^{-1}(\hat{\mu}_R-\hat{B}\gamma)$, where $\hat{B}=[1_N,\ \hat{\beta}]$, the ML estimator becomes equivalent

[3] In a multi-factor model, acceptance or rejection of $\gamma_{1,i}=0$ does not indicate whether the ith factor makes an incremental contribution to the model's overall explanatory power, given the presence of the other factors. See Kan et al. (2013) for a discussion of this subtle point.

to the asymptotic least squares estimator of Gourieroux et al. (1985) and Kodde et al. (1990).

The test for correct model specification of Shanken (1985) is given by

$$\mathcal{S} = T \min_{\gamma} \frac{(\hat{\alpha}-1_N\gamma_0-\hat{\beta}(\gamma_1-\hat{\mu}_f))'\hat{\Sigma}^{-1}(\hat{\alpha}-1_N\gamma_0-\hat{\beta}(\gamma_1-\hat{\mu}_f))}{1+\gamma_1'\hat{V}_f^{-1}\gamma_1} \tag{7}$$

and is asymptotically distributed as $\mathcal{S} \xrightarrow{d} \chi^2_{N-K}$ under the null H_0: $\alpha = 1_N\gamma_0 + \beta(\gamma_1 - \mu_f)$ when the model is identified.

Due to the special structure of this objective function, the ML estimator of γ^* can be obtained explicitly as the solution to an eigenvector problem.[4] Letting $v = [-\gamma_0,\ 1,\ -(\gamma_1 - \hat{\mu}_f)']'$ and $\hat{G} = [1_N,\ \hat{\alpha},\ \hat{\beta}]$, and noting that $\hat{\alpha} - 1_N\gamma_0 - \hat{\beta}(\gamma_1 - \hat{\mu}_f) = \hat{G}v$, we can write the objective function of the ML estimator as

$$\min_{v} \frac{v'\hat{G}'\hat{\Sigma}^{-1}\hat{G}v}{v'A(X'X/T)^{-1}A'v}, \tag{8}$$

where $A = [0_K,\ I_K]'$ and X is a $T \times K$ matrix with a typical row $x_t' = [1,\ f_t']$. Let $\hat{v}$ be the eigenvector associated with the largest eigenvalue of

$$\hat{\Omega} = (\hat{G}'\hat{\Sigma}^{-1}\hat{G})^{-1}[A(X'X/T)^{-1}A']. \tag{9}$$

Then, the ML estimator of γ^* can be constructed as

$$\hat{\gamma}_0^{ML} = -\frac{\hat{v}_1}{\hat{v}_2} \tag{10}$$

and

$$\hat{\gamma}_{1,i}^{ML} = \hat{\mu}_{f,i} - \frac{\hat{v}_{i+2}}{\hat{v}_2}, \quad i = 1, \ldots, K-1. \tag{11}$$

When the model is correctly specified and B is of full column rank, we have that $Gv^* = 0_N$ for $v^* = [-\gamma_0^*,\ 1,\ -(\gamma_1^* - \hat{\mu}_f)']'$ and

$$\sqrt{T}\begin{bmatrix} \hat{\gamma}_0^{ML} - \gamma_0^* \\ \hat{\gamma}_1^{ML} - \gamma_1^* \end{bmatrix} \xrightarrow{d} \mathcal{N}\left(0_K,\ (1+\gamma_1^{*\prime}V_f^{-1}\gamma_1^*)(B'\Sigma^{-1}B)^{-1} + \begin{bmatrix} 0 & 0_{K-1}' \\ 0_{K-1} & V_f \end{bmatrix}\right). \tag{12}$$

As a result, the t-statistics for statistical significance of $\hat{\gamma}_0^{ML}$ and $\hat{\gamma}_{1,i}^{ML}$ $(i = 1, \ldots, K-1)$ are constructed as

$$t(\hat{\gamma}_0^{ML}) = \frac{\sqrt{T}\hat{\gamma}_0^{ML}}{s(\hat{\gamma}_0^{ML})} \tag{13}$$

and

$$t(\hat{\gamma}_{1,i}^{ML}) = \frac{\sqrt{T}\hat{\gamma}_{1,i}^{ML}}{s(\hat{\gamma}_{1,i}^{ML})}, \tag{14}$$

where $s(\hat{\gamma}_0^{ML}),\ s(\hat{\gamma}_{1,1}^{ML}), \ldots, s(\hat{\gamma}_{1,K-1}^{ML})$ denote the square roots of the diagonal elements of

$$V_{\hat{\gamma}} = (1+\hat{\gamma}_1^{ML\prime}\hat{V}_f^{-1}\hat{\gamma}_1^{ML})(\hat{B}'\hat{\Sigma}^{-1}\hat{B})^{-1} + \hat{V}_x, \tag{15}$$

where $\hat{B} = [1_N,\ \hat{\beta}]$ and $\hat{V}_x = \begin{bmatrix} 0 & 0_{K-1}' \\ 0_{K-1} & \hat{V}_f \end{bmatrix}$. Using the ML estimates, $\hat{\gamma}_0^{ML}$ and $\hat{\gamma}_1^{ML}$, the ML estimate of β, $\hat{\beta}^{ML}$, and the fitted expected returns on the test assets, $\hat{\mu}_R^{ML}$, are obtained as

$$\hat{\beta}^{ML} = \hat{\beta} + \frac{[\hat{\alpha} - 1_N\hat{\gamma}_0^{ML} - \hat{\beta}(\hat{\gamma}_1^{ML} - \hat{\mu}_f)]\hat{\gamma}_1^{ML\prime}\hat{V}_f^{-1}}{1+\hat{\gamma}_1^{ML\prime}\hat{V}_f^{-1}\hat{\gamma}_1^{ML}} \tag{16}$$

and

$$\hat{\mu}_R^{ML} = 1_N\hat{\gamma}_0^{ML} + \hat{\beta}^{ML}\hat{\gamma}_1^{ML}. \tag{17}$$

An equivalent way to obtain $\hat{\beta}^{ML}$ is by running an OLS regression of $R_t - 1_N\hat{\gamma}_0^{ML}$ on $f_t + \hat{\gamma}_1^{ML} - \hat{\mu}_f$.[5]

Because the empirical evidence strongly suggests that linear asset pricing models are misspecified (as emphasized in our empirical application and many papers in the literature), in the following analysis we present results only for the misspecified model case. The analytical and simulation results for the correctly specified model case are available upon request. We briefly summarize some of these results in Sections 2.3 and 3.

The following theorem and Auxiliary Lemma 1 in the Appendix characterize the limiting behavior of the ML estimates $\hat{\gamma}^{ML}$, the t-statistics $t(\hat{\gamma}_0^{ML})$ and $t(\hat{\gamma}_{1,i}^{ML})$ $(i = 1, \ldots, K-1)$, and the R^2 statistic $R^2 = \text{Corr}(\hat{\mu}_R^{ML}, \hat{\mu}_R)^2$ in misspecified models that contain a spurious factor.

Without loss of generality, we assume that the spurious factor is the last element of the vector f_t with $\beta_{K-1} = 0_N$ and is independent of the test asset returns and the other factors. Our analysis can be easily modified to deal with the case in which the betas of the factors are constant across assets instead of being equal to zero and the case of a model with two (or more) factors that are noisy versions of the same underlying factor. In these scenarios, B is also of reduced rank. Let $\bar{Z}_i$, $i = 0, \ldots, K-2$, denote a bounded random variable defined in the Appendix.

Theorem 1. Assume that Y_t is independent and identically normally distributed (i.i.d.). Suppose that the model is misspecified (that is, $\mu_R \neq B\gamma$ for all γ) and contains a spurious factor [that is, $\text{rank}(B) = K-1$]. Then, as $T \to \infty$,

(a) *(i) $t(\hat{\gamma}_0^{ML}) \xrightarrow{d} \bar{Z}_0$; (ii) $t(\hat{\gamma}_{1,i}^{ML}) \xrightarrow{d} \bar{Z}_i$ for $i = 1, \ldots, K-2$; (iii) $t^2(\hat{\gamma}_{1,K-1}^{ML}) \xrightarrow{d} \chi^2_{N-K+1}$; and*

(b) $R^2 \xrightarrow{p} 1$.

Proof. See the Appendix. □

2.3. Discussion of results and intuition

Theorem 1 establishes the limiting behavior of the t-tests and R^2 statistic in misspecified models with identification failure. The t-tests for the useful factors converge to bounded random variables and, hence, are inconsistent. As our simulations illustrate, the tests $t(\hat{\gamma}_{1,i}^{ML})$ for $i = 1, \ldots, K-2$ tend to exhibit power that is close to their size. In contrast, the t-test for the spurious factor will over-reject substantially (with the probability of

[4] See also Zhou (1995) and Bekker et al. (1996) for expressing the beta-pricing model as a reduced-rank regression whose estimated parameters are obtained as an eigenvalue problem.

[5] We are grateful to an anonymous referee for pointing this out.

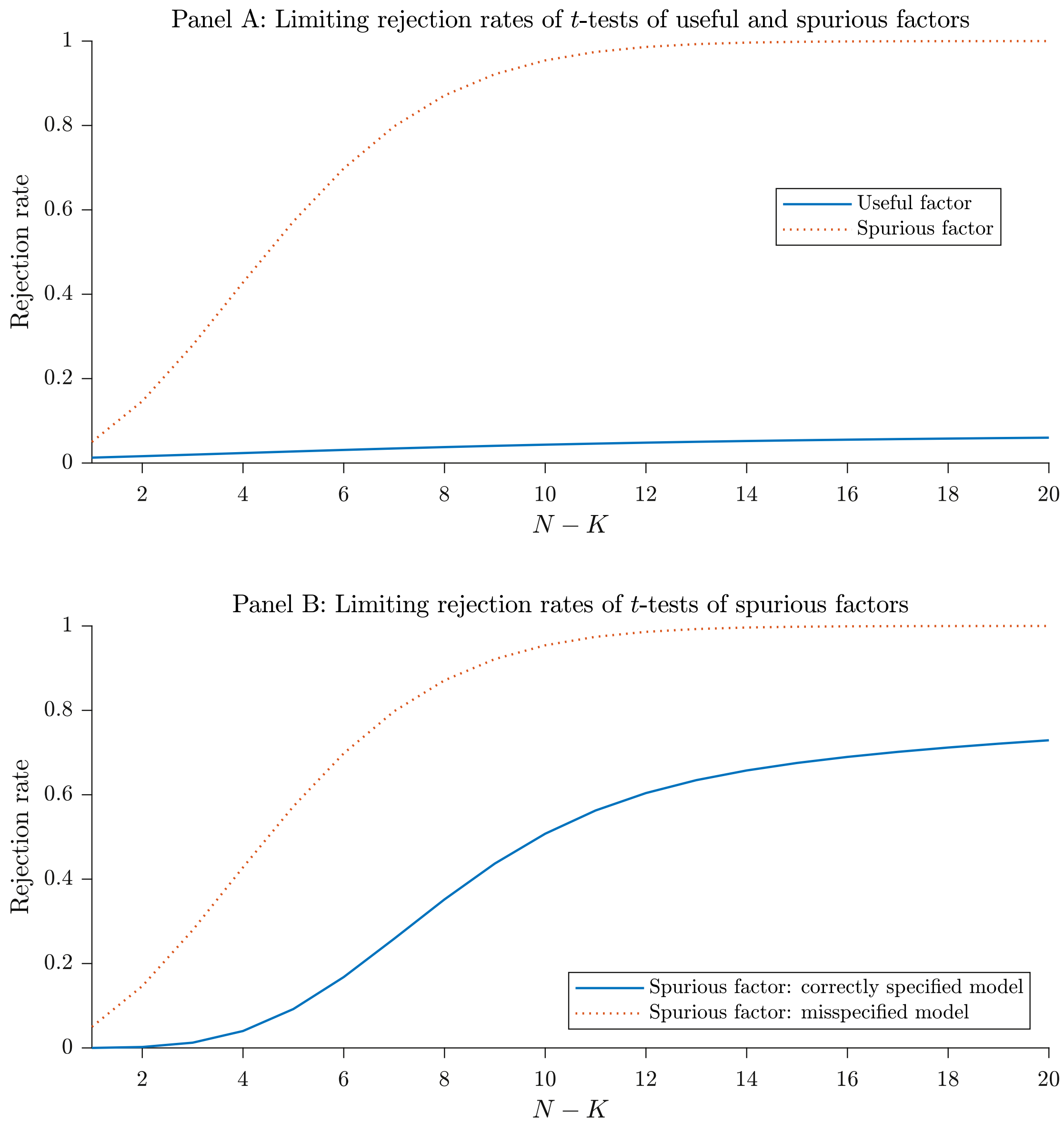

Fig. 3. Limiting rejection rates of t-tests of statistical significance. Panel A plots the limiting rejection rates under misspecified models of $t(\hat{\gamma}_{1,i}^{ML})$ and $t(\hat{\gamma}_{1,K-1}^{ML})$ as functions of $N-K$ when one uses the standard normal critical values. Panel B plots the limiting rejection rates under correctly specified and misspecified models of $t(\hat{\gamma}_{1,K-1}^{ML})$ as functions of $N-K$ when one uses the standard normal critical values.

rejection rapidly approaching one as N increases) when $\mathcal{N}(0,1)$ critical values are used. Furthermore, part (b) of Theorem 1 shows that the R^2 of a misspecified model that contains a spurious factor approaches one. This leads to a completely spurious inference procedure as the spurious factors do not contribute to the pricing performance of the model and yet the R^2 would indicate that the model perfectly explains the cross-sectional variation in the expected returns on the test assets.

To visualize the limiting behavior of the t-statistics in part (a), Panel A of Fig. 3 plots the limiting rejection rates of $t(\hat{\gamma}_{1,i}^{ML})$ and $t(\hat{\gamma}_{1,K-1}^{ML})$ as functions of $N-K$ for a misspecified model with a spurious factor when one uses the standard normal critical values. The sample quantities that enter the computation of the t-statistics for the useful factor are calibrated to the CAPM. Fig. 3 confirms that $t(\hat{\gamma}_{1,i}^{ML})$ is inconsistent as its power does not go to one asymptotically. The over-rejection of $t(\hat{\gamma}_{1,K-1}^{ML})$ increases with $N-K$, and the probability of rejecting $H_0 : \gamma_{1,K-1} = 0$ for the spurious factor is effectively one when $N-K \geq 15$.

When the model is correctly specified, the limiting distribution of the t-statistics for the useful factors is still nonstandard but, unlike the misspecified model case, useful factors that are priced are maintained in the model with probability approaching one. Although less pronounced than in the misspecified model case, using

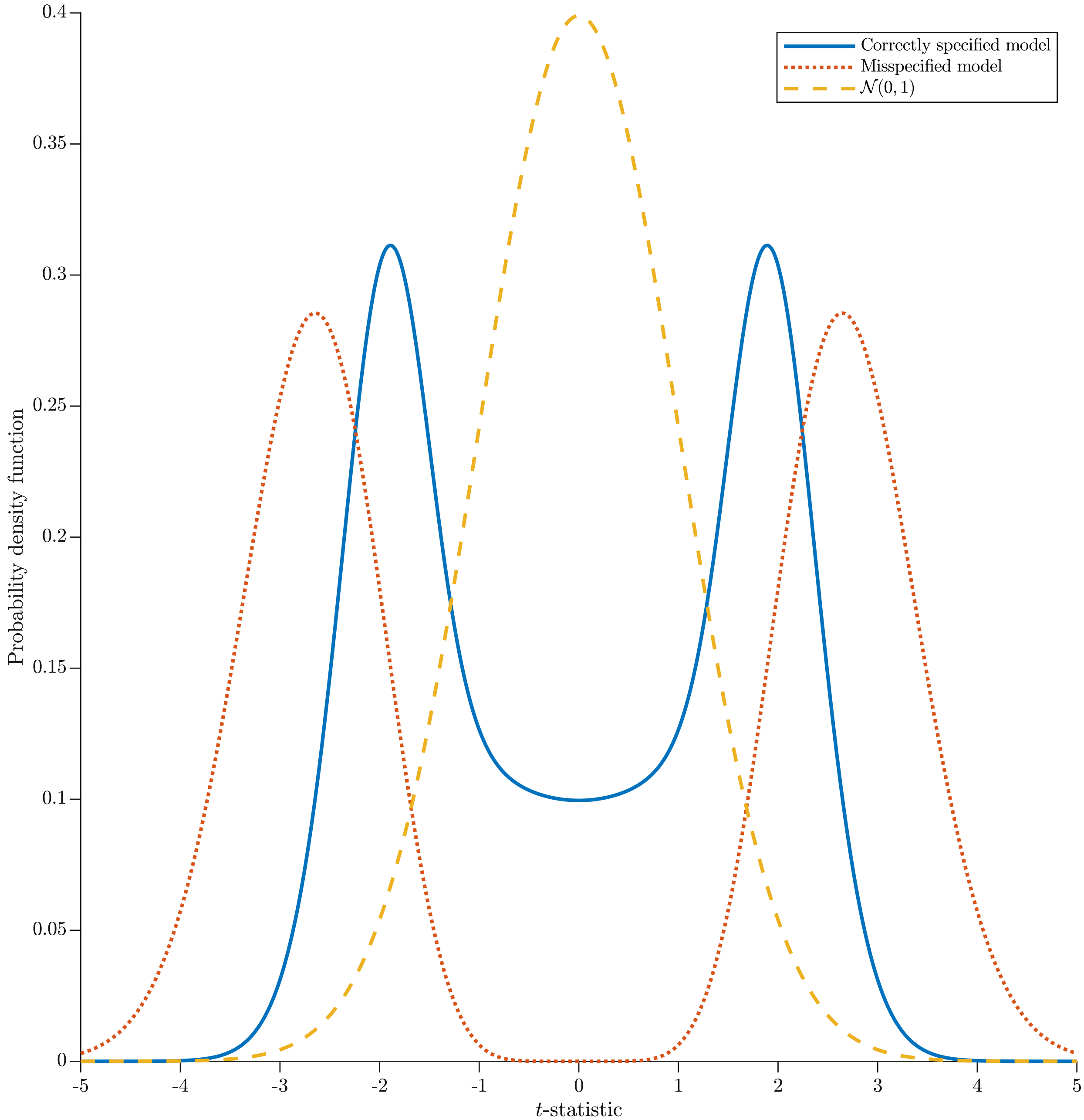

Fig. 4. Limiting distributions of $t(\hat{\gamma}_{1,K-1}^{ML})$ under correctly specified and misspecified models. The figure plots the limiting densities of $t(\hat{\gamma}_{1,K-1}^{ML})$ for correctly specified and misspecified models that contain a spurious factor (for $N-K=7$), along with the standard normal density.

$\mathcal{N}(0,1)$ critical values will still lead to substantial over-rejections of $H_0: \gamma_{1,K-1}=0$ for the spurious factor. This is revealed by Panel B of Fig. 3.

The reason for the over-rejection for the parameter on the spurious factor is clearly illustrated in Fig. 4, which plots the limiting probability density functions of $t(\hat{\gamma}_{1,K-1}^{ML})$ under correctly specified and misspecified models ($N-K=7$), along with the standard normal density. Given the bi-modal shape and large variance of the probability density function of the limiting distribution of $t(\hat{\gamma}_{1,K-1}^{ML})$ under correctly specified models (which arises from the model's under-identification), using $\mathcal{N}(0,1)$ critical values will lead to an over-rejection of the hypothesis that the spurious factor is not priced. This over-rejection is further exacerbated by model misspecification, as illustrated by the outward shift of the probability density function. Hence, with lack of identification, misleading inference arises in correctly specified models as well as in misspecified models, although the inference problems are more pronounced in the latter case.

Assuming that $\mu_f = 0_{K-1}$, some further intuition behind these results can be gained from considering the simpler case of a model without γ_0.[6] In this case, the eigenvector associated with the largest eigenvalue of the matrix in Eq. (9) is identical to the eigenvector associated with the smallest root of the characteristic polynomial

$$\left|\xi(X'X) - \hat{\mathcal{B}}'\hat{\Sigma}^{-1}\hat{\mathcal{B}}\right| = 0. \tag{18}$$

Under correct model specification, $\alpha = \beta\gamma_1$, and absence of spurious factors, $\hat{\mathcal{B}}$ converges to the reduced-rank matrix $\mathcal{B}_0 = [\beta\gamma_1^*,\ \beta]$ as the sample size increases, and the smallest root of the above characteristic polynomial converges to zero with its corresponding eigenvector $\hat{v} = [\hat{v}_1, \ldots, \hat{v}_K]'$ being proportional to $[1,\ -\gamma_1^{*\prime}]'$. Then, $\hat{\gamma}_1^{ML} = -[\hat{v}_2, \ldots, \hat{v}_K]'/\hat{v}_1$ is a consistent estimator of γ_1^*, and the usual limiting characterization applies.

Under the conditions of Theorem 1, a misspecified model with one spurious factor (ordered last), matrix $\mathcal{B}$ takes a different form, $\mathcal{B} = [\alpha,\ \beta_1, \ldots, \beta_{K-2},\ 0_N]$, and it is still of reduced column rank $K-1$. The rank deficiency here is not caused by correct model specification but by the reduced rank of the β matrix. An immediate consequence is that the specification test $\mathcal{S}$ has asymptotic power that is equal to its size, and a researcher who ignores this rank failure in the β matrix will likely conclude that the model is correctly specified even when the degree of misspecification is arbitrarily large (see Gospodinov et al., 2014b; Gospodinov et al., 2017a). Furthermore, the limiting properties of the ML estimator, significance tests, and goodness-of-fit statistic are highly nonstandard. The smallest root of the characteristic polynomial in Eq. (18) again approaches zero, but its corresponding eigenvector $\hat{v}$ is now proportional to $[0_{K-1}',\ 1]'$ because $[\alpha,\ \beta_1, \ldots, \beta_{K-2},\ 0_N][0_{K-1}',\ 1]' = 0_N$. Then, $\sqrt{T}[\hat{v}_1, \ldots, \hat{v}_{K-1},\ \hat{v}_K - 1]' \stackrel{d}{\to} z$, where z is a mean-zero normally distributed random vector. Hence, $\hat{\gamma}_{1,i}^{ML} \stackrel{d}{\to} -z_{i+1}/z_1$ for $i = 1, \ldots, K-2$ and $T^{-1/2}\hat{\gamma}_{1,K-1}^{ML} \stackrel{d}{\to} 1/z_1$.

These results suggest that when a spurious factor is present, the estimates for the useful factors are inconsistent and converge to ratios of normal random variables, the estimate for the spurious factor, $\hat{\gamma}_{1,K-1}^{ML}$, diverges at rate root T, and the standardized estimator converges to the reciprocal of a normal random variable (see also Kan and Zhang, 1999, and Kleibergen, 2009, for similar results for non-invariant two-pass CSR estimators). In contrast, when the model is correctly specified, $\hat{\gamma}_0^{ML}$ and $\hat{\gamma}_{1,i}^{ML}$ ($i = 1, \ldots, K-2$) for the useful factors are consistent estimators (although with a non-normal asymptotic limit) of γ_0^* and $\gamma_{1,i}^*$, respectively, and $\hat{\gamma}_{1,K-1}^{ML}$ for the spurious factor is inconsistent but has a limiting Cauchy distribution. These nonstandard properties of the ML estimator give rise to the nonstandard asymptotic distribution of the t-statistics in part (a) of Theorem 1.

The limiting behavior of R^2, which measures the squared correlation between $\hat{\mu}_R^{ML}$ and $\hat{\mu}_R$, is also directly driven by $\hat{\gamma}_{1,i}^{ML} = O_p(1)$ (for $i = 1, \ldots, K-2$) and the divergent behavior of $\hat{\gamma}_{1,K-1}^{ML} = O_p(T^{\frac{1}{2}})$. Because $\hat{\mu}_R^{ML} = \hat{\beta}^{ML}\hat{\gamma}_1^{ML} = \hat{\mu}_R - \frac{\hat{\mu}_R - \hat{\beta}\hat{\gamma}_1^{ML}}{1+\hat{\gamma}_1^{ML\prime}\hat{V}_f^{-1}\hat{\gamma}_1^{ML}} = \hat{\mu}_R + o_p(1)$ from Eq. (16) and the limiting properties of $\hat{\gamma}_1^{ML}$, it immediately follows that the R^2 converges to one in large samples. These limiting characterizations, albeit at the expense of some technicalities, provide guidance and a conceptual framework for explaining the seemingly abnormal empirical results presented in the introduction and Sections 3 and 4.

Qualitatively similar results extend to other invariant estimators. An earlier version of the paper (Gospodinov et al., 2017b) contained results for the continuously updated generalized method of moments estimator. Other popular likelihood-based estimators, such as generalized empirical likelihood and Bayesian estimators, are also not immune to this problem as they exhibit heightened sensitivity to departures from full identification and correct model specification.

[6] We would like to thank an anonymous referee for suggesting this.

3. Simulation experiment

In this section, we undertake a Monte Carlo simulation experiment to study the empirical rejection rates of the t-tests for the ML estimator as well as the finite-sample distribution of the goodness-of-fit measure. We consider three linear models: (1) a model with a constant term and a useful factor, (2) a model with a constant term and a spurious factor, and (3) a model with a constant term, a useful, and a spurious factor. All three models are misspecified.

The returns on the test assets and the useful factor are drawn from a multivariate normal distribution. In all simulation designs, the covariance matrix of the simulated test asset returns is set equal to the sample covariance matrix from the January 1967 to December 2012 sample of monthly returns on the 25 Fama–French size and book-to-market ranked portfolios (from Kenneth French's website at http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html). The means of the simulated returns are set equal to the sample means of the actual returns, and they are not exactly linear in the chosen betas for the useful factor. As a result, the models are misspecified in all three cases. The mean and variance of the simulated useful factor are calibrated to the sample mean and variance of the value-weighted market excess return. The covariances between the useful factor and the returns are chosen based on the sample covariances estimated from the data. The spurious factor is generated as a standard normal random variable, which is independent of the returns and the useful factor. The time series sample size is $T = 200$, 600, and 1000, and all results are based on 100,000 Monte Carlo replications, with the exception of the results in Fig. 5, which are based on 500,000 Monte Carlo simulations to obtain a smoother plot of the cumulative distribution function of the R^2. We also report the limiting rejection probabilities (denoted by $T = \infty$) for the t-tests based on our asymptotic results in Section 2.

A popular way to assess the performance of the model is to compute the squared correlation between the fitted expected returns of the model and the average realized returns. The empirical distribution of this R^2 is reported in Fig. 5. As our theoretical analysis suggests, the empirical distribution of the R^2 in misspecified models with a spurious factor collapses to one as the sample size gets

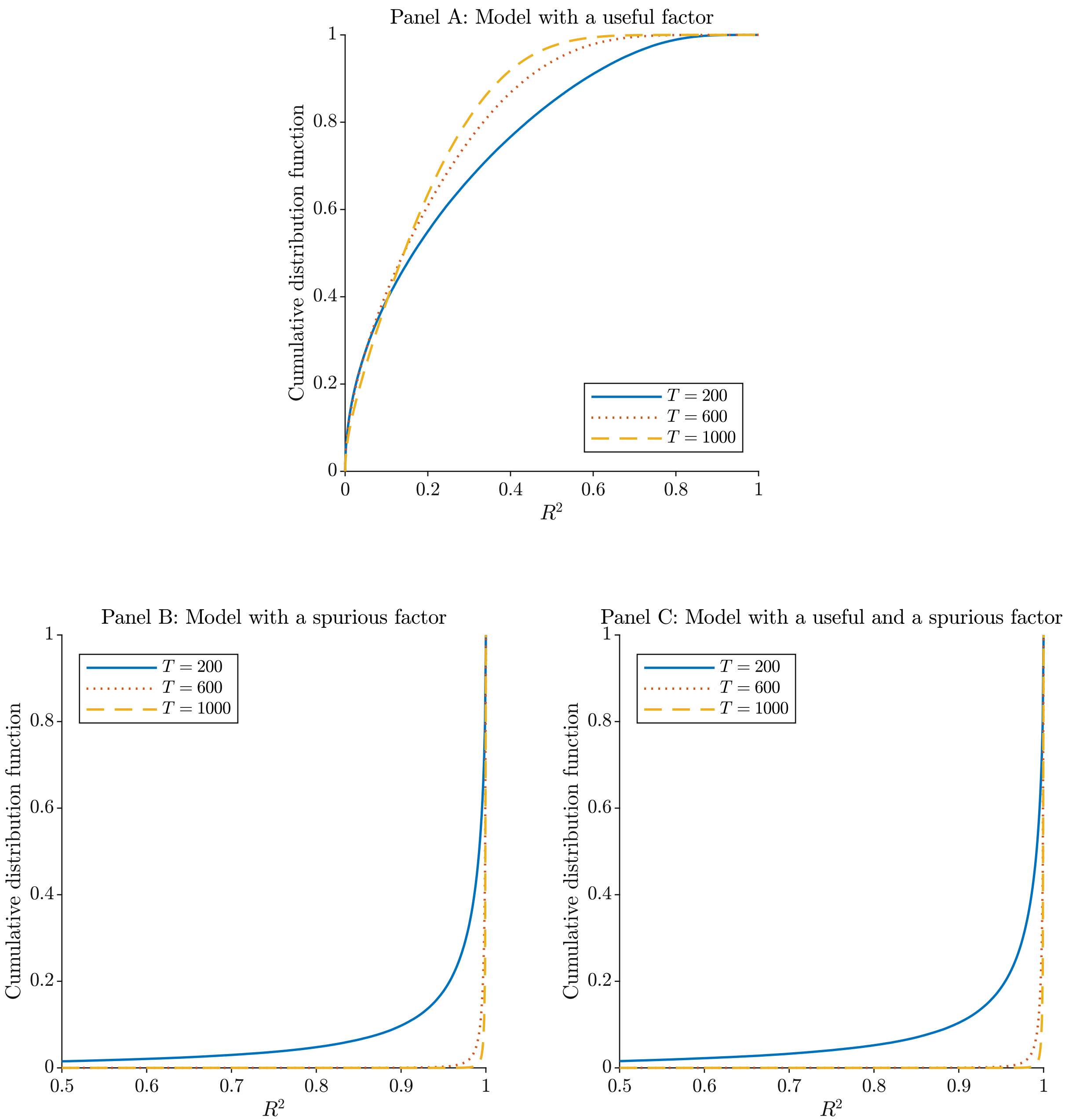

Fig. 5. Cumulative distribution function of the R^2. The figure plots the cumulative distribution function of the R^2 computed as the squared correlation between the realized and fitted expected returns based on the maximum likelihood estimator.

large.[7] As a result, this measure will indicate a perfect fit for models that include a factor that is independent of the returns on the test assets. These spurious results should serve as a warning signal in applied work in which many macroeconomic factors are only weakly correlated with the returns on the test assets.

Table 3 presents the rejection probabilities of the t-tests of $H_0 : \gamma_{1,i} = 0$ (tests of parameter significance) for the useful and the spurious factors in models (1)–(3). The t-statistics are computed under the assumption that the model is correctly specified and are compared against the critical values from the standard normal distribution, as is

[7] In the presence of spurious factors, the empirical distributions of the R^2s in correctly specified and misspecified models are very different. For example, when the model is correctly specified with a spurious factor and $T = 600$, the 10%, 50%, and 90% percentiles of the R^2 distribution are 0.049, 0.686 and 0.989, respectively, and the corresponding ones for the misspecified model case are 0.993, 0.999 and 1.000. This holds true even though we expect the R^2 for correctly specified models to be higher than the R^2 for misspecified models.

Table 3
Rejection rates of t-tests.

The table presents the rejection rates of t-tests of statistical significance under misspecified models for the maximum likelihood estimator. The null hypothesis is that the parameter of interest is equal to zero. The results are reported for different levels of significance (10%, 5%, and 1%) and for different values of the number of time series observations (T). The t-statistics with standard errors computed under the assumption of correct model specification are compared with the critical values from a standard normal distribution. The rejection rates for the limiting case ($T=\infty$) in Panels B and C are based on the asymptotic distributions in part (a) of Theorem 1.

	Useful			Spurious		
T	10%	5%	1%	10%	5%	1%
Panel A: Model with a useful factor only						
200	0.698	0.616	0.442	–	–	–
600	0.959	0.934	0.848	–	–	–
1000	0.996	0.992	0.971	–	–	–
∞	1.000	1.000	1.000	–	–	–
Panel B: Model with a spurious factor only						
200	–	–	–	0.996	0.996	0.993
600	–	–	–	1.000	1.000	1.000
1000	–	–	–	1.000	1.000	1.000
∞	–	–	–	1.000	1.000	1.000
Panel C: Model with a useful and a spurious factor						
200	0.271	0.185	0.075	0.992	0.991	0.986
600	0.171	0.099	0.025	1.000	1.000	1.000
1000	0.152	0.083	0.019	1.000	1.000	1.000
∞	0.124	0.062	0.011	1.000	1.000	1.000

commonly done in the literature. Table 3 reveals that for models with a spurious factor, the t-tests will give rise to spurious results, suggesting that these completely irrelevant factors are priced. Moreover, the spurious factor (which, by construction, does not contribute to the pricing performance of the model) drives out the useful factor and leads to the grossly misleading conclusion to keep the spurious factor and drop the useful factor from the model (see Panel C of Table 3).

The spuriously high R^2 values and the perils of relying on the traditional t-tests of parameter significance in unidentified models suggest that the decision regarding the model specification should be augmented with additional diagnostics. One approach to restoring the validity of the standard inference is based on the following model reduction procedure.[8] To assess the degree of identification of the model, the matrix $B=[1_N,\ \beta]$ is subjected to a rank test. To this end, we employ a version of the rank test of Cragg and Donald (1997) denoted by $\mathcal{CD}_B(L)$, where $1\le L\le K-1$ is the reduced rank under the null. Under our assumptions, we have $\sqrt{T}\text{vec}(\hat\beta-\beta)\xrightarrow{d}\mathcal{N}(0_{N(K-1)},V_f^{-1}\otimes\Sigma)$. Because the covariance matrix of $\hat\beta$ is in Kronecker form, the rank test of $H_0:\text{rank}(B)=L$ reduces to a solution to an eigenvalue problem and takes the form

$$\mathcal{CD}_B(L)=T(\lambda_{L+1}+\cdots+\lambda_K), \tag{19}$$

where $\lambda_{L+1},\dots,\lambda_K$ are the $K-L$ smallest generalized eigenvalues of the square matrices $\hat B'\hat\Sigma^{-1}\hat B$ and $\hat V_x$. Under the null $H_0:\text{rank}(B)=L$, $\mathcal{CD}_B(L)\xrightarrow{d}\chi^2_{(N-L)(K-L)}$ (Cragg and Donald, 1997).

If the null hypothesis of a reduced rank is rejected, the researcher can proceed with the standard inference although the t-tests of parameter significance could still need to be robustified against possible model misspecification. If the null of a reduced rank is not rejected, the researcher needs to estimate consistently the reduced rank L of B. The estimation of the rank of B can be performed using the modified Bayesian information criterion (MBIC) of Ahn et al. (2018) by choosing the value of L (for $L=1,\dots,K-1$) that minimizes

$$MBIC(L)=\mathcal{CD}_B(L)-T^{0.2}(N-L)(K-L), \tag{20}$$

where $\mathcal{CD}_B(L)$ is the test in Eq. (19) of the null that the rank of B is equal to L. To minimize the probability of rejecting the null of a reduced rank when the true rank of B is deficient and to guard the procedure against the selection of nearly spurious factors (see also Wright, 2003), we fix the level of the rank test on B to be the same and small (say, 1%) for all levels of the subsequent tests.

This step of the model reduction procedure can be implemented using any available rank test. In our setup, the intercept is always included in the model. If the rank is estimated to be $1\le l\le K-1$, construct $N\times l$ matrices $\tilde B$ by selecting all possible combinations of $l-1$ risk factors, $\tilde f$, and perform a rank test on each $\tilde B$. Then, choose the $\tilde f$ that gives rise to the largest rejection of the reduced-rank hypothesis. See also Bryzgalova (2016) and Feng et al. (2017) for alternative model selection methods based on the lasso estimator in a two-pass setting.

Columns 2–4 of Table 4 report the probabilities of retaining factors in the proposed model reduction procedure. We fix the significance level of the rank test on B to be 1%. In addition, we denote by P_A, P_B, and P_C the marginal probability of retaining the useful factors, the marginal probability of eliminating the spurious factors, and the joint probability of retaining the useful factors and eliminating the spurious factors, respectively. The reported probabilities are numerically identical for the correctly specified and misspecified versions of each model.

To make the simulation design more challenging for the model reduction procedure, we consider, in addition to the three models described above, a model with a constant term, three useful factors, and two spurious factors. The results for all models suggest that our model reduction procedure is very effective in retaining the useful factors and eliminating the spurious factors from the analysis. For sample sizes $T\ge 600$, the most challenging scenario, the model with three useful and two spurious factors, retains (removes) the useful (spurious) factors with probability one.

Assessing the empirical rejection probabilities of the parameter significance tests before and after the identification-inducing reduction procedure is implemented could be desirable. The Wald test provides a

[8] If the integrity of the model needs to be preserved, one could use the limiting distribution in Theorem 1 to conduct inference on the risk premia parameters that is valid under possible lack of identification and model misspecification. However, this requires knowledge of which factor is spurious. Kleibergen (2009) develops alternative test procedures for constructing confidence intervals that are asymptotically valid irrespective of the degree of identification. When the model is of reduced rank, the corresponding confidence intervals are unbounded.

Table 4
Probabilities of retaining factors in the model reduction procedure.

The table presents the probabilities of retaining factors in our proposed model reduction procedure. The results are reported for different values of the number of time series observations (T). The level of the rank test on B is 1%. P_A, P_B, and P_C are the marginal probability of retaining the useful factors, the marginal probability of eliminating the spurious factors, and the joint probability of retaining the useful factors and eliminating the spurious factors, respectively. The table also reports the size of the Wald test with weighting matrix constructed under correct model specification when all the factors are included in the model ($W_c^{(all)}$), the size of the Wald test with weighting matrix constructed under correct model specification when only the selected factors are included in the model ($W_c^{(selected)}$), and the size of the Wald test with weighting matrix constructed under potential model misspecification when only the selected factors are included in the model ($W_m^{(selected)}$).

	Selection probabilities			$W_c^{(all)}$			$W_c^{(selected)}$			$W_m^{(selected)}$		
T	P_A	P_B	P_C	10%	5%	1%	10%	5%	1%	10%	5%	1%
Panel A: One useful factor only												
200	1.000	–	1.000	0.288	0.207	0.098	0.288	0.207	0.098	0.110	0.057	0.013
600	1.000	–	1.000	0.219	0.142	0.055	0.219	0.142	0.055	0.103	0.053	0.011
1000	1.000	–	1.000	0.207	0.132	0.048	0.207	0.132	0.048	0.103	0.052	0.011
Panel B: One spurious factor only												
200	–	0.948	0.948	0.996	0.994	0.990	0.219	0.153	0.082	0.122	0.066	0.016
600	–	0.982	0.982	1.000	1.000	0.998	0.157	0.097	0.039	0.107	0.055	0.012
1000	–	0.986	0.986	1.000	1.000	0.997	0.147	0.088	0.034	0.104	0.053	0.011
Panel C: One useful and one spurious factor												
200	1.000	0.951	0.951	0.991	0.988	0.972	0.318	0.238	0.134	0.108	0.056	0.012
600	1.000	0.982	0.982	1.000	0.999	0.985	0.232	0.157	0.071	0.102	0.051	0.010
1000	1.000	0.986	0.986	1.000	0.999	0.983	0.217	0.145	0.061	0.103	0.052	0.010
Panel D: Three useful and two spurious factors												
200	1.000	0.998	0.998	0.942	0.864	0.560	0.268	0.187	0.085	0.111	0.058	0.013
600	1.000	1.000	1.000	0.913	0.812	0.489	0.197	0.122	0.042	0.104	0.053	0.011
1000	1.000	1.000	1.000	0.903	0.792	0.456	0.181	0.111	0.037	0.103	0.052	0.011

convenient way to perform this comparison. In the evaluation of the empirical size of the Wald test, we use γ^* as the pseudo-true values for the useful factors and zero as the reference values for the spurious factors. We denote the augmented parameter vector by $\tilde{\gamma}^*$. The Wald test for all parameters prior to the reduction procedure takes the form $W_c^{(all)} = T(\hat{\gamma} - \tilde{\gamma}^*)' V_{\hat{\gamma}}^{-1} (\hat{\gamma} - \tilde{\gamma}^*)$, where $V_{\hat{\gamma}}$ is the covariance matrix of $\hat{\gamma}$ defined in Eq. (15). The subscript in $W_c^{(all)}$ indicates that the covariance matrix of the parameter estimates is obtained under the assumption that the model is correctly specified. The corresponding Wald test for the factors selected by the identification-inducing procedure is denoted by $W_c^{(selected)}$. Finally, we present results for the Wald test $W_m^{(selected)}$, where the covariance matrix is computed allowing for potential model misspecification (see Gospodinov et al., 2018). The empirical rejection rates of these Wald tests are reported in Table 4.

In line with our theoretical results, when all factors are included and the model contains spurious factors, the empirical size of the Wald test is characterized by strong over-rejections. When the model does not contain spurious factors [$W_c^{(all)}$ in Panel A of Table 4] or after the identification-inducing model reduction procedure is performed [$W_c^{(selected)}$], the tests also exhibit over-rejections that are due to the fact that the true model is misspecified while $W_c^{(all)}$ and $W_c^{(selected)}$ are constructed under the assumption of correct model specification. The test $W_m^{(selected)}$ accounts for the misspecification uncertainty and has the correct size, after the full rank condition for the model is ensured. Finally, while not reported in Table 4 to conserve space, the power of the ML specification test, $\mathcal{S}$, is very low and bounded by the size of the test when a spurious factor is present (see Gospodinov et al., 2014b; Gospodinov et al., 2017a), but it increases to one in large samples when the full identification of the model is restored.

In unreported experiments, we also consider intermediate cases in which priced factors, that is, factors that carry nonzero risk premia, are only weakly correlated with the returns on the test assets. In these scenarios, the Wald test exhibits some size distortions in small samples, but these distortions tend to disappear as the sample size increases. A more rigorous treatment of these intermediate cases is a promising direction for future research.

4. Empirical analysis

We evaluate the performance of several prominent asset pricing models with traded and non-traded factors in light of our analytical and simulation results in Sections 2 and 3. First, we describe the data used in the empirical analysis and outline the different specifications of the asset pricing models considered. Next, we present our results.

4.1. Data and asset pricing models

The return data are from Kenneth French's website and consist of the monthly value-weighted gross returns on the 25 Fama–French size and book-to-market ranked portfolios, 25 Fama–French size- and momentum-ranked portfolios, and 32 Fama–French size-, operating profitability-, and investment-ranked portfolios. The data are from January 1967 to December 2012 (552 monthly observations). The beginning date of our sample period is dictated by profitability and investment data availability.[9]

[9] We thank Lu Zhang for sharing his data with us.

We analyze six asset pricing models starting with the conditional labor model (C-LAB) of Jagannathan and Wang (1996). This model incorporates measures of the return on human capital as well as the change in financial wealth and allows the conditional moments to vary with a state variable, *prem*, the lagged yield spread between Baa- and Aaa-rated corporate bonds from the Board of Governors of the Federal Reserve System. The cross-sectional specification for this model is

$$\mu_R^{C-LAB} = 1_N\gamma_0 + \beta_{vw}\gamma_{vw} + \beta_{labor}\gamma_{labor} + \beta_{prem}\gamma_{prem}, \quad (21)$$

where *vw* is the excess return (in excess of the one-month T-bill rate from Ibbotson Associates) on the value-weighted stock market index (NYSE, Amex, and Nasdaq) from Kenneth French's website, and *labor* is the growth rate in per capita labor income, L, defined as the difference between total personal income and dividend payments, divided by the total population (from the Bureau of Economic Analysis). Following Jagannathan and Wang (1996), we use a two-month moving average to construct the growth rate $labor_t = (L_{t-1} + L_{t-2})/(L_{t-2} + L_{t-3}) - 1$, for the purpose of minimizing the influence of measurement error.

Our second model (CC-CAY) is a conditional version of the consumption CAPM of Lettau and Ludvigson (2001). The relation is

$$\mu_R^{CC-CAY} = 1_N\gamma_0 + \beta_{cg}\gamma_{cg} + \beta_{cay}\gamma_{cay} + \beta_{cg\cdot cay}\gamma_{cg\cdot cay}, \quad (22)$$

where *cg* is the growth rate in real per capita nondurable consumption (seasonally adjusted at annual rates) from the Bureau of Economic Analysis, and *cay*, the conditioning variable, is a consumption-aggregate wealth ratio. This specification is obtained by scaling the constant term and the *cg* factor of a linearized consumption CAPM by a constant and *cay*. Following Vissing-Jørgensen and Attanasio (2003), we linearly interpolate the quarterly values of *cay* to permit analysis at the monthly frequency.

The third model (ICAPM) is an empirical implementation of the Merton (1973) intertemporal extension of the CAPM based on Campbell (1996), who argues that innovations in state variables that forecast future investment opportunities should serve as the factors. The cross-sectional relation for the five-factor specification proposed by Petkova (2006) is

$$\mu_R^{ICAPM} = 1_N\gamma_0 + \beta_{vw}\gamma_{vw} + \beta_{term}\gamma_{term} + \beta_{def}\gamma_{def} + \beta_{div}\gamma_{div} + \beta_{rf}\gamma_{rf}, \quad (23)$$

where *term* is the difference between the yields of ten- and one-year government bonds (from the Board of Governors of the Federal Reserve System), *def* is the difference between the yields of long-term corporate Baa bonds and long-term government bonds (from Ibbotson Associates), *div* is the dividend yield on the Center for Research in Security Prices (CRSP) value-weighted stock market portfolio, and *rf* is the one-month T-bill yield (from CRSP, Fama risk-free rates). The actual factors for *term, def* , *div*, and *rf* are their innovations from a VAR(1) system of seven state variables that also includes *vw*, *smb*, and *hml* [the market, size, and value factors, respectively, of the three-factor model of Fama and French (1993)].

We complete our list of models with traded and non-traded factors by considering the (D-CCAPM) specification of Yogo (2006), which highlights the cyclical role of durable consumption in asset pricing. The asset pricing restriction is

$$\mu_R^{D-CCAPM} = 1_N\gamma_0 + \beta_{vw}\gamma_{vw} + \beta_{cg}\gamma_{cg} + \beta_{cgdur}\gamma_{cgdur}, \quad (24)$$

where *cgdur* is the growth rate in real per capita durable consumption (seasonally adjusted at annual rates) from the Bureau of Economic Analysis.

Our fifth model (FF3) is the three-factor model of Fama and French (1993). The cross-sectional relation is given by

$$\mu_R^{FF3} = 1_N\gamma_0 + \beta_{vw}\gamma_{vw} + \beta_{smb}\gamma_{smb} + \beta_{hml}\gamma_{hml}, \quad (25)$$

where *smb* is the return difference between portfolios of stocks with small and large market capitalizations and *hml* is the return difference between portfolios of stocks with high and low book-to-market ratios (from Kenneth French's website).

Finally, we consider the newly proposed empirical specification (HXZ) of Hou et al. (2015), which is built on the neoclassical q theory of investment. The beta representation of the model is

$$\mu_R^{HXZ} = 1_N\gamma_0 + \beta_{vw}\gamma_{vw} + \beta_{me}\gamma_{me} + \beta_{roe}\gamma_{roe} + \beta_{ia}\gamma_{ia}, \quad (26)$$

where *me* is the difference between the return on a portfolio of small size stocks and the return on a portfolio of big size stocks, *roe* is the difference between the return on a portfolio of high profitability stocks and the return on a portfolio of low profitability stocks, and *ia* is the difference between the return on a portfolio of low investment stocks and the return on a portfolio of high investment stocks. This four-factor model has been shown to successfully explain many asset pricing anomalies.

Empirical results for the five-factor model of Fama and French (2015), the three-factor model of Fama and French (1993) augmented with the momentum factor of Carhart (1997), and the three-factor model of Fama and French (1993) augmented with the non-traded liquidity factor of Pastor and Stambaugh (2003) are available upon request.

4.2. Results

The results for all models are reported for both the invariant (ML) and non-invariant (GLS) estimators. Starting with C-LAB (Table 5), we investigate whether this model is well identified. The outcomes of the rank test suggest that C-LAB is poorly identified across different sets of test assets. The p-values of these tests are large, ranging from 0.53 to 0.72, and indicate that the null hypothesis of a deficient column rank for the B matrix cannot be rejected. Similar concerns were also raised by Kleibergen and Paap (2006) using the original data in Jagannathan and Wang (1996). This identification failure results in the inability of the specification test to reject the model (see Gospodinov et al., 2014b) and in spuriously high R^2s (indistinguishable from one in the columns labeled "All" in Table 5) for ML. Based on $\mathcal{S}$ and the R^2 for ML, C-LAB appears to have a spectacular fit, and a researcher would likely proceed with t-tests of parameter significance with standard errors computed under the assumption of correct model specification. This would lead us to conclude that the *labor* and *prem*

Table 5
Test statistics for conditional labor model (C-LAB).

The table reports test statistics for C-LAB. $\mathcal{CD}_B$ denotes the Cragg and Donald (1997) test for the null of a reduced rank. $\mathcal{Q}$ and $\mathcal{S}$ denote the Shanken (1985) tests of correct model specification based on the generalized least squares (GLS) and maximum likelihood (ML) estimators, respectively. The rows for the different factors report the t-tests of statistical significance with standard errors computed under the assumption of correct model specification and the misspecification-robust t-tests (in square brackets). R^2 denotes the squared correlation coefficient between the fitted expected returns and the average realized returns.

Method	Factor	All	Selected
Panel A: 25 portfolios formed on size and book-to-market			
Rank test	$\mathcal{CD}_B$ (p-value)	20.87 (0.5290)	465.03 (0.0000)
ML	$\mathcal{S}$ (p-value)	20.87 (0.4672)	68.79 (0.0000)
	R^2	1.0000	0.1447
	vw	1.42 [0.01]	−3.65 [−2.92]
	labor	−3.14 [−0.01]	–
	prem	−4.07 [−0.01]	–
GLS	$\mathcal{Q}$ (p-value)	69.68 (0.0000)	71.96 (0.0000)
	R^2	0.1111	0.0993
	vw	−2.66 [−2.09]	−3.14 [−2.97]
	labor	−1.05 [−0.47]	–
	prem	0.56 [0.23]	–
Panel B: 25 portfolios formed on size and momentum			
Rank test	$\mathcal{CD}_B$ (p-value)	17.71 (0.7231)	505.71 (0.0000)
ML	$\mathcal{S}$ (p-value)	17.71 (0.6674)	105.80 (0.0000)
	R^2	1.0000	0.1128
	vw	−0.67 [−0.00]	−0.95 [−0.68]
	labor	2.82 [0.00]	–
	prem	3.97 [0.00]	–
GLS	$\mathcal{Q}$ (p-value)	97.23 (0.0000)	106.09 (0.0000)
	R^2	0.6890	0.0963
	vw	−0.49 [−0.42]	−0.75 [−0.68]
	labor	1.89 [0.98]	–
	prem	−0.89 [−0.37]	–
Panel C: 32 portfolios formed on size, profitability, and investment			
Rank test	$\mathcal{CD}_B$ (p-value)	25.57 (0.6486)	575.04 (0.0000)
ML	$\mathcal{S}$ (p-value)	25.56 (0.5972)	159.18 (0.0000)
	R^2	1.0000	0.1055
	vw	0.73 [0.06]	−2.61 [−1.68]
	labor	−4.86 [−0.06]	–
	prem	−1.74 [−0.06]	–
GLS	$\mathcal{Q}$ (p-value)	161.93 (0.0000)	161.99 (0.0000)
	R^2	0.0679	0.0717
	vw	−1.88 [−1.69]	−1.90 [−1.71]
	labor	−0.18 [−0.07]	–
	prem	0.07 [0.03]	–

factors are often priced in the cross section of expected returns, as emphasized by the high traditional t-ratios on the *prem* and *labor* factors for ML. The evidence of pricing for the market factor is rather weak, with traditional absolute t-ratio values ranging from 0.67 to 1.42 for ML. These empirical findings are consistent with our methodological results and reveal the spurious nature of inference as factors that are spurious are selected with high probability, while factors that are useful (such as the market factor) are driven out of the model.

Applying the model reduction procedure, described in Section 3 to C-LAB reveals that only the market factor survives the identification-inducing procedure. Essentially, C-LAB reduces to CAPM, and the $\mathcal{S}$ test now has power to reject the model (see columns labeled "Selected" in Table 5). In turn, the R^2s provide a completely different and more realistic assessment of the goodness of fit of the model, ranging from 0.11 to 0.14. The high misspecification-robust t -ratios on *vw* in Panel A (see Gospodinov et al., 2018, for the derivation of misspecification-robust t-ratios for ML) suggest some strong pricing ability for the market factor when the test portfolios are formed on size and book-to-market. In contrast, when considering portfolios formed on size and momentum, the evidence of pricing for *vw* is very limited, consistent with the uncontroversial finding that CAPM cannot explain the returns on portfolios formed on momentum. Panel C also shows that the pricing ability of *vw* is somewhat weak when employing misspecification-robust t-ratios and considering portfolios formed on size, operating profitability, and investment.

Non-invariant estimators, such as the GLS estimator, provide a less optimistic picture of C-LAB compared with ML. The p-values of $\mathcal{Q}$ in Table 5 are always zero even before applying the model reduction procedure. Therefore, even if $\mathcal{Q}$ is inconsistent under identification failure (Gospodinov et al., 2014b), it seems to be more robust to lack of identification and can detect model misspecification with higher probability than $\mathcal{S}$. In sharp contrast with the R^2s based on ML, the R^2s for the GLS estimator are much smaller (see the columns labeled "All" in the table). Finally, after applying our model selection procedure, the pricing implications for *vw* (based on misspecification-robust t-ratios as in Kan et al. (2013)) are largely consistent across invariant and non-invariant estimators.

The spurious nature of the results analyzed in this paper are probably best illustrated with CC-CAY in Table 6. The rank tests in all three panels provide strong evidence that the model is not identified. Ignoring the outcome of the rank tests would lead us to conclude that the model estimated by ML is correctly specified and that scaled consumption growth, $cg \cdot cay$, is highly significant. However, none of the factors survives after applying the proposed model reduction procedure because none of the factors (or a subset of factors) in this model satisfies the rank condition. Supporting evidence for this conclusion is provided in Kleibergen (2009). Kleibergen (2009) finds that the identification-robust confidence intervals for the risk premia on *cg, cay*, and $cg \cdot cay$ are unbounded, which suggests that these factors are likely to be spurious.

The results for ICAPM and D-CCAPM in Tables 7 and 8 further reveal the fragility of statistical inference in models with factors that are only weakly correlated with the test asset returns. Similar to the case of C-LAB in Table 5, only the market factor survives the identification-inducing procedure in ICAPM and D-CCAPM.

Turning to models with traded factors only, the results for the rank tests in Tables 9 and 10 for FF3 and HXZ suggest that these models are well identified, albeit misspecified. Our main empirical findings can be summarized as follows. Models with non-traded factors are often poorly identified and tend to produce highly misleading inference in terms of spuriously high statistical significance and lack of power in rejecting the null of correct model specification. In addition to the outcome of the rank tests, two observations cast doubts on the validity of the results for these models: the difference between the t-statistics computed under the assumption of correct specification and the misspecification-robust

Table 6
Test statistics for a conditional version of the consumption capital asset pricing model (CC-CAY).

The table reports test statistics for CC-CAY. $\mathcal{CD}_B$ denotes the Cragg and Donald (1997) test for the null of a reduced rank. $\mathcal{Q}$ and $\mathcal{S}$ denote the Shanken (1985) tests of correct model specification based on the generalized least squares (GLS) and maximum likelihood (ML) estimators, respectively. The rows for the different factors report the t-tests of statistical significance with standard errors computed under the assumption of correct model specification and the misspecification-robust t-tests (in square brackets). R^2 denotes the squared correlation coefficient between the fitted expected returns and the average realized returns.

Method	Factor	All	Selected
Panel A: 25 portfolios formed on size and book-to-market			
Rank test	$\mathcal{CD}_B$ (p-value)	14.10 (0.8978)	–
ML	$\mathcal{S}$ (p-value)	13.85 (0.8758)	81.90 (0.0000)
	R^2	0.9995	–
	cg	−2.23 [−0.12]	–
	cay	−0.77 [−0.04]	–
	cg · *cay*	3.63 [0.19]	–
GLS	$\mathcal{Q}$ (p-value)	71.77 (0.0009)	81.90 (0.0000)
	R^2	0.0475	–
	cg	0.70 [0.40]	–
	cay	1.34 [0.70]	–
	cg · *cay*	1.84 [0.94]	–
Panel B: 25 portfolios formed on size and momentum			
Rank test	$\mathcal{CD}_B$ (p-value)	21.10 (0.5146)	–
ML	$\mathcal{S}$ (p-value)	20.79 (0.4721)	106.65 (0.0000)
	R^2	0.9977	–
	cg	−0.63 [−0.05]	–
	cay	−0.20 [−0.01]	–
	cg · *cay*	4.69 [0.16]	–
GLS	$\mathcal{Q}$ (p-value)	73.91 (0.0098)	106.65 (0.0000)
	R^2	0.0368	–
	cg	1.71 [1.27]	–
	cay	3.59 [2.51]	–
	cg · *cay*	1.64 [1.01]	–
Panel C: 32 portfolios formed on size, profitability, and investment			
Rank test	$\mathcal{CD}_B$ (p-value)	23.79 (0.7394)	–
ML	$\mathcal{S}$ (p-value)	23.68 (0.6985)	165.59 (0.0000)
	R^2	0.9999	–
	cg	−4.21 [−0.15]	–
	cay	1.21 [0.05]	–
	cg · *cay*	4.04 [0.10]	–
GLS	$\mathcal{Q}$ (p-value)	163.80 (0.0000)	165.59 (0.0000)
	R^2	0.0009	–
	cg	−0.37 [−0.16]	–
	cay	1.07 [0.41]	–
	cg · *cay*	0.72 [0.31]	–

Table 7
Test statistics for intertemporal capital asset pricing model (ICAPM).

The table reports test statistics for ICAPM. $\mathcal{CD}_B$ denotes the Cragg and Donald (1997) test for the null of a reduced rank. $\mathcal{Q}$ and $\mathcal{S}$ denote the Shanken (1985) tests of correct model specification based on the generalized least squares (GLS) and maximum likelihood (ML) estimators, respectively. The rows for the different factors report the t-tests of statistical significance with standard errors computed under the assumption of correct model specification and the misspecification-robust t-tests (in square brackets). R^2 denotes the squared correlation coefficient between the fitted expected returns and the average realized returns.

Method	Factor	All	Selected
Panel A: 25 portfolios formed on size and book-to-market			
Rank test	$\mathcal{CD}_B$ (p-value)	23.85 (0.2488)	465.03 (0.0000)
ML	$\mathcal{S}$ (p-value)	21.46 (0.3118)	68.79 (0.0000)
	R^2	0.9942	0.1447
	vw	1.79 [0.61]	−3.65 [−2.92]
	term	4.78 [1.05]	–
	def	1.16 [0.41]	–
	div	−2.14 [−0.60]	–
	rf	−3.19 [−0.90]	–
GLS	$\mathcal{Q}$ (p-value)	63.72 (0.0016)	71.96 (0.0000)
	R^2	0.3692	0.0993
	vw	−1.77 [−1.38]	−3.14 [−2.97]
	term	2.13 [1.16]	–
	def	−0.23 [−0.14]	–
	div	1.00 [0.64]	–
	rf	−1.66 [−0.91]	–
Panel B: 25 portfolios formed on size and momentum			
Rank test	$\mathcal{CD}_B$ (p-value)	25.36 (0.1879)	505.71 (0.0000)
ML	$\mathcal{S}$ (p-value)	24.55 (0.1757)	105.80 (0.0000)
	R^2	0.9976	0.1128
	vw	2.02 [0.62]	−0.95 [−0.68]
	term	3.41 [0.57]	–
	def	−3.72 [−0.60]	–
	div	−3.44 [−0.66]	–
	rf	−1.63 [−0.49]	–
GLS	$\mathcal{Q}$ (p-value)	99.90 (0.0000)	106.09 (0.0000)
	R^2	0.1048	0.0963
	vw	0.17 [0.13]	−0.75 [−0.68]
	term	1.05 [0.56]	–
	def	−0.25 [−0.13]	–
	div	−1.33 [−0.74]	–
	rf	−1.35 [−0.84]	–
Panel C: 32 portfolios formed on size, profitability, and investment			
Rank test	$\mathcal{CD}_B$ (p-value)	18.70 (0.8805)	575.04 (0.0000)
ML	$\mathcal{S}$ (p-value)	18.39 (0.8611)	159.18 (0.0000)
	R^2	0.9996	0.1055
	vw	−2.42 [−0.45]	−2.61 [−1.68]
	term	−3.64 [−0.47]	–
	def	−3.58 [−0.46]	–
	div	2.70 [0.43]	–
	rf	1.55 [0.35]	–
GLS	$\mathcal{Q}$ (p-value)	152.35 (0.0000)	161.99 (0.0000)
	R^2	0.2528	0.0717
	vw	−1.62 [−0.98]	−1.90 [−1.71]
	term	−1.56 [−0.61]	–
	def	−0.79 [−0.32]	–
	div	0.23 [0.10]	–
	rf	0.30 [0.14]	–

t-statistics (with the misspecification-robust t-statistics being typically small) and the unrealistically high value of the R^2. The models that perform the best are FF3 and especially HXZ in which all the factors appear to contribute to pricing and are characterized by statistically significant risk premia. Out of the different sets of test portfolios, the portfolios formed on size and momentum, and size and short- and long-term reversal appear to be the most challenging from a pricing perspective.

In unreported empirical investigations, we explore the performance of these six models using the 25 Fama–French portfolios formed on size and short-term reversal, 25 Fama–French portfolios formed on size and long-term reversal, and 25 Fama–French portfolios formed on size and book-to-market plus 17 industry portfolios (all the test assets are from Kenneth French's website). The results based on these three additional sets of test asset returns are largely consistent with the results reported in this section.

5. Concluding remarks

In this paper, we study the limiting properties of ML-based tests of statistical significance and goodness of fit in

Table 8
Test statistics for durable consumption capital asset pricing model (D-CCAPM).

The table reports test statistics for D-CCAPM. $\mathcal{CD}_B$ denotes the Cragg and Donald (1997) test for the null of a reduced rank. $\mathcal{Q}$ and $\mathcal{S}$ denote the Shanken (1985) tests of correct model specification based on the generalized least squares (GLS) and maximum likelihood (ML) estimators, respectively. The rows for the different factors report the t-tests of statistical significance with standard errors computed under the assumption of correct model specification and the misspecification-robust t-tests (in square brackets). R^2 denotes the squared correlation coefficient between the fitted expected returns and the average realized returns.

Method	Factor	All	Selected
Panel A: 25 portfolios formed on size and book-to-market			
Rank test	$\mathcal{CD}_B$ (p-value)	37.87 (0.0190)	465.03 (0.0000)
ML	$\mathcal{S}$ (p-value)	28.37 (0.1299)	68.79 (0.0000)
	R^2	0.9762	0.1447
	vw	−2.17 [−1.77]	−3.65 [−2.92]
	cg	5.31 [1.77]	–
	cgdur	2.35 [0.54]	–
GLS	$\mathcal{Q}$ (p-value)	61.96 (0.0023)	71.96 (0.0000)
	R^2	0.3642	0.0993
	vw	−3.09 [−2.93]	−3.14 [−2.97]
	cg	2.61 [1.69]	–
	cgdur	1.09 [0.79]	–
Panel B: 25 portfolios formed on size and momentum			
Rank test	$\mathcal{CD}_B$ (p-value)	30.73 (0.1019)	505.71 (0.0000)
ML	$\mathcal{S}$ (p-value)	30.15 (0.0889)	105.80 (0.0000)
	R^2	0.9926	0.1128
	vw	−0.81 [−0.62]	−0.95 [−0.68]
	cg	−0.07 [−0.00]	–
	cgdur	5.52 [0.22]	–
GLS	$\mathcal{Q}$ (p-value)	95.14 (0.0000)	106.09 (0.0000)
	R^2	0.0143	0.0963
	vw	−1.10 [−0.96]	−0.75 [−0.68]
	cg	2.21 [1.24]	–
	cgdur	2.04 [1.07]	–
Panel C: 32 portfolios formed on size, profitability, and investment			
Rank test	$\mathcal{CD}_B$ (p-value)	29.93 (0.4175)	575.04 (0.0000)
ML	$\mathcal{S}$ (p-value)	29.82 (0.3717)	159.18 (0.0000)
	R^2	0.9990	0.1055
	vw	0.08 [0.00]	−2.61 [−1.68]
	cg	−2.89 [−0.02]	–
	cgdur	4.02 [0.08]	–
GLS	$\mathcal{Q}$ (p-value)	154.38 (0.0000)	161.99 (0.0000)
	R^2	0.5117	0.0717
	vw	−2.14 [−1.83]	−1.90 [−1.71]
	cg	1.07 [0.50]	–
	cgdur	2.32 [1.13]	–

Table 9
Test statistics for the Fama and French three-factor model (FF3).

The table reports test statistics for FF3. $\mathcal{CD}_B$ denotes the Cragg and Donald (1997) test for the null of a reduced rank. $\mathcal{Q}$ and $\mathcal{S}$ denote the Shanken (1985) tests of correct model specification based on the generalized least squares (GLS) and maximum likelihood (ML) estimators, respectively. The rows for the different factors report the t-tests of statistical significance with standard errors computed under the assumption of correct model specification and the misspecification-robust t-tests (in square brackets). R^2 denotes the squared correlation coefficient between the fitted expected returns and the average realized returns.

Method	Factor	All	Selected
Panel A: 25 portfolios formed on size and book-to-market			
Rank test	$\mathcal{CD}_B$ (p-value)	321.18 (0.0000)	321.18 (0.0000)
ML	$\mathcal{S}$ (p-value)	51.05 (0.0003)	51.05 (0.0003)
	R^2	0.7337	0.7337
	vw	−3.80 [−3.03]	−3.80 [−3.03]
	smb	1.73 [1.72]	1.73 [1.72]
	hml	3.04 [3.03]	3.04 [3.03]
GLS	$\mathcal{Q}$ (p-value)	55.61 (0.0004)	55.61 (0.0004)
	R^2	0.6901	0.6901
	vw	−3.29 [−3.02]	−3.29 [−3.02]
	smb	1.73 [1.73]	1.73 [1.73]
	hml	3.04 [3.04]	3.04 [3.04]
Panel B: 25 portfolios formed on size and momentum			
Rank test	$\mathcal{CD}_B$ (p-value)	111.25 (0.0000)	111.25 (0.0000)
ML	$\mathcal{S}$ (p-value)	77.55 (0.0000)	77.55 (0.0000)
	R^2	0.8805	0.8805
	vw	−5.32 [−1.76]	−5.32 [−1.76]
	smb	4.06 [2.84]	4.06 [2.84]
	hml	−4.63 [−1.48]	−4.63 [−1.48]
GLS	$\mathcal{Q}$ (p-value)	93.49 (0.0000)	93.49 (0.0000)
	R^2	0.4934	0.4934
	vw	−1.88 [−1.48]	−1.88 [−1.48]
	smb	2.99 [2.76]	2.99 [2.76]
	hml	−1.30 [−0.95]	−1.30 [−0.95]
Panel C: 32 portfolios formed on size, profitability, and investment			
Rank test	$\mathcal{CD}_B$ (p-value)	256.43 (0.0000)	256.43 (0.0000)
ML	$\mathcal{S}$ (p-value)	133.50 (0.0000)	133.50 (0.0000)
	R^2	0.5981	0.5981
	vw	−0.46 [−0.20]	−0.46 [−0.20]
	smb	0.94 [0.88]	0.94 [0.88]
	hml	4.66 [2.85]	4.66 [2.85]
GLS	$\mathcal{Q}$ (p-value)	141.97 (0.0000)	141.97 (0.0000)
	R^2	0.5394	0.5394
	vw	−0.92 [−0.77]	−0.92 [−0.77]
	smb	1.12 [1.09]	1.12 [1.09]
	hml	3.96 [3.46]	3.96 [3.46]

asset pricing models and show that the inference based on these tests can be spurious when the models are unidentified. The spurious results in these models arise from the combined effect of identification failure and model misspecification, which is not an isolated problem limited to a particular sample (data frequency), test assets, and asset pricing models. This suggests that the statistical evidence on the pricing ability of many macro factors and their usefulness in explaining the cross section of asset returns should be interpreted with caution. Some warning signs about this problem (for example, the outcome of a rank test) are often ignored by researchers. While the non-invariant GLS estimator also suffers from similar problems, the invariant ML estimator turns out to be much more sensitive to model misspecification and lack of identification.

Given the severity of the inference problems associated with invariant estimators of possibly unidentified and misspecified asset pricing models that we show in this paper, our recommendations for empirical practice can be summarized as follows. Any model should be subjected to a rank test that will provide evidence on whether the model parameters are identified or not. If the null hypothesis of a reduced rank is rejected, a researcher can proceed with the standard tools for inference in analyzing and evaluating the model. If the null of a reduced rank is not rejected, a researcher needs to estimate consistently the reduced rank of the model and select the combination of factors that delivers the largest rejection of the reduced rank hypothesis. This procedure would restore the standard inference, although it could still need to be robustified against possible model misspecification as in Gospodinov et al. (2018). An alternative empirical strategy is to work with non-invariant estimators and pursue

Table 10
Test statistics for the Hou, Xue, and Zhang model (HXZ).

The table reports test statistics for HXZ. $\mathcal{CD}_B$ denotes the Cragg and Donald (1997) test for the null of a reduced rank. $\mathcal{Q}$ and $\mathcal{S}$ denote the Shanken (1985) tests of correct model specification based on the generalized least squares (GLS) and maximum likelihood (ML) estimators, respectively. The rows for the different factors report the t-tests of statistical significance with standard errors computed under the assumption of correct model specification and the misspecification-robust t-tests (in square brackets). R^2 denotes the squared correlation coefficient between the fitted expected returns and the average realized returns.

Method	Factor	All	Selected
Panel A: 25 portfolios formed on size and book-to-market			
Rank test	$\mathcal{CD}_B$ (p-value)	138.27 (0.0000)	138.27 (0.0000)
ML	$\mathcal{S}$ (p-value)	50.72 (0.0002)	50.72 (0.0002)
	R^2	0.7607	0.7607
	vw	−3.29 [−2.25]	−3.29 [−2.25]
	me	2.53 [2.37]	2.53 [2.37]
	roe	1.65 [1.03]	1.65 [1.03]
	ia	2.72 [2.05]	2.72 [2.05]
GLS	$\mathcal{Q}$ (p-value)	56.09 (0.0003)	56.09 (0.0003)
	R^2	0.6938	0.6938
	vw	−2.98 [−2.67]	−2.98 [−2.67]
	me	2.38 [2.34]	2.38 [2.34]
	roe	1.24 [1.06]	1.24 [1.06]
	ia	2.54 [2.35]	2.54 [2.35]
Panel B: 25 portfolios formed on size and momentum			
Rank test	$\mathcal{CD}_B$ (p-value)	65.73 (0.0000)	65.73 (0.0000)
ML	$\mathcal{S}$ (p-value)	51.56 (0.0001)	51.56 (0.0001)
	R^2	0.9347	0.9347
	vw	3.21 [0.75]	3.21 [0.75]
	me	3.80 [3.60]	3.80 [3.60]
	roe	3.71 [1.81]	3.71 [1.81]
	ia	3.35 [0.66]	3.35 [0.66]
GLS	$\mathcal{Q}$ (p-value)	65.79 (0.0009)	65.79 (0.0009)
	R^2	0.8784	0.8784
	vw	0.74 [0.58]	0.74 [0.58]
	me	3.46 [3.41]	3.46 [3.41]
	roe	3.23 [3.17]	3.23 [3.17]
	ia	0.96 [0.71]	0.96 [0.71]
Panel C: 32 portfolios formed on size, profitability, and investment			
Rank test	$\mathcal{CD}_B$ (p-value)	169.09 (0.0000)	169.09 (0.0000)
ML	$\mathcal{S}$ (p-value)	69.36 (0.0000)	69.36 (0.0000)
	R^2	0.8810	0.8810
	vw	2.94 [1.80]	2.94 [1.80]
	me	3.59 [3.40]	3.59 [3.40]
	roe	6.74 [4.45]	6.74 [4.45]
	ia	6.42 [5.65]	6.42 [5.65]
GLS	$\mathcal{Q}$ (p-value)	102.44 (0.0001)	102.44 (0.0001)
	R^2	0.7499	0.7499
	vw	0.90 [0.82]	0.90 [0.82]
	me	2.97 [2.86]	2.97 [2.86]
	roe	4.63 [4.56]	4.63 [4.56]
	ia	5.73 [5.57]	5.73 [5.57]

misspecification-robust inference that is asymptotically valid regardless of the degree of identification (Kleibergen, 2009; Gospodinov et al., 2014a).

Appendix

A.1. Auxiliary lemma

Auxiliary Lemma 1. *Let* $z = [z_1, z_2, \dots, z_K]' \sim \mathcal{N}(0_K, (G_1'\Sigma^{-1}G_1)^{-1}/\sigma^2_{f,K-1})$, *where* $G_1 = [1_N, \alpha, \beta_1, \dots, \beta_{K-2}]$ *and* $\sigma^2_{f,K-1} = \text{Var}[f_{K-1,t}]$. *Assume that* Y_t *is i.i.d. normal. Suppose that the model is misspecified and it contains a spurious factor (that is,* $\text{rank}(B) = K-1$*). Then, as* $T \to \infty$, *we have (i)* $\hat{\gamma}_0^{ML} \xrightarrow{d} -\frac{z_1}{z_2}$; *(ii)* $\hat{\gamma}_{1,i}^{ML} \xrightarrow{d} \mu_{f,i} - \frac{z_{i+2}}{z_2}$ *for* $i = 1, \dots, K-2$; *and (iii)* $\frac{\hat{\gamma}_{1,K-1}^{ML}}{\sqrt{T}} \xrightarrow{d} \frac{1}{z_2}$.

Proof. When the model is misspecified and contains a spurious factor (ordered last), we have $Gv^* = 0_N$ for $v^* = [0_K', 1]'$. Let $\hat{v}$ be the eigenvector associated with the largest eigenvalue of

$$\hat{\Omega} = (\hat{G}'\hat{\Sigma}^{-1}\hat{G})^{-1}[A(X'X/T)^{-1}A']. \tag{A.1}$$

Define $\hat{\psi} = [\hat{\psi}_1, \hat{\psi}_2, \dots, \hat{\psi}_K]'$ as

$$\hat{\psi}_i = -\frac{\hat{v}_i}{\hat{v}_{K+1}}, \quad i = 1, \dots, K, \tag{A.2}$$

which is asymptotically equivalent to the estimator

$$\tilde{\psi} = (\hat{G}_1'\hat{\Sigma}^{-1}\hat{G}_1)^{-1}(\hat{G}_1'\hat{\Sigma}^{-1}\hat{\beta}_{K-1}). \tag{A.3}$$

Because $\sqrt{T}\hat{\beta}_{K-1} \xrightarrow{d} \mathcal{N}(0_N, \Sigma/\sigma^2_{f,K-1})$, we have

$$\sqrt{T}\tilde{\psi} \xrightarrow{d} \mathcal{N}(0_K, (G_1'\Sigma^{-1}G_1)^{-1}/\sigma^2_{f,K-1}), \tag{A.4}$$

and $\sqrt{T}\hat{\psi}$ also has the same asymptotic distribution. Therefore, we can write

$$\hat{\gamma}_0^{ML} = -\frac{\sqrt{T}\hat{\psi}_1}{\sqrt{T}\hat{\psi}_2} \xrightarrow{d} -\frac{z_1}{z_2}, \tag{A.5}$$

$$\hat{\gamma}_{1,i}^{ML} = \hat{\mu}_{f,i} - \frac{\sqrt{T}\hat{\psi}_{i+2}}{\sqrt{T}\hat{\psi}_2} \xrightarrow{d} \mu_{f,i} - \frac{z_{i+2}}{z_2}, \quad i = 1, \dots, K-2, \tag{A.6}$$

and

$$\frac{\hat{\gamma}_{1,K-1}^{ML}}{\sqrt{T}} = \frac{\hat{\mu}_{f,K-1}}{\sqrt{T}} + \frac{1}{\sqrt{T}\hat{\psi}_2} \xrightarrow{d} \frac{1}{z_2}. \tag{A.7}$$

This completes the proof of the lemma. □

A.2. Proof of Theorem 1

part (a). Let $\sigma_i^2 = \text{Var}[z_i]$, $\sigma_{i,j} \equiv \text{Cov}[z_i, z_j]$, $\rho_{i,j} = \sigma_{i,j}/(\sigma_i\sigma_j)$, $G_2 = [1_N, \beta_1, \dots, \beta_{K-2}]$, and $\hat{G}_2 = [1_N, \hat{\beta}_1, \dots, \hat{\beta}_{K-2}]$, and define the random variables $\tilde{z}_2 \equiv z_2/\sigma_2 \sim \mathcal{N}(0,1)$, $x \sim \chi^2_{N-K}$, and $q_i \sim \mathcal{N}(0,1)$, where x and q_i are independent of $\tilde{z}_2$, and $b_i = (x + \tilde{z}_2^2)/(x + \tilde{z}_2^2 + q_i^2)$ for $i = 1, \dots, K-1$. We start with the squared t-ratio of the spurious factor, $t^2(\hat{\gamma}_{1,K-1}^{ML})$. Using the formula for the inverse of a partitioned matrix, we obtain

$$\begin{aligned} s^2(\hat{\gamma}_{1,K-1}^{ML}) &= (1 + \hat{\gamma}_1^{ML\prime}\hat{V}_f^{-1}\hat{\gamma}_1^{ML})\big(\hat{\beta}_{K-1}'[\hat{\Sigma}^{-1} \\ &\quad - \hat{\Sigma}^{-1}\hat{G}_2(\hat{G}_2'\hat{\Sigma}^{-1}\hat{G}_2)^{-1}\hat{G}_2'\hat{\Sigma}^{-1}]\hat{\beta}_{K-1}\big)^{-1} + \hat{\sigma}^2_{f,K-1} \\ &= \left(\frac{\hat{\gamma}_{1,K-1}^{ML}}{\hat{\sigma}_{f,K-1}}\right)^2\big(\hat{\beta}_{K-1}'\big[\hat{\Sigma}^{-1} \\ &\quad - \hat{\Sigma}^{-1}\hat{G}_2(\hat{G}_2'\hat{\Sigma}^{-1}\hat{G}_2)^{-1}\hat{G}_2'\hat{\Sigma}^{-1}\big]\hat{\beta}_{K-1}\big)^{-1} \\ &\quad + O_p(T^{\frac{1}{2}}) \end{aligned} \tag{A.8}$$

by using the fact that $\hat{\gamma}_{1,i}^{ML} = O_p(1)$ for $i = 1, \dots, K-2$ and $\hat{\gamma}_{1,K-1}^{ML} = O_p(T^{\frac{1}{2}})$. In addition, by defining u as

$$\sqrt{T}\hat{\sigma}_{f,K-1}\hat{\Sigma}^{-\frac{1}{2}}\hat{\beta}_{K-1} \xrightarrow{d} u \sim \mathcal{N}(0_N, I_N), \tag{A.9}$$

we obtain

$$
\begin{aligned}
&t^2(\hat{\gamma}^{ML}_{1,K-1}) \\
&= \frac{T(\hat{\gamma}^{ML}_{1,K-1})^2\hat{\beta}'_{K-1}[\hat{\Sigma}^{-1}-\hat{\Sigma}^{-1}\hat{G}_2(\hat{G}_2'\hat{\Sigma}^{-1}\hat{G}_2)^{-1}\hat{G}_2\hat{\Sigma}^{-1}]\hat{\beta}_{K-1}}{(\hat{\gamma}^{ML}_{1,K-1}/\hat{\sigma}_{f,K-1})^2} \\
&\quad + O_p(T^{-\frac{1}{2}}) \\
&= u'[I_N-\hat{\Sigma}^{-\frac{1}{2}}\hat{G}_2(\hat{G}_2'\hat{\Sigma}^{-1}\hat{G}_2)^{-1}\hat{G}_2'\hat{\Sigma}^{-\frac{1}{2}}]u + O_p(T^{-\frac{1}{2}}) \\
&\xrightarrow{d} u'[I_N-\Sigma^{-\frac{1}{2}}G_2(G_2'\Sigma^{-1}G_2)^{-1}G_2'\Sigma^{-\frac{1}{2}}]u \sim \chi^2_{N-K+1}.
\end{aligned} \tag{A.10}
$$

For the limiting distributions of $t(\hat{\gamma}^{ML}_0)$ and $t(\hat{\gamma}^{ML}_{1,i})$, $i=1,\ldots,K-2$, we use the formula for the inverse of a partitioned matrix to obtain the upper left $(K-1)\times(K-1)$ block of $(\hat{B}'\hat{\Sigma}^{-1}\hat{B})^{-1}$ as

$$
\begin{aligned}
&(\hat{G}_2'\hat{\Sigma}^{-1}\hat{G}_2)^{-1} \\
&\quad + \frac{(\hat{G}_2'\hat{\Sigma}^{-1}\hat{G}_2)^{-1}\hat{G}_2'\hat{\Sigma}^{-1}\hat{\beta}_{K-1}\hat{\beta}'_{K-1}\hat{\Sigma}^{-1}\hat{G}_2(\hat{G}_2'\hat{\Sigma}^{-1}\hat{G}_2)^{-1}}{\hat{\beta}'_{K-1}\hat{\Sigma}^{-1}\hat{\beta}_{K-1}-\hat{\beta}'_{K-1}\hat{\Sigma}^{-1}\hat{G}_2(\hat{G}_2'\hat{\Sigma}^{-1}\hat{G}_2)^{-1}\hat{G}_2'\hat{\Sigma}^{-1}\hat{\beta}_{K-1}} \\
&= (G_2'\Sigma^{-1}G_2)^{-1} \\
&\quad + \frac{(G_2'\Sigma^{-1}G_2)^{-1}G_2'\Sigma^{-\frac{1}{2}}uu'\Sigma^{-\frac{1}{2}}G_2(G_2'\Sigma^{-1}G_2)^{-1}}{u'[I_N-\Sigma^{-\frac{1}{2}}G_2(G_2'\Sigma^{-1}G_2)^{-1}G_2'\Sigma^{-\frac{1}{2}}]u} \\
&\quad + O_p(T^{-\frac{1}{2}}).
\end{aligned} \tag{A.11}
$$

We can write

$$
\begin{aligned}
&I_N-\Sigma^{-\frac{1}{2}}G_1(G_1'\Sigma^{-1}G_1)^{-1}G_1'\Sigma^{-\frac{1}{2}} \\
&= I_N-\Sigma^{-\frac{1}{2}}G_2(G_2'\Sigma^{-1}G_2)^{-1}G_2'\Sigma^{-\frac{1}{2}}-hh',
\end{aligned} \tag{A.12}
$$

where

$$
h = \frac{[I_N-\Sigma^{-\frac{1}{2}}G_2(G_2'\Sigma^{-1}G_2)^{-1}G_2'\Sigma^{-\frac{1}{2}}]\Sigma^{-\frac{1}{2}}\alpha}{\left(\alpha'\Sigma^{-\frac{1}{2}}[I_N-\Sigma^{-\frac{1}{2}}G_2(G_2'\Sigma^{-1}G_2)^{-1}G_2'\Sigma^{-\frac{1}{2}}]\Sigma^{-\frac{1}{2}}\alpha\right)^{\frac{1}{2}}}. \tag{A.13}
$$

With this expression, we can write

$$
\begin{aligned}
&u'[I_N-\Sigma^{-\frac{1}{2}}G_2(G_2'\Sigma^{-1}G_2)^{-1}G_2'\Sigma^{-\frac{1}{2}}]u \\
&= u'[I_N-\Sigma^{-\frac{1}{2}}G_1(G_1'\Sigma^{-1}G_1)^{-1}G_1'\Sigma^{-\frac{1}{2}}]u + (h'u)^2 \\
&= x+\tilde{z}_2^2,
\end{aligned} \tag{A.14}
$$

where $x\sim\chi^2_{N-K}$ and it is independent of $\tilde{z}_2\sim\mathcal{N}(0,1)$. To establish the last equality, we need to show that $h'u=\tilde{z}_2$. Denote by $\boldsymbol{\iota}_{m,i}$ an m-vector with its ith element equal to one and zero elsewhere, and let $\sigma_{i,j}\equiv\mathrm{Cov}[z_i,z_j]=\boldsymbol{\iota}'_{K,i}(G_1'\Sigma^{-1}G_1)^{-1}\boldsymbol{\iota}_{K,j}/\sigma^2_{f,K-1}$. Using the formula for the inverse of a partitioned matrix, we obtain

$$
\begin{aligned}
z_2 &= \frac{1}{\sigma_{f,K-1}}\boldsymbol{\iota}'_{K,2}(G_1'\Sigma^{-1}G_1)^{-1}G_1'\Sigma^{-\frac{1}{2}}u \\
&= \frac{1}{\sigma_{f,K-1}} \\
&\quad\times\frac{\alpha'\Sigma^{-\frac{1}{2}}[I_N-\Sigma^{-\frac{1}{2}}G_2(G_2'\Sigma^{-1}G_2)^{-1}G_2'\Sigma^{-\frac{1}{2}}]u}{\alpha'\Sigma^{-\frac{1}{2}}[I_N-\Sigma^{-\frac{1}{2}}G_2(G_2'\Sigma^{-1}G_2)^{-1}G_2'\Sigma^{-\frac{1}{2}}]\Sigma^{-\frac{1}{2}}\alpha}.
\end{aligned} \tag{A.15}
$$

It follows that

$$
\begin{aligned}
&\sigma_2^2 \\
&= \frac{1}{\sigma^2_{f,K-1}\alpha'\Sigma^{-\frac{1}{2}}[I_N-\Sigma^{-\frac{1}{2}}G_2(G_2'\Sigma^{-1}G_2)^{-1}G_2'\Sigma^{-\frac{1}{2}}]\Sigma^{-\frac{1}{2}}\alpha}
\end{aligned} \tag{A.16}
$$

and $h'u=z_2/\sigma_2=\tilde{z}_2$.

Denote by w_i the ith diagonal element of $(\hat{B}'\hat{\Sigma}^{-1}\hat{B})^{-1}$, $i=1,\ldots,K-1$. Using Eq. (A.11), we have

$$
\begin{aligned}
w_i &\xrightarrow{d} \boldsymbol{\iota}'_{K-1,i}(G_2'\Sigma^{-1}G_2)^{-1}\boldsymbol{\iota}_{K-1,i} \\
&\quad + \frac{\boldsymbol{\iota}'_{K-1,i}(G_2'\Sigma^{-1}G_2)^{-1}G_2'\Sigma^{-\frac{1}{2}}uu'\Sigma^{-\frac{1}{2}}G_2(G_2'\Sigma^{-1}G_2)^{-1}\boldsymbol{\iota}_{K-1,i}}{x+\tilde{z}_2^2} \\
&= \boldsymbol{\iota}'_{K-1,i}(G_2'\Sigma^{-1}G_2)^{-1}\boldsymbol{\iota}_{K-1,i}\left(1+\frac{q_i^2}{x+\tilde{z}_2^2}\right),
\end{aligned} \tag{A.17}
$$

where

$$
q_i = \frac{\boldsymbol{\iota}'_{K-1,i}(G_2'\Sigma^{-1}G_2)^{-1}G_2'\Sigma^{-\frac{1}{2}}u}{[\boldsymbol{\iota}'_{K-1,i}(G_2'\Sigma^{-1}G_2)^{-1}\boldsymbol{\iota}_{K-1,i}]^{\frac{1}{2}}} \sim \mathcal{N}(0,1). \tag{A.18}
$$

Using the fact that $\mathrm{Var}[u]=I_N$ and

$$
(G_1'\Sigma^{-1}G_1)^{-1}G_1'\Sigma^{-1}G_2 = [\boldsymbol{\iota}_{K,1},\ \boldsymbol{\iota}_{K,3},\ldots,\boldsymbol{\iota}_{K,K}], \tag{A.19}
$$

it is straightforward to show that

$$
\begin{aligned}
&\mathrm{Cov}[z_1,q_1] \\
&= \frac{\boldsymbol{\iota}'_{K,1}(G_1'\Sigma^{-1}G_1)^{-1}G_1'\Sigma^{-1}G_2(G_2'\Sigma^{-1}G_2)^{-1}\boldsymbol{\iota}_{K-1,1}}{\sigma_{f,K-1}[\boldsymbol{\iota}'_{K-1,1}(G_2'\Sigma^{-1}G_2)^{-1}\boldsymbol{\iota}_{K-1,1}]^{\frac{1}{2}}} \\
&= [\boldsymbol{\iota}'_{K-1,1}(G_2'\Sigma^{-1}G_2)^{-1}\boldsymbol{\iota}_{K-1,1}/\sigma^2_{f,K-1}]^{\frac{1}{2}}
\end{aligned} \tag{A.20}
$$

and

$$
\begin{aligned}
&\mathrm{Cov}[z_2,q_1] \\
&= \frac{\boldsymbol{\iota}'_{K,2}(G_1'\Sigma^{-1}G_1)^{-1}G_1'\Sigma^{-1}G_2(G_2'\Sigma^{-1}G_2)^{-1}\boldsymbol{\iota}_{K-1,1}}{\sigma_{f,K-1}[\boldsymbol{\iota}'_{K-1,1}(G_2'\Sigma^{-1}G_2)^{-1}\boldsymbol{\iota}_{K-1,1}]^{\frac{1}{2}}} = 0.
\end{aligned} \tag{A.21}
$$

From the formula for the inverse of a partitioned matrix, we have

$$
\begin{aligned}
&\frac{1}{\sigma^2_{f,K-1}}\boldsymbol{\iota}'_{K-1,1}(G_2'\Sigma^{-1}G_2)^{-1}\boldsymbol{\iota}_{K-1,1} \\
&\quad = \sigma_1^2-\frac{\sigma_{1,2}^2}{\sigma_2^2} = \sigma_1^2(1-\rho_{1,2}^2).
\end{aligned} \tag{A.22}
$$

It follows that

$$
\begin{aligned}
&\mathrm{Cov}\left[z_1-\frac{\sigma_{1,2}}{\sigma_2^2}z_2,q_1\right] \\
&= [\boldsymbol{\iota}'_{K-1,1}(G_2'\Sigma^{-1}G_2)^{-1}\boldsymbol{\iota}_{K-1,1}/\sigma^2_{f,K-1}]^{\frac{1}{2}} \\
&= \sigma_1\sqrt{1-\rho_{1,2}^2}.
\end{aligned} \tag{A.23}
$$

Therefore, $z_1-(\sigma_{1,2}/\sigma_2^2)z_2$ is perfectly correlated with q_1, and we can write

$$z_1 = \frac{\sigma_{1,2}}{\sigma_2^2} z_2 + \sqrt{1-\rho_{1,2}^2}\sigma_1 q_1 = \sigma_1\left(\rho_{1,2}\tilde{z}_2 + \sqrt{1-\rho_{1,2}^2} q_1\right). \tag{A.24}$$

Similarly,

$$\begin{aligned} z_{i+1} &= \frac{\sigma_{i+1,2}}{\sigma_2^2} z_2 + \sqrt{1-\rho_{i+1,2}^2}\sigma_{i+1} q_i \\ &= \sigma_{i+1}\left(\rho_{i+1,2}\tilde{z}_2 + \sqrt{1-\rho_{i+1,2}^2} q_i\right), \\ & i = 2, \dots, K-1. \end{aligned} \tag{A.25}$$

Let

$$b_i = \frac{x + \tilde{z}_2^2}{x + \tilde{z}_2^2 + q_i^2}, \quad i = 1, \dots, K-1. \tag{A.26}$$

With the above results, we can now write the limiting distribution of the t -ratios as

$$\begin{aligned} t(\hat{\gamma}_0^{ML}) &\xrightarrow{d} -\frac{z_1 |z_2| b_1^{\frac{1}{2}}}{z_2[\boldsymbol{\iota}'_{K-1,1}(G_2'\Sigma^{-1}G_2)^{-1}\boldsymbol{\iota}_{K-1,1}/\sigma_{f,K-1}^2]^{\frac{1}{2}}} \\ &= -\left(\frac{\rho_{1,2}|\tilde{z}_2|}{\sqrt{1-\rho_{1,2}^2}} + q_1\right) b_1^{\frac{1}{2}} \end{aligned} \tag{A.27}$$

and

$$\begin{aligned} t(\hat{\gamma}_{1,i}^{ML}) &\xrightarrow{d} \frac{\left(\mu_{f,i} - \frac{z_{i+2}}{z_2}\right)|z_2| b_{i+1}^{\frac{1}{2}}}{[\boldsymbol{\iota}'_{K-1,i+1}(G_2'\Sigma^{-1}G_2)^{-1}\boldsymbol{\iota}_{K-1,i+1}/\sigma_{f,K-1}^2]^{\frac{1}{2}}} \\ &= \left(\frac{\frac{\mu_{f,i}\sigma_2}{\sigma_{i+2}} - \rho_{i+2,2}}{\sqrt{1-\rho_{i+2,2}^2}}|\tilde{z}_2| - q_{i+1}\right) b_{i+1}^{\frac{1}{2}}, \\ & i = 1, \dots, K-2. \end{aligned} \tag{A.28}$$

Defining $\bar{Z}_0 = -\left(\frac{\rho_{1,2}|\tilde{z}_2|}{\sqrt{1-\rho_{1,2}^2}} + q_1\right) b_1^{\frac{1}{2}}$ and $\bar{Z}_i = \left(\frac{\frac{\mu_{f,i}\sigma_2}{\sigma_{i+2}} - \rho_{i+2,2}}{\sqrt{1-\rho_{i+2,2}^2}}|\tilde{z}_2| - q_{i+1}\right) b_{i+1}^{\frac{1}{2}}$, for $i = 1, \dots, K-2$, delivers the desired result. This completes the proof of part (a). part (b). Let $\hat{e} = \hat{\mu}_R - 1_N\hat{\gamma}_0^{ML} - \hat{\beta}\hat{\gamma}_1^{ML}$, and note that the fitted expected returns can be rewritten as

$$\begin{aligned} \hat{\mu}_R^{ML} &= 1_N\hat{\gamma}_0^{ML} + \hat{\beta}^{ML}\hat{\gamma}_1^{ML} \\ &= 1_N\hat{\gamma}_0^{ML} + \hat{\beta}\hat{\gamma}_1^{ML} + \hat{e}\frac{\hat{\gamma}_1^{ML\prime}\hat{V}_f^{-1}\hat{\gamma}_1^{ML}}{1+\hat{\gamma}_1^{ML\prime}\hat{V}_f^{-1}\hat{\gamma}_1^{ML}} \\ &= \hat{\mu}_R - \hat{e} + \hat{e}\frac{\hat{\gamma}_1^{ML\prime}\hat{V}_f^{-1}\hat{\gamma}_1^{ML}}{1+\hat{\gamma}_1^{ML\prime}\hat{V}_f^{-1}\hat{\gamma}_1^{ML}} \\ &= \hat{\mu}_R - \hat{e}\frac{1}{1+\hat{\gamma}_1^{ML\prime}\hat{V}_f^{-1}\hat{\gamma}_1^{ML}}. \end{aligned} \tag{A.29}$$

Using the result from Auxiliary Lemma 1 that $\hat{\gamma}_{1,i}^{ML} = O_p(1)$ for $i = 1, \dots, K-2$ and $\hat{\gamma}_{1,K-1}^{ML} = O_p(T^{\frac{1}{2}})$, we have $\hat{\mu}_R^{ML} - \hat{\mu}_R \xrightarrow{p} 0_N$ and

$$R^2 = \text{Corr}(\hat{\mu}_R^{ML}, \hat{\mu}_R)^2 \xrightarrow{p} 1 \tag{A.30}$$

as $T \to \infty$. This completes the proof of part (b).

References

Ahn, S.C., Horenstein, A.R., Wang, N., 2018. Beta matrix and common factors in stock returns. J. Financial Quant. Anal. 53, 1417–1440.

Almeida, C., Garcia, R., 2012. Assessing misspecified asset pricing models with empirical likelihood estimators. J. Economet. 170, 519–537.

Almeida, C., Garcia, R., 2017. Economic implications of nonlinear pricing kernels. Manag. Sci. 63, 3361–3380.

Barillas, F., Shanken, J., 2017. Which alpha? Rev. Financial Stud. 30, 1316–1338.

Barillas, F., Shanken, J., 2018. Comparing asset pricing models. J. Finance 73, 715–754.

Bekker, P., Dobbelstein, P., Wansbeek, T., 1996. The APT model as reduced-rank regression. J. Bus. Econ. Stat. 14, 199–202.

Bryzgalova, S., 2016. Spurious factors in linear asset pricing models. Stanford University, Stanford, CA Unpublished working paper.

Burnside, C., 2016. Identification and inference in linear stochastic discount factor models with excess returns. J. Financial Economet. 14, 295–330.

Campbell, J.Y., 1996. Understanding risk and return. J. Polit. Econ. 104, 298–345.

Carhart, M.M., 1997. On persistence in mutual fund performance. J. Finance 52, 57–82.

Cragg, J.G., Donald, S.G., 1997. Inferring the rank of a matrix. J. Economet. 76, 223–250.

Fama, E.F., French, K.R., 1993. Common risk factors in the returns on stocks and bonds. J. Financial Econ. 33, 3–56.

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

Feng, G., Giglio, S., Xiu, D., 2017. Taming the factor zoo. University of Chicago, Chicago, IL Unpublished working paper.

Ghosh, A., Julliard, C., Taylor, A.P., 2017. What is the consumption-CAPM missing? An information-theoretic framework for the analysis of asset pricing models. Rev. Financial Stud. 30, 442–504.

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

Gospodinov, N., Kan, R., Robotti, C., 2014b. Spurious inference in unidentified asset-pricing models. Unpublished working paper. Federal Reserve Bank of Atlanta, GA.

Gospodinov, N., Kan, R., Robotti, C., 2017a. Spurious inference in reduced-rank asset-pricing models. Econometrica 85, 1613–1628.

Gospodinov, N., Kan, R., Robotti, C., 2017b. Too good to be true? fallacies in evaluating risk factor models. Unpublished working paper. Federal Reserve Bank of Atlanta, GA.

Gospodinov, N., Kan, R., Robotti, C., 2018. Asymptotic variance approximations for invariant estimators in uncertain asset-pricing models. Econom. Rev. 37, 695–718.

Gourieroux, C., Monfort, A., Trognon, A., 1985. Moindres carrés asymptotiques. Ann. de l'INSEE 58, 91–121.

Harvey, C.R., Liu, Y., Zhu, H., 2016. … and the cross-section of expected returns. Rev. Financial Stud. 29, 5–68.

Hou, K., Xue, C., Zhang, L., 2015. Digesting anomalies: an investment approach. Rev. Financial Stud. 28, 650–705.

Jagannathan, R., Wang, Z., 1996. The conditional CAPM and the cross-section of expected returns. J. Finance 51, 3–53.

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

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

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

Kleibergen, F., Paap, R., 2006. Generalized reduced rank tests using the singular value decomposition. J. Econom. 133, 97–126.

Kleibergen, F., Zhan, Z., 2015. Unexplained factors and their effects on second pass r-squared's. J. Econom. 189, 101–116.

Kodde, D.A., Palm, F.C., Pfann, G.A., 1990. Asymptotic least-squares estimation efficiency considerations and applications. J. Appl. Econom. 5, 229–243.

Lettau, M., Ludvigson, S.C., 2001. Resurrecting the (c)CAPM: a cross-sectional test when risk premia are time-varying. J. Polit. Econ. 109, 1238–1287.

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

Manresa, E., Peñaranda, F., Sentana, E., 2017. Empirical evaluation of overspecified asset pricing models. Unpublished working paper. Center for Monetary and Financial Studies, Madrid, Spain.

Merton, R.C., 1973. An intertemporal capital asset pricing model. Econometrica 41, 867–887.

Pastor, L., Stambaugh, R.F., 2003. Liquidity risk and expected stock returns. J. Polit. Econ. 111, 642–685.
Peñaranda, F., Sentana, E., 2015. A unifying approach to the empirical evaluation of asset pricing models. Rev. Econ. Stat. 97, 412–435.
Petkova, R., 2006. Do the Fama-French factors proxy for innovations in predictive variables? J. Finance 61, 581–612.
Shanken, J., 1985. Multivariate tests of the zero-beta CAPM. J. Financial Econ. 14, 327–348.
Shanken, J., 1992. On the estimation of beta-pricing models. Rev. Financial Stud. 5, 1–33.
Shanken, J., Zhou, G., 2007. Estimating and testing beta pricing models: alternative methods and their performance in simulations. J. Financial Econ. 84, 40–86.
Vissing-Jørgensen, A., Attanasio, O.P., 2003. Stock-market participation, intertemporal substitution, and risk-aversion. Am. Econ. Rev. Pap. Proc. 93, 383–391.
White, H.L., 1994. Estimation, Inference, and Specification Analysis. Cambridge University Press, New York.
Wright, J.H., 2003. Detecting lack of identification in GMM. Econom. Theory 19, 322–330.
Yogo, M., 2006. A consumption-based explanation of expected stock returns. J. Finance 61, 539–580.
Zhou, G., 1995. Small sample rank tests with applications to asset pricing. J. Empir. Finance 2, 71–93.