

On the statistical properties of tests of parameter restrictions in beta-pricing models with a large number of assets*

Amedeo Andriollo^a, Cesare Robotti^{b,†} and Giulio Rossetti^b

^a*University of Warwick–Department of Economics, Coventry, CV4 7AL, United Kingdom*

^b*Warwick Business School, Coventry, CV4 7AL, United Kingdom*

January 24, 2024

ABSTRACT

We study the size and power properties of t -tests of parameter restrictions for newly-designed methods that aim at reliably estimating risk premia in linear asset pricing models when the cross-sectional dimension is large. By simulating a variety of empirically calibrated data generating processes for sample sizes that are typically encountered in empirical work, we evaluate the finite-sample performance of the test statistics for scenarios where the factor structure is (i) strong and pervasive; (ii) spurious; (iii) weak/semi-strong and pervasive; (iv) weak/semi-strong and not pervasive; and (v) sparse. PCA-based methods such as those of Lettau and Pelger (2020), Giglio and Xiu (2021), and Giglio et al. (2022) work best when the factors are strong and pervasive, and they continue to exhibit good finite-sample properties when the factors are spurious. However, when the factor structure is semi-strong and pervasive, the split-sample estimator of Anatolyev and Mikusheva (2021) performs substantially better than the PCA-based estimators listed above. In the case of sparse loadings or when the factors are semi-strong and not pervasive, none of the candidate methods displays satisfactory finite-sample properties.

Keywords: PCA; Risk premium; Factor models; Two-pass methodology; Strong and weak factors; Spurious factors; Local factors; Asset pricing tests.

JEL classification: C12, C38, C53, G12.

*We gratefully acknowledge comments and suggestions from Eric Renault, and participants at Warwick Business School, the Warwick Macro and International Economic Seminar, the Warwick-Turing Economics Data Science Workshop, the 15th annual SoFiE conference. We also thank Stefano Giglio, Dacheng Xiu, Anna Mikusheva, and Elisa Ossola for sharing their codes.

1 Introduction

Based on the theoretical arguments provided by Sharpe (1964), Merton (1973), and Ross (1976), linear asset pricing models have laid the groundwork for understanding the relationship between test assets' expected returns and a set of risk factors - i.e, traded or nontraded factors that aim at capturing the systematic risk that investors are subject to and cannot diversify away.

The most popular methodology for estimating risk premia is considered to be the two-step procedure of Fama and MacBeth (1973), which still remains the standard tool for cross-sectional asset pricing. Despite its simplicity, this methodology potentially suffers from problems that have been widely documented in the asset pricing literature. These shortcomings can be broadly grouped into two categories. First, when factors are weak or completely spurious, the risk premia from the second-pass method of Fama and MacBeth (1973) are undefined. (See Kan and Zhang (1999), Kleibergen (2009) and Gospodinov et al. (2014), among others) Second, even if the factors were to be strong and pervasive, there could still be substantial (global) model misspecification due, for example, to the omission of important factors in the model. This type of model misspecification would not vanish asymptotically, and it would need to be addressed by robustifying the risk premium inference against potential misspecification. (See Kan et al. (2013)) In the meanwhile, the empirical asset pricing literature has come up with hundreds of factors (see the factor zoo of Harvey et al. (2016)) that are deemed to be priced on a variety of test assets. This lends support to the view that not all models can be correct at the same time and that model misspecification needs to be taken seriously into account.

In light of these two challenges, and concurrently with easier access to bigger datasets over time, new methods have been recently proposed for estimating risk premia when the number of test assets is large. Traditionally, researchers have mostly focused on a reduced number of test assets/portfolios to infer whether a factor is priced in the cross-section of asset expected returns. From here, most of the emphasis, from a statistical standpoint, has been on large- T asymptotics. More recently, several researchers have proposed large- T and large- N methods to accommodate a large number of test asset returns in the analysis. (See, for example, Gagliardini et al. (2016)) In addition to factor weakness and fixed model misspecification, accommodating a large cross-sectional dimension in the analysis represents a major challenge for financial econometricians.

In this paper, we consider five recent advances in large- N asset pricing: (i) the two-pass methodology of Gagliardini et al. (2016); (ii) the three-step procedure of Giglio and Xiu (2021); (iii) the three-step method of Lettau and Pelger (2020); (iv) the four-split procedure of Anatolyev and Mikusheva (2021); and (v) the supervised-PCA of Giglio et al. (2022). Our objective is to investigate the finite-sample performance of these methods by examining the size and power properties of the t -tests of zero risk premia for realistic sample sizes. In our large-scale Monte Carlo simulations, we consider a variety of empirically relevant scenarios with a particular focus on the following cases: (i) all factors are strong; (ii) a spurious factor is included in the data generating process; (iii) the strength of the factors decreases; (iv) the factor structure is sparse; and (v) the strength and composition of the factors vary.

Our contribution to the literature is threefold. To our knowledge, we are the first to provide a comprehensive comparison of the statistical properties of the aforementioned methodologies with respect to a plethora of relevant scenarios. In other words, we analyze the statistical properties of the t -tests associated with the various risk premium estimators that have been proposed in the large- N literature. Particular emphasis is put on weak identification, which might undermine the validity of standard statistical inference.

Second, differently from existing studies, we implement our Monte Carlo simulations by keeping as close as possible to the actual features of the data. In doing so, we develop a general procedure that can be easily applied to large data sets and generate artificial returns whose (large- N) cross-sectional structure mimics the one observed in the data. In other words, we are careful in preserving the factor structure of the residuals, which is often ignored in Monte Carlo designs that aim at evaluating the statistical properties of the proposed estimators.

Third, we give researchers some practical recommendations for reliable risk premium inference. In fact, our simulations indicate that no estimator enjoys satisfactory statistical properties in all of the alternative scenarios that we consider. Therefore, we argue that the choice of a specific estimator should be based on the relevant scenario at hand. In particular, we show that PCA-based estimators perform well when the return factor structure is strong and pervasive (e.g., market factor) and when the proposed factors are spurious. On the other hand, when the factors that drive returns are semi-strong and pervasive (e.g., consumption factor), the split-sample method of Anatolyev and Mikusheva (2021) seems to provide a rather reliable (although still not ideal) inference. Finally,

we show that when returns are generated from sparse factors (i.e., factors associated with sparse loadings) and when the strength and composition of factors vary heterogeneously in a large panel of asset returns, none of the estimators analyzed in this study exhibit reliable statistical properties in terms of size and power of the t -tests.

The rest of the paper is organized as follows. Section 2 describes the various large- N estimators we consider in the paper. Section 3 reports a detailed description of realistic scenarios that a researcher might encounter when dealing with the estimation of factors' risk premia, and it provides details on how we generate artificial returns. This section also highlights conditions under which lack of identification arises and renders standard tools invalid. Our simulation results are in Section 4. Section 5 addresses a few issues that arise in the setting of Giglio and Xiu (2021). Section 6 concludes. Additional simulation results are provided in the Appendix and an online Appendix.

2 A Review of Recent Methods

This section provides an overview of the strategies that have been proposed in the asset pricing literature to estimate risk premia when we have a large cross-section of assets. In section 2.1 we discuss three methodologies in which observable factors are proxies for the unobservable latent drivers of return. Next, section 2.2 discusses two approaches in which the factor structure of return is observable.

2.1 PCA-based methods: Latent and observable factors

2.1.1 Three-pass methods: Giglio and Xiu (2021) and Giglio et al. (2022)

Giglio and Xiu (2021) and Giglio et al. (2022) (henceforth GX and SPCA) consider a linear factor model with p latent factors:

$$r_t = \alpha + \beta\gamma + \beta v_t + u_t, \quad \mathbb{E}[v_t] = \mathbb{E}[u_t] = \mathbb{E}[v_t u_t'] = 0 \quad (1)$$

where r_t are returns on N assets in excess of the risk-free rate, β are the factor loadings, v_t are the innovations of the p latent factors, u_t are idiosyncratic errors, α are deviations from the model, and γ is the vector of risk premia. If the model is correct, then $\alpha = 0$ and any unit-beta portfolio

with respect to factor v_p delivers a risk premium of γ_p .¹

$$\mathbb{E}[r_t] = \alpha + \beta\gamma \tag{2}$$

$$= 0 + \beta\gamma \tag{3}$$

The set of d observable factors g_t , whose risk premia are the object of interest, are assumed to be proxies for the latent unobservable factors. Observable and latent factors are related as follows:

$$g_t = \delta + \eta v_t + w_t, \quad \mathbb{E}[w_t] = \mathbb{E}[w_t v_t'] = \mathbb{E}[w_t u_t'] = 0, \tag{4}$$

where η characterizes the relation between the latent and observable factors, and w_t is the measurement error in g_t .²

Equation 4 represents the projection of observable factors onto the span of latent factors, augmented with a constant and resembles the usual factor mimicking projection that is used to build mimicking portfolios of nontraded factors (see Huberman et al. (1987)). The risk premium on the observable factors g_t implied by Eqs. (1)-(4) corresponds to $\gamma_g = \eta\gamma$.

The main challenge posed by this framework is the identification of the unobservable factors in Eq.(1). These latent factors need to (i) have sufficient time variation to correlate with the test assets (so that $\beta'\beta$ is not rank deficient); (ii) adequately explain the cross-section of the underlying test assets; and (iii) be distinguishable from the error terms u_t . While the first two features are standard in the asset pricing literature, the last one is inherited from the PCA literature.

GX propose to estimate the latent factor space via PCA. In this regard, the latent factors v_t are the estimated principal components extracted from the $T \times T$ covariance matrix of asset returns (alternatively, one can estimate principal components from the cross-section covariance matrix. The two approaches take different normalizations but yield the same risk premia). Once the latent factors are estimated, a two-step Fama-Macbeth regression is used to estimate the factor loadings β and the risk premia of the unobservable factors γ (see Appendix A.1 for a detailed description of GX three-step procedure).

Under fairly common assumptions (see Bai (2003)), the estimated latent factors are consistent estimates (up to a rotation) for their population counterparts as long as $N, T \rightarrow \infty$. Such latent

¹GX provide an alternative set of results in the Online Appendix when the cross-sectional pricing errors α are i.i.d., independent of β , u , and v , with a standard deviation $\sigma_\alpha > 0$ and a finite fourth moment. In Section 3.3, we briefly discuss the case of mispricing.

²In fact, $\hat{g}_t = \hat{\delta} + \hat{\eta}\hat{v}_t$, is defined as the denoised factor in GX.

factors or PCs are the unique factors in population that price the test assets r_t . In other words, if all the relevant factors can be uncovered by PCA, then the methodology can successfully eliminate strong missing factors (Anatolyev and Mikusheva (2021)). Although their results are derived under similar if not weaker assumptions than those in Bai (2003), they crucially depend on the pervasive condition for a factor model. In that sense, the identification assumption that weaker and less pervasive factors should not be priced is in clear contrast with the empirical findings of Lettau and Pelger (2020).

In light of these considerations, we can think of the GX’s strategy as an *identification strategy* of the latent factor model defined from the restrictions on the error terms’ moments. These restrictions allow us to separate factors from idiosyncratic noise and guarantee a consistent estimator for a candidate latent factor model which might not be the true one, as the distinction between signal and noise is between unobserved quantities.

The crucial assumption posed on the strength of the latent factors is discussed in Giglio et al. (2022), wherein an alternative approach is proposed to accommodate for a weak (asymptotically sparse) factor structure to characterize the latent asset pricing formulation in Eq. (1). They show that if $N/\lambda_{\min}(\beta'\beta)T \not\rightarrow 0$, the majority of the estimators proposed in the existing literature may fail to produce a consistent estimation of the risk premium, as they do not potentially recover the latent factor structure under such circumstances.

To address this issue, Giglio et al. (2022) propose to estimate the risk premium through supervised PCA (SPCA). Before extracting the latent factors, a test asset selection procedure is performed. The first step of the proposed methodology is about selecting an appropriate subset of assets according to a specific metric, which, in this case, is presumed to be the maximal correlation/covariance with the considered set of observed factors g_t . After this screening procedure, Giglio and Xiu (2021)’s three-pass method is applied with a single principal component. Before iterating again on the first step, the test assets and the observable factors are residualized with respect to the explained part of the principal component. Finally, when at the k^{th} iteration of the procedure a stopping rule is met, the risk premium of the observed factor is estimated as the sum of the resulting k risk premia from the various iterations.³ In practice, the implementation proposed by Giglio et al. (2022) is based on a grid search of the optimal tuning parameters through cross-

³See Appendix A.2 for a thorough description of the SPCA method.

validation by considering the out-of-sample R^2 of the mimicking portfolio Eq. (4): “[We] compute the weights of the hedging portfolio built by SPCA using the training data only, and calculate the mean-squared-error of hedging g_t over the validation period using that portfolio [...] We apply this criterion to pick q using cross-validation (CV) within the training sample”. Note that the estimation of a parameter based on a hedging strategy might not be the optimal value one would estimate for pricing. In fact, Antoine et al. (2020) argue that “since all models are misspecified [...] one can only elicit a pseudo-true SDF. Its definition can only be objective-driven; the elicited pseudo-true value [...] is the one that does the best job in respect to a specific application like pricing, hedging, forecasting, explaining, etc. ”.

Furthermore, while in Giglio and Xiu (2021) the estimation of the parameters of the latent model does not hinge on the observable factors, in Giglio et al. (2022) the estimation of the latent factor model crucially depends on g_t . The parameters of the latent factor model are functions of the PCs, and their extraction depends on the considered test assets, which in turn are selected based on the sample correlation/covariance between r_t and g_t . It is then clear that if we were to consider two observable factors, say $g_{1,t}$ and $g_{2,t}$, that are not identical and correlate differently with the test assets, we would generally have different estimates of the risk premium on $g_{1,t}$ when the set of observable factors consists either of $g_{1,t}$ alone or of $\{g_{1,t}, g_{2,t}\}$. In the first case, the shrinkage is uniquely determined by the correlation between $g_{1,t}$ and r_t , while in the second case, it is based on the maximum between the two correlations - i.e., the test asset i is retained if $\max\{corr[g_{1,t}, r_{i,t}], corr[g_{2,t}, r_{i,t}]\} \geq threshold$.

Thus, differently from Giglio and Xiu (2021), SPCA may suffer from the omitted factor problem in the risk premium analysis. Moreover, for the asymptotically normal behavior of the estimator, Giglio et al. (2022) need to assume that the number of variables in the set of observable factors g_t must be at least equal to the (unknown) number of true factors. For the identification of the latent factor model, it is then of the utmost importance to determine which and how many observable factors should be included in the analysis. Potentially, we deal with a plethora of asset pricing models, with corresponding risk premia that are sensitive to the chosen specification of the model.

Since Giglio et al. (2022) assume the existence of a subset of assets where the factors are strong, the possibility of weak factors with varying rates of strength affecting the factor structure of returns, as discussed for instance in Freyaldenhoven (2021) is ruled out.

From a more technical standpoint, it is important to note that the assumptions made to ensure the favourable asymptotic behavior of the proposed estimator may not be considered mild. Firstly, in order to incorporate the three-step methodology of Giglio and Xiu (2021) in their strategy, they assume that the subset of test assets for which the latent factor is strong is non-random.

Secondly, to ensure consistency in the asset selection procedure, they assume that the proposed selection algorithm consistently recovers, at each step, the uniquely identified subset of assets (and thus the factors).

Lastly, the assumptions in place bound the trade-off in the asymptotic rate between the total number of assets N and the selected number of assets N_0 : N_0 needs to explode fast enough to guarantee consistency of the estimated principal components but slow enough to guarantee the correct subset of assets is selected.

2.1.2 Three-pass methods: Lettau and Pelger (2020)

Lettau and Pelger (2020) argue that factors with large Sharpe-ratios but with a relatively small variance-covariance structure are potentially not recovered by conventional PCA. In fact, principal components are usually extracted from the covariance (or alternatively, from the correlation) matrix of returns, therefore information contained in the mean is ignored. Given that conventional PCA often performs poorly and in some cases structure in the first moments of the data is motivated by economic theory, they propose to combine the first and second moments in a unique objective function to estimate unobservable factors and show that including this information improves the estimation of such latent factors. The intuition stems from asset pricing theory, which suggests that priced risk factors should simultaneously show explanatory power in both the time-series and the cross section (Pukthuanthong et al. (2019)).

Denote by R the $T \times N$ matrix of excess returns and \bar{R} the $N \times 1$ vector of average excess returns. The estimation strategy consists in recovering the latent factors, $\{f_t\}$, by applying PCA to the covariance matrix of the returns with overweighted mean:

$$\frac{1}{T}R^\top R + \mu\bar{R}\bar{R}^\top, \tag{5}$$

where $\mu \in [-1, \infty)$ is a tuning parameter. Lettau and Pelger (2020) show that this is equivalent to

minimizing the following objective function:

$$\min_{\{\beta_i\}, f_t} \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T (\tilde{r}_{i,t} - \beta_i v_t)^2 + (1 + \mu) \frac{1}{N} \sum_{i=1}^N (\bar{r}_i - \beta_i \bar{f})^2, \quad (6)$$

where the first term is related to the unexplained time variation whilst the second one to the (average) cross-sectional pricing errors, $\{N^{-1} \sum_{i=1}^N \alpha_i\}$, with $\tilde{r}_t = r_t - \bar{r}$ and $v_t = f_t - \bar{f}$ being the demeaned returns and factors, respectively. When $\mu = -1$, the recovered latent factors coincide with the ones of Giglio and Xiu (2021)’s three-pass methodology.

The generalized objective function is as follows:

$$\tilde{Q}^\top R^\top (I + \frac{\mu}{T} \mathbf{1}\mathbf{1}^\top) R \tilde{Q},$$

where the case in Eq. (5) occurs when $\tilde{Q} = I_T$.

In our simulations, we set $\tilde{Q} = I_N$ as the asymptotic results in Lettau and Pelger (2020) rely on this particular choice of the weighting matrix \tilde{Q} . Additionally, we also consider the case where \tilde{Q} is the inverse of the diagonal matrix of standard deviations of the test asset returns.

In fact, as argued in Lettau and Pelger (2020) and Choi and Yang (2022), factors estimated from the correlation matrix are more robust, as they are not influenced by few outliers that display large variances.

Following Anatolyev and Mikusheva (2021) and Giglio et al. (2022), we incorporate the intuition of Lettau and Pelger (2020) in the first step of the procedure of Giglio and Xiu (2021), with the second and third step being unchanged. We denote this strategy as GXLP.

Given the similarity of this estimation procedure (and the asymptotic theory) with the one proposed by Giglio and Xiu (2021), the previous comments on Giglio and Xiu (2021) apply. Moreover, we wish to discuss the following points.

Firstly, the authors do not propose any estimator for the tuning parameter, γ , which is pivotal for the trade-off between making all weak factors detectable or attaining the highest correlation for a particular factor. In particular, the authors emphasize that, on one side, “*if too much weight is given to an uninformative mean, the estimator will pick up some of the non-zero residuals.*” On the other side, “*increasing the signal strength for detecting weak factors becomes more relevant for correlated residuals,*” which are unobservable quantities. This amounts to having a potential candidate model

for the latent space for each $\mu \in [-1, \infty)$. Interestingly, when $\mu \rightarrow \infty$, the estimator can perform well for weak factors if the information from return covariances is dominated by the information about β derived from the expected return $\beta\gamma$. (Giglio et al. (2022) Proposition 4)

Secondly, the authors derive the statistical theory separately for the strong factor and the weak factor scenarios, since in the mixed case, once the strong factors are estimated, “*the residuals from the strong factor model can then be described by a weak factor model.*” (Lettau and Pelger (2020)) Hence, despite the clear implementation, consistent estimation of the latent weak factors heavily relies on the consistent estimation of the number of strong factors, which in general is hard to pin down. (See, e.g., figure 8 in Lettau and Pelger (2020) and figures 1-2 in Anatolyev and Mikusheva (2021))

Lastly, the (high-level) assumptions on the risk exposures β and on the time dependence of the residuals set in place for achieving desirable asymptotic properties are stronger than the common assumptions usually found in the literature.

2.2 Methods with observable factors

2.2.1 Four-split method: Anatolyev and Mikusheva (2021)

Similar to Eq. (1), Anatolyev and Mikusheva (2021) (henceforth AM) consider the linear beta-pricing model

$$r_t = \beta\lambda + \beta(F_t - \mathbb{E}[F_t]) + u_t, \quad (7)$$

where the $p \times 1$ vector of factors, F_t , is observable. The model is assumed to be correctly specified, thus the asset pricing restriction of Eq.2) holds.

To estimate risk premia associated to the observable factors when the number of assets N is large, they provide a methodology that is robust to two scenarios: (i) the observable factors are weak; (ii) there is (sufficiently) strong cross-sectional dependence among the error terms, u_t .

Specifically, they posit that some of the factors may enter the model with a local-to-zero (i.e., of order $O(T^{-1/2})$) risk exposure. Additionally, they impose a latent factor structure on the (unobserved) error terms:

$$u_{i,t} = v_t' \kappa_i + e_{i,t}, \quad \forall i, t \quad (8)$$

for some unobserved factors v_t with loadings κ_i , and where the $e_{i,t}$'s are the “*clean errors*” whose dependence is nearly negligible. The reason behind this assumption is that numerous studies show that the post-estimation residuals usually have a non-trivial factor structure which, however, is thought to carry no compensation for systematic risk.⁴

To estimate the parameters of interest, the authors propose a sample-splitting IV methodology. They argue that if the sources of potential bias coming from the risk exposures (i.e., the betas estimated in the first step of the Fama and MacBeth (1973) procedure) are conditionally independent, then a sub-sample estimate of the exposures can be potentially a valid instrument when estimating risk premia in the cross-sectional regression. In other words, they propose to first split the sample into four distinct, non-overlapping subsets, and then, for each of the four subsamples, estimate the risk exposures using a time-series regression. Once the risk exposures are estimated, they suggest running an IV cross-sectional regression, using the estimated betas as instrument. This procedure is repeated for each subsample in a circular fashion, to get four estimates of the risk premia. Lastly, the average of these four risk premia estimates represents the overall risk premia obtained with this estimation strategy (We provide a detailed description of the algorithm in Appendix A.4).

Anatolyev and Mikusheva (2021) assume the missing factor structure to be strong (and therefore stronger than the potential weak factors) and not to carry any mispricing. Under these assumptions, the observable factors are the unique drivers of the risk premia. The procedure, therefore, is affected by the usual omitted variable problem encountered in standard asset pricing models. Moreover, the validity of the four-split estimation method is jeopardized when the observable factors are completely spurious, as they assume that the observable factors are drifting to, but not zero in population. Lastly, the asymptotic behavior of the estimator relies on two additional parameters: the number of unobserved strong factors in the error factor structure, and the weighting matrix used in the IV regression.

⁴“One may wonder whether [...] the pricing model is misspecified. The answer is “no”; the linear factor pricing model describes the expectations of excess returns, while the factor structure in the error is related to their covariances or co-movements [...] Not all co-movements of returns must carry non-zero risk premia; those co-movements can be placed in the error term without causing misspecification of the pricing model” (Anatolyev and Mikusheva (2021))

2.2.2 Two-pass method of Gagliardini et al. (2016)

Different from the previous models, Gagliardini et al. (2016) (GOS) consider the following conditional factor model for the excess return, $r_t(\gamma)$, of asset $\gamma \in [0, 1]$:

$$r_t(\gamma) = \beta_t(\gamma)' \lambda_t + \beta_t(\gamma)' (F_t - \mathbb{E}[F_t | \mathcal{F}_{t-1}]) + \varepsilon_t(\gamma), \quad \forall t \in \mathbb{Z}, \quad (9)$$

where the filtration \mathcal{F}_t entails the flow of information available to investors up to date t , F_t is the $K \times 1$ vector of observable factors, $\beta_t(\gamma)$ are the factor sensitivities (or time-varying loadings), and $\varepsilon_t(\gamma)$ are the error terms. The random variable $[\lambda_t(\gamma)]_{\gamma \in [0,1]}$ is the $K \times 1$ vector of the conditional risk premia.⁵

To estimate the parameters of interest, they further characterize the functional specification of the model coefficients, which amounts to expressing the conditioning information with respect to a set of instruments (see their Assumptions FS. 1-2). Since our analysis focuses on unconditional risk premia,

$$\mathbb{E}[r_t(\gamma)] = \beta(\gamma)' \lambda,$$

the estimation procedure reduces to two steps that are similar to the ones in Fama and MacBeth (1973). The first stage consists of an OLS time series regression of excess returns on the factors. The second stage consists of a cross-sectional weighted least squares (WLS) regression (augmented with the factor mean) with respect to a trimming device, that is, an indicator function that guarantees the avoidance of ill-conditioning. The trimming device depends on a function of the eigenvalues corresponding to the covariance matrix of the first-pass estimates.

Note that the focus of Gagliardini et al. (2016) is on time-varying risk premia with respect to a filtration generated by a set of instruments. It is clear that the performance of the procedure in finite sample, depending on the identification strength, relies on the choice of such instruments. Nonetheless, when the instrument set is empty, this risk premium estimator could be used in a large- N environment as a potential benchmark against which the performance of the various methods described in previous sections could be compared. In fact, under this setting, the methodology of Gagliardini et al. (2016) can be seen as a large- N generalization of the standard Fama-Macbeth regression. Second, the properties of this estimation strategy when spurious factors are present in the analysis have not been investigated, which calls for a careful examination.

⁵For further details, please refer to Section 2.1 of Gagliardini et al. (2016).

3 Data Generating Processes

This section presents the Data Generating Processes (DGPs) we use to characterize the statistical properties of the t -tests of parameter restrictions associated with the estimation strategies discussed in Section 2.

In Section 3.1, we begin with a scenario that follows closely the simulations in Giglio and Xiu (2021), where we have a relatively large number of assets and the factor structure is strong. In Section 3.2, we consider scenarios in which the factor structure might not be strong. The analysis is then carried out for the weak, spurious, and semi-strong factor cases. Notably, throughout these two sections, we assume that the underlying model is correctly specified. Finally, in Section 3.3, we briefly discuss a more general framework in which mispricing is allowed for in the model.

Additionally, in Appendix C, we study all the aforementioned scenarios calibrating the DGPs to align with the methodology of Anatolyev and Mikusheva (2021).

3.1 Strong and pervasive factors in a large- N setting

We start our analysis with the benchmark case in which the factor structure of our simulated returns is strong and pervasive. In this setting, all estimators should perform relatively well, in particular when the sample size is sufficiently large.

We simulate a five-factor DGP based on the five-factor model of Fama and French (2015). We calibrate population parameters following two different approaches. The first approach mirrors the methodology outlined in Giglio and Xiu (2021), whereas the second one follows Anatolyev and Mikusheva (2021) and is described in Appendix C.

In the first calibration strategy, the returns are generated according to Eq. (1) where the latent factors are calibrated to match the denoised⁶ version of the five Fama and French (2015) factors. The factors (v_t^{FF5}) are latent and we observe a noisy version of them (as specified in Eq. (4)). We calibrate the η and the measurement error w_t as in Giglio and Xiu (2021).

Given a number \check{N} of test assets returns $r_t^{(\check{N})}$, the realizations of the $\check{N} \times 1$ simulated returns

⁶See Giglio and Xiu (2021) for details about the denoising procedure. In our simulation setting, we generate artificial returns with different degrees of denoising, i.e., $p = \{5, 7, 9\}$. However, we report the results only for the case when the number of principal components used to denoise factors is five. The complete set of simulation results is available upon request.

r_t^\diamond , together with the 5×1 simulated factors v_t^\diamond , and the 5×1 simulated measurement errors w_t^\diamond , are generated from a multivariate normal as follows:

$$\begin{pmatrix} r_t^\diamond \\ v_t^\diamond \\ w_t^\diamond \end{pmatrix} \sim \mathcal{N} \left(\begin{pmatrix} \bar{r} \\ 0_{(2 \times 5) \times 1} \end{pmatrix}, \begin{pmatrix} \widehat{\Sigma}^r & \widehat{\Sigma}^{r,v} & 0 \\ \widehat{\Sigma}^{v,r} & \widehat{\Sigma}^v & 0 \\ 0 & 0 & \widehat{\Sigma}^w \end{pmatrix} \right), \quad (10)$$

where \bar{r} is the $\check{N} \times 1$ vector of sample average returns, $\widehat{\Sigma}^r$ is the $\check{N} \times \check{N}$ sample covariance matrix of returns, and $\widehat{\Sigma}^v$ and $\widehat{\Sigma}^w$ are the 5×5 sample covariance matrices of the empirically calibrated denoised factors v_t^{FF5} and measurement errors. $\widehat{\Sigma}^{r,v}$ is the $\check{N} \times 5$ matrix of sample covariances between the returns and v_t^{FF5} . The simulated observable factors (i.e., noisy proxies of the true priced factors) are $g_t^\diamond = v_t^\diamond + w_t^\diamond$. The other relevant quantities for the model in Eqs. (1)-(4), i.e., $\{\check{\beta}, \check{\gamma}, \hat{u}_t^{(\check{N})}\}$, are then obtained as

$$\begin{aligned} \check{\beta} &= (\widehat{\Sigma}^v)^{-1} (\widehat{\Sigma}^{r,v})', & \check{\gamma} &= (\check{\beta}' \check{\beta})^{-1} \check{\beta}' \bar{r} \\ \hat{u}_t^{(\check{N})} &= r_t - \check{\beta} \check{\gamma} - \check{\beta} v_t^{FF5}. \end{aligned}$$

These quantities are taken as the population ones in the various simulation designs. In the experiments, we consider $\check{T} = \{100, 200, 400, 600, 1200\}$ and $\check{N} = \{50, 100, 200, 400, 600\}$.

This design allows us to differentiate our analysis from the simulation studies that were proposed in the papers mentioned in Section 3. We are not imposing particular restrictions on the second moments, such as zero non-diagonal entries or artificial factor structures (e.g., Giglio and Xiu (2021) and Lettau and Pelger (2020)). Instead, we aim at maintaining the cross-sectional covariance structure of the error terms *as close as possible* to the empirical one. This is an important feature to account for while simulating returns for two main reasons. First, there is strong evidence of high cross-sectional dependence, or strong factor structure, in the residuals (See Kleibergen and Zhan (2015) and Anatolyev and Mikusheva (2021)). Second, standard PCA might not properly take into account the heteroskedastic feature of the data. (See, for comparison, the heteroPCA method of Zhang et al. (2022)).

To preserve the heterogeneous factor structure of the original dataset, we always include bonds and currencies portfolios in the selected asset returns when calibrating the parameters.

Finally, in our simulations, we maintain the condition $\check{N} \leq \check{T}$ when generating our artificial returns. For instance, when $\check{N} = 400$, we then generate samples corresponding to $\check{T} =$

{400, 600, 1200}.⁷ In the interest of space, the simulation results for the small- N cases are relegated to the Online Appendix.

3.2 Spurious, weak, and semi-strong factors

Most of the large- N methods described above should be robust to a weak factor structure. To study their performance in terms of size and power, we need to first define the various DGPs.

Factors' weakness arises when the betas are small (or zero). In this case, time variation is not sufficient to ensure the identification of the factor risk premium. Nonetheless, on one hand, Anatolyev and Mikusheva (2021) suggest to use the betas as instruments, and on the other hand, the methodology of Giglio and Xiu (2021) hinges upon a Fama-Macbeth procedure applied to the latent factor model. An exhaustive theoretical discussion can be found in Giglio et al. (2022), in which it is formally shown that some of the estimators presented in the previous section are inconsistent in a sparse-beta scenario.

Following Gospodinov et al. (2019), we define the observed factor $g_{k,t}$ to be *spurious* if the population betas associated with $g_{k,t}$ are zero, i.e., $\beta_k = \text{cov}(g_{k,t}, r_t) / \text{var}(g_{k,t}) = 0_{n \times 1}$. As for *weak* factors, we need to consider a variety of cases, all related to different specifications of a convergence rate. (Freyaldenhoven (2021)):

- Pervasive/global weak or semi-strong loadings (Onatski (2012)), $\beta = O_p(N^{-1/2})$. While Onatski (2012) discusses explicitly a single rate of convergence, Freyaldenhoven (2021) generalizes the argument to various rates of convergence. In particular, if $\beta = O_p(N^{-\alpha})$, with $\alpha < 1/2$, then we are in the case of semi-strong factors (or loadings).⁸
- Locally strong loadings (Freyaldenhoven (2021) and Giglio et al. (2022)): $\beta_i = \{\beta_{i,0}, \beta_{i,\vartheta}\}$, with $[\beta_{i,0}]_i = 0$ and $[\beta_{i,\vartheta}]_i = O_p(1)$. The betas are sparse, since $[\beta_{i,0}]_i = 0$, but, if the model is restricted to the assets corresponding to the exposure β_ϑ , such factors are strong. In other words, among all the assets, there exists a subset of test assets whose factor structure is strong.

⁷In untabulated work, we also run simulations when $\check{T} < \check{N}$. The results are very similar except for extreme cases such as $\check{T} = 100, \check{N} = 600$. The results are available from the authors upon request.

⁸See Connor and Korajczyk (2022) and Antoine et al. (2020).

In summary, the problem of identifying risk premia could be traced back to four scenarios: *spurious factors*, *weak factors* (or loadings), *sparse* loadings, and *semi-strong* factors (or loadings).

In our simulations, we consider the following scenarios where returns are generated under a weak factor structure.

1. SPURIOUS FACTORS

We consider the same scenario of Section 3.1, except that we now augment the set of observable factors \check{g}_t with a factor that is uncorrelated with the test asset returns. In practice, we consider $g_{6,t}^\diamond$ such that

$$\mathbb{E}[g_{6,t}^\diamond r_t^{(\check{N})}] = 0_{\check{N}}, \quad g_{6,t}^\diamond \sim \mathcal{N}(0, \check{\sigma}^2), \quad \check{\sigma} = \hat{\sigma}_{MKT} \times \{1/15, 1/3, 4/3, 8/3\},$$

where $\hat{\sigma}_{MKT}$ is the sample variance of the MKT factor estimated from the data. In other words, we generate an independent factor, with variance that is a multiple of the market variance. The rest remains unchanged with respect to Section 3.1, as the spurious factor does not enter the DGP.

2. WEAK/SEMI-STRONG LOADINGS

Our starting point for this scenario is the one discussed in Section 3.1. However, we include an additional factor in the DGP. This additional factor is the sixth principal component extracted when calibrating the parameters, so that the denoised factors are six in total: the five Fama and French (2015) and the sixth PC.⁹ This weaker factor is observed with precision (i.e., the measurement error w_t is set equal to zero to avoid additional distortions) and is orthogonal to the other observable factors. This setup is similar to the setting in Anatolyev and Mikusheva (2021). Nonetheless, in our simulations, we impose the Sharpe ratio (SR) of this factor to be relatively high. We impose the Sharpe Ratio of such factor to be a multiple of the market's Sharpe ($SR \in SR_{MKT} \times \{1, 2, 8\}$). In other words, by design, the factor is weak but important to price the cross-section of asset returns. Since we want to impose weaker loadings, we scale down the betas by the (inverse) rate of convergence.¹⁰ Since $\check{\beta}\check{\beta}'/\check{N}^{1-2\theta_1} = I_{\check{N}}$, we have the

⁹We emphasize that Giglio and Xiu (2021) set $\hat{p} = 7$ in their empirical application, therefore assuming that the factor structure is strong and pervasive up to the seventh PC.

¹⁰If one considers $\theta = 0$, then it is the case of the pervasive strong factors, meaning the rate of convergence is fast enough. Vice versa, when $\theta = 1/2$, then we have the case envisaged by Onatski (2012), $\check{\beta}\check{\beta}' \rightarrow I_{\check{N}}$, i.e., $\check{\beta} = O_p(\check{N}^{-\theta})$.

following relationship between the rate and the signal-to-noise ratio:

$$\text{tr}(\Sigma_r - \Sigma_{\hat{u}})/\text{tr}(\Sigma_{\hat{u}}) = \text{tr}(\check{\beta}\Sigma_{\hat{v}}\check{\beta}')/\text{tr}(\Sigma_{\hat{u}}) = \check{N}^{1-2\theta} \times \text{signal-to-noise},$$

where the signal-to-noise = $\text{tr}(\Sigma_{\hat{v}})/\text{tr}(\Sigma_{\hat{u}})$. Thus, changing the different rates of convergence boils down to different specifications of the signal-to-noise ratio. In particular, the relationship between rates of convergence and the signal-to-noise ratio is as follows:

$$\theta = 1/2 \left(1 - \frac{\ln[\text{tr}(\Sigma_r - \Sigma_{\hat{u}})] - \ln(\text{tr}(\Sigma_{\hat{u}})) - \ln(\text{signal-to-noise})}{\ln(\check{N})} \right)$$

We consider $\theta \in 0.05 \times \{1 : 10\}$, i.e., from moderately weak to extremely weak loadings. Given the calibrations we discussed in the previous section, we compute the $\check{N} \times T$ matrix of residuals $\hat{u}_t^{(\check{N})}$ (and its $(\check{N} \times \check{N})$ variance-covariance matrix). Then, factor loadings $\check{\beta}$ and residuals $\hat{u}_t^{(\check{N})}$ are scaled as follows:

$$\begin{aligned} \check{\beta}^* &= \check{\beta} \times \check{N}^{-\theta} \\ \hat{u}_t^* &= \hat{u}_t^{(\check{N})} \times \check{N}^{-\theta_{u_t}}. \end{aligned}$$

where $\check{\beta}^*$ (and \hat{u}_t^*) are the rescaled exposures (and error terms), and $\theta_{u_t} \in 0.05 \times \{1 : 10\}$. This creates a 10×10 grid of possible combinations of rates of convergence, giving a clearer and more comprehensive picture of the behaviour of estimators under the weak loading scenarios. Note that we are implicitly assuming that the initial betas are strong loadings, which, given that the factors to generate returns are calibrated to match the (denoised) Fama and French (2015) factors, is a plausible assumption.

Subsequently, we generate the simulated returns, \check{r}_t^* , according to the asset pricing model in Eq. (1) with the rescaled exposures, denoised six factors, and rescaled error terms, while maintaining the (unscaled) risk premia $\check{\gamma}$:

$$\check{r}_t^* = \check{\beta}^* \check{\gamma} + \check{\beta}^* \hat{v}_t + \hat{u}_t^*. \quad (11)$$

The simulations are then based on Eq. (12) and $\{\check{r}_t^*\}$.

3. SPARSE LOADINGS

We follow a similar procedure as before, as we include in the DGP for returns the same

six factors. However, for this particular case, we impose a level of sparsity on the betas by randomly setting to zero the loadings of a number of assets with respect to some group of factors. We consider the exposures associated with the market factor.¹¹ In practice, we multiply entry-wise the $\check{N} \times 1$ loading $\check{\beta}_{Mkt}$ by a random indicator function vector, $\iota(\theta_{sparse})$, whose entries are generated according to a Bernoulli distribution with the probability of success being $1 - \theta_{sparse}$, with $\theta_{sparse} \in 0.05 \times \{1 : 6\}$. In the worst case ($\theta_{sparse} = 0.3$), therefore, up to 30% of the simulated assets have a risk exposure with respect to the market factor that is equal to zero. Then, we have

$$\begin{aligned}\check{\beta}_i^* &= \check{\beta}_{i,MKT} \times \iota(\theta_{sparse})_i \quad i = 1, \dots, \check{N} \\ \hat{u}_t^* &= \hat{u}_t^{(\check{N})} \times \check{N}^{-\theta_{u_t}},\end{aligned}$$

with θ_{u_t} ranging as before. We generate the new returns, \check{r}_t^* , accordingly to Eq. (11). The simulations are then based on Eq. (12) with respect to $\{\check{r}_t^*\}$.

4. LOCAL AND STRONG/SEMI-STRONG LOADINGS

Similar to the spurious factor case, we do not include the sixth factor in the DGP. This last case relaxes the pervasiveness and the strength of the factors by considering a heterogeneous environment. In Section 4.5 we motivate why this scenario is, in our opinion, relevant. After ranking the returns with respect to their volatilities, we impose a strong factor structure for the lowest-volatility asset returns and a semi-strong factor structure for the remaining ones. Considering the exposures associated with the *SMB* factor, we multiply entry-wise the $\check{N} \times 1$ loadings $\check{\beta}_{SMB}$ by the vector $\zeta(q, \theta)$, whose entries are either 1 for the corresponding bottom $q\%$ of the volatility-ranked returns (i.e., *SMB* is a strong factor for such assets) or $\check{N}^{-\theta}$ (i.e., *SMB* is a semi-strong factor for such assets). We consider $\theta \in 0.05 \times \{1 : 10\}$.¹² We then have

$$\begin{aligned}\check{\beta}_i^* &= \begin{cases} \check{\beta}_{i,SMB} \times \check{N}^{-\theta} & i \notin \text{bottom-50\% volatile} \\ \check{\beta}_{i,SMB} & i \in \text{bottom-50\% volatile} \end{cases} \\ \hat{u}_t^* &= \hat{u}_t^{(\check{N})} \times \check{N}^{-\theta_{u_t}},\end{aligned}$$

¹¹We have results that extend the analysis to other groups of factors. The results are in line with those presented in the paper and are available upon request.

¹²The results with different percentiles are in line with those presented in the paper and thus omitted and available upon request.

with θ_{u_t} ranging as before. We then generate the new returns, \check{r}_t^* , according to Eq. (11). The simulations are then performed as before, via Eq. (12), with respect to $\{\check{r}_t^*\}$.

3.3 Mispricing

Up to this point, our analysis has been carried out under the scenario in which the asset pricing restriction (Eq. (2)) holds, or in other words, by imposing $\alpha = 0$ in the linear factor asset pricing model in Eq. (1).

It is well-known in the asset pricing literature that models are very likely to be misspecified, which in turn translates into non-trivial mispricing (see, for instance, the influential work of Hansen and Jagannathan (1997)).

We then consider the case of non-zero mispricing in three scenarios: STRONG AND PERVASIVE (Section 3.1), WEAK/SEMI-STRONG LOADINGS, and SPARSE LOADINGS (Section 3.2). The simulation setting is the same as before, except that the mean of the multivariate normal for generating the sample (Eq. (12)) is translated by a $(\check{N} + 2 \times K) \times 1$ vector $\check{\alpha} = (\hat{\alpha}', 0_{2 \times K})'$ with variance $\sigma_{\check{\alpha}}^2 = \check{N}^{-1} \check{\alpha}' \check{\alpha}$, with respect to $K = \{5, 6\}$, the number of factors (depending on the scenario), and $\hat{\alpha}$ is the estimated pricing errors, i.e., $\hat{\alpha} = \hat{\mu} - \check{\beta} \check{\gamma}$. As these results are similar to, and often worse than, those of the benchmark scenarios, we relegate these simulation results to Appendix B.

3.4 DGP calibration (Giglio and Xiu (2021))

We use the same set of portfolios employed by Giglio and Xiu (2021). Our calibrations for the simulations are based on a monthly dataset of 647 portfolio returns that span a period of 35 years, from 1976 to 2010. This data set includes portfolios sorted by a different number of characteristics commonly linked to variation in expected equity returns, US Treasury and corporate bonds portfolios, as well as currencies portfolios.

We refer the readers to the supplemental material provided by Giglio and Xiu (2021) for a detailed description of the data sources, test asset portfolios, and factors.

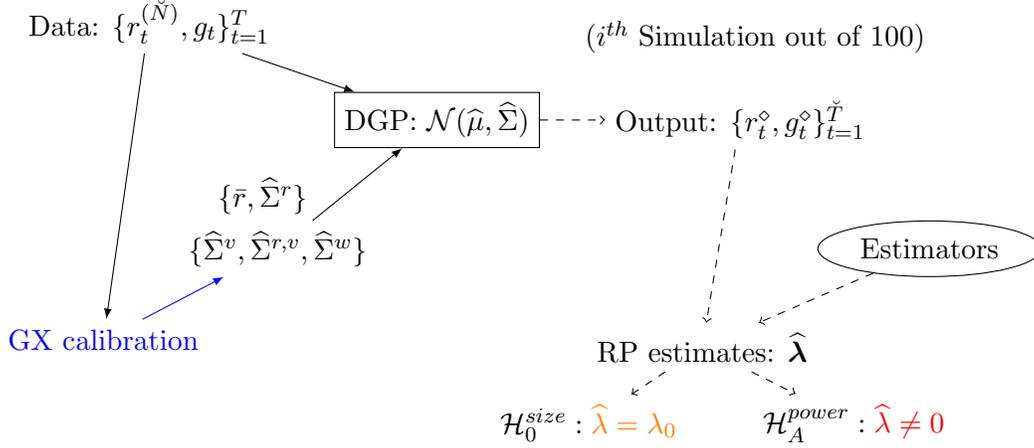


Figure 1: A visual representation of the Monte Carlo simulations for size and power.

4 Results

In this section, we discuss the main findings of our Monte Carlo simulations, while providing an extensive version of our simulation work in the Online Appendix. For ensuring good Monte Carlo properties, all the results are based on 1000 Monte Carlo replications. In our simulations, once we draw a sample of asset returns and observable factors (i.e., noisy proxies of the true factors), $\{r_t^\diamond, g_t^\diamond\}_{t=1}^T$, we can employ one of the estimation strategy presented in Section 2 to get an estimate of the risk premia. Having such estimate, say $\hat{\lambda}$, when studying size and power, we are testing two different hypotheses:

$$\begin{aligned} \mathcal{H}_0^{size} : \hat{\lambda} &= \check{\lambda} \\ \mathcal{H}_A^{power} : \hat{\lambda} &\neq 0 \end{aligned}$$

which are, for size, testing whether the risk premia estimates are statistically close to their true values (that is known in our simulation to be $\check{\lambda}$), and for power, testing whether the risk premia estimates are statistically different with respect to zero.

In practice, when the sample is sufficiently large, this translates into comparing the t -statistics to the critical values of a standard normal distribution with respect to a 5% significance level. Drawing and repeating the simulation experiments will give us an empirical distribution of rejection rates: if the asymptotic normality of the estimators holds correctly, we expect the empirical rejection rates under the null to be around 5%, while having power as large as possible.

In the next subsection, we report the results for different simulation scenarios presented in Section 3. As already mentioned in Section 3.1, the focus of our results is to the case of large- N (and large- T), as the estimators assessed are specifically designed to estimate the risk premium on factors with a large panel of returns. We relegate the remaining results to Online Appendix.¹³

Three remarks are in order. First, if the test is not reliable in terms of size, then its power becomes an immaterial feature to assess. Thus, when the test is severely oversized, we do not comment on results related to power.

Second, if it is not explicitly clear when discussing the PCA-based estimators, we implicitly estimate the risk premia when the number¹⁴ of latent factors is set equal to 7 (i.e., $\hat{p} = 7$), aligning our estimation strategies to the one reported in the work of Giglio and Xiu (2021).

Third, since Anatolyev and Mikusheva (2021)’s estimator requires to split the sample into four non-overlapping subsets, we consider two possible choices: sequential or nonrandom splitting (e.g., the first subsample corresponds to the one where $t = (1 : \check{T}/4)$), and random splitting (e.g., the first subsample is extracted via random draws without replacement).

4.1 Strong and pervasive factors

Table 1 to Table 4 report the rejection rates of the asymptotic t -statistics for the various estimators described in Section 2, when the factor structure of the simulated returns is strong and pervasive.

We start by considering returns generated from the Fama and French (2015)’s 5-factor model. As discussed in Section 3, our set of observable factors includes a noisy version of these five factors. Empirical results for the estimation methodologies of Giglio and Xiu (2021) and Lettau and Pelger (2020) include different specifications of the number of latent factors included in the estimation of the risk premium ($\hat{p} = \{1 : 12\}$).

The first five columns of each table present the results of the test on the estimated risk premium being equal to the true value, thus displaying the empirical size of the test. Columns 6 to 10 report the results when testing the estimated risk premium being equal to 0 (which in population is not).

¹³All other simulations are available upon request.

¹⁴We have a comprehensive study of the number of latent factors in our DGPs, with respect to different Information Criteria estimators. For the sake of exposition, we do not report this here, and it is available upon request.

We refer to these latter as displaying the empirical power of the test.

Table 1 and Table 2 about here

Clearly, the empirical size for the t -tests of the PCA-based methodologies is in line with the 5% nominal level of the test, especially when \check{N} and \check{T} increase. We find little to no difference between the GX and GXLP in terms of size and power, even when the GXLP's tuning parameter μ increases. This finding is consistent across different specifications of the number of assets and time series observations. When the risk premia on factors are estimated using supervised PCA (SPCA), again the size of the t -statistics is in line with the one of the other two PCA-based estimators. However, this methodology exhibits larger power for the *SMB*, *HML*, and *CMA* factors, especially in finite samples (showing an overperformance of around 5 to 10 percentage points).

When the risk premia are estimated using the sample splitting IV methodology proposed by Anatolyev and Mikusheva (2021), our simulation results suggest that only the size of the t statistic associated with the market factor is in line with the nominal level of the test, with power increasing as the time series sample size increases. The t -test for the *SMB* factor is slightly oversized (ranging from 8% to 13% for different sample sizes), while the power, although lower, exhibits similar behaviour as in the case of the market factor (increasing with \check{T}).

Table 3 about here

For the remaining factors that are assessed, we find that the test performs poorly under the null, with rejection rates above 35% for the statistics related to *HML* and *RMW*, and well above 10% for *CMA*. The results are qualitatively the same for different specifications of the number of potentially missing factors and the tuning parameter A . Overall, we can conclude that the asymptotic properties associated with this estimator tend to be not ideal, leading to misleading inference on priced factors when used to price such factor structure.

Turning to the performance of the test built on Gagliardini et al. (2016), we note a similar behaviour as for the case of AM. In fact, the test presents good properties in terms of size when testing the risk premium of the market factor (although the test exhibits lower power), while it

breaks down for the other factors, in which the test is clearly oversized with size close to the power for CMA and RMW. As the asymptotics for this estimator are grounded on the assumption of the factors being strong, our results might suggest that potentially the last two factors in the Fama and French (2015) five-factor model might be considered as (relatively) weak.

Table 4 about here

4.2 Spurious factors

We also examine the statistical properties of the test when a spurious factor, that is, a factor that is uncorrelated with test asset returns, is included in the set of observed factors. As the distortions due to spuriousness when conducting inference might depend on the factor’s variance, we evaluate the asymptotic behaviour of the t-statistic for different specifications of the spurious factor’s standard deviation. (See Section 3.2) We report the tables in the Online Appendix.

Not surprisingly, when we assess the statistical properties of the PCA-based estimators, we find that all tests (GX, GXLP, and SPCA) perform discretely well, with a rejection rate that, on average, does not exceed 5%, if not undersized.

The three-step procedure of Giglio and Xiu (2021) allows us to estimate risk premium for any observable factor separately, even when not all the true risk factors are included in the model. Therefore, including the spurious factor in the analysis does not affect the test results, as for the other five factors we see similar behaviour as in Section 4.1. This conclusion comes from realizing that, if one of the elements in g_t is uncorrelated with the asset returns r_t , then it is necessarily uncorrelated with the latent factors v_t as well. This, in turn, means that in the second step of GX (Eq. (4)), the projection of the spurious factor onto the space spanned by the v_t (i.e., the η ’s), will be (extremely close to) zero, making the estimated risk premium on the observable factor practically zero.

When we turn to the statistical properties of the test for the AM estimator, we provide evidence that this estimator would suffer identification problems: one would tend to conclude more often that a factor is priced when in fact is spurious. Rejection rates for the null of the risk premium being zero for the spurious factor are above 15% in all the simulations we perform. The over-rejection of the null (i.e., oversize) of a factor being spurious is more severe when \check{N} increases (higher than 26%

for $\check{N} = 600$, while around 18% for $\check{N} = 400$). Including additional factors in the factor structure of the error terms can help a little in mitigating the problem: although still above the nominal level of the test, when we include two potential missing factors in the error terms when we estimate risk premia, the rejection rates of the test for the spurious factor reduce from 26% (and 18%) to around 22% (and 12%) for $\check{N} = 600$ (and $\check{N} = 400$).

Furthermore, we note that including a spurious factor in the estimation of the risk premia does not affect the test size (and power) of the other priced factors, as the results of the simulations are in line with those discussed previously.

This pattern is not found when we look at the results for the GOS estimator. In fact, in addition to the extreme oversize of the test for the spurious factor (rejecting the null of the risk premium being zero more than 90% of the times), the inclusion of such spurious factor badly affects the test size for the priced factors, driving very large distortions in conducting overall inference. Thus, we are prone to rank the GOS estimator last in terms of performance.

In general, for all the estimators covered in this work, we find that the standard deviation of the spurious factor does not induce stronger distortions to the size and power of the t-tests, as the test properties do not change relevantly when we increase the volatility of the spurious factor.

4.3 Weak/semi-strong loadings

Next, we examine the size and power of the t-tests for the case of weak factors. Recall that our generated returns have a weak factor structure, i.e., the exposure of the assets to the factors is scaled according to a rate θ . We consider the case in which the process that generates returns consists of a 6-factor model, in which the Fama and French (2015) five-factor model is augmented with the sixth principal component extracted from our test asset returns. Recall that the last factor is already relatively weaker by construction since the denoising is done with respect to the first 5 PCs. In the different simulations, we impose different Sharpe ratios for our sixth factor, in a way to accommodate the framework of Lettau and Pelger (2020), where one of the priced factors is weak but has a relatively high Sharpe Ratio: the Sharpe ratio is set equal to 0.96 (8 times the one of the market factor).¹⁵

¹⁵The results for different Sharpe ratio specifications are available upon request.

We apply a convergence rate to the empirical error term to provide results in terms of a plethora of signal-to-noise ratios. The top left corner of the (x, y) -grid of the figures depicts the case in which the factors are extremely weak ($\theta \rightarrow 0.5$) and the error terms are not scaled ($\theta_{u_t} \rightarrow 0$). Vice versa, towards the bottom centre of the (x, y) -grid, we have the opposite scenario ($\theta \rightarrow 0, \theta_{u_t} \rightarrow 0.5$), meaning that we tend towards the previous scenario of a strong and pervasive factor structure. Regarding the z -axis, warmer colours signal values that are closer to 1. Figure 2 below provides a graphical example of how the charts are set up.

Table 2 about here

Figure 3 and Figure 4 plot respectively the size and power (the z -axis) of the t-statistics for the various estimators as a function of θ on the y -axis and θ_{u_t} on the x -axis (see Section 3.2).

Figure 3 and Figure 4 about here

Clearly, in the extreme scenario of very low signal-to-noise ratio, the size properties of the test for the GX estimator are far from being optimal. The rejection rates under the null of correct risk premium are over 90% for all the 6 factors. Given a θ of 0.45 (i.e., a close-to-be weak factor structure), the empirical rejection rates are close to the nominal level of the test for $\theta_{u_t} > 0.35$ for the FF3 factors, while we reach the same level of size for all factors only when $\theta_{u_t} > 0.45$. As we expected, in first approximation, we can conclude that, if the latent factor structure is not strong enough, inference on the estimated risk premia can be severely misleading.

A similar conclusion is reached when we turn our analysis to the GXLP estimator. Despite the size properties of t-statistic showing improvements for lower level of signal-to-noise, rejection rates are still very far from the nominal level of the test (on average, around 70% with peaks of 95% and 100% when testing the market factor and the sixth principal component). Again, as we saw for the GX estimator, when we move from the top left to the bottom right corner of the figure, hence increasing the strength of the factor structure of our generated returns, the test becomes fairly reliable. No sign of improvement is shown when test asset selection is performed and the SPCA estimator is adopted. This comes with little to no surprise as the returns in this scenario share the same factor structure. Hence, ideally, the subset of assets that is selected in the first step of the procedure would have the same factor structure as the starting dataset.

In general, when the factor structure is extremely weak ($\theta \rightarrow 0.5$), the tests for all the PCA-based estimators considered in this work show poor asymptotic properties, with the size being fairly close to the power of the test, a condition that suggests that none of these estimators can provide reliable inference when the factors that drive the cross-section of returns are extremely weak.

When we look at the performance of AM, we do not see a common pattern across factors when the signal-to-noise ratio is very low. In fact, when doing inference on the estimated risk premium on the market factor, the size of the test is 5.3% while for the other Fama and French (2015)'s factors the empirical sizes are 11.4%, 36.4%, 31.6%, and 17.3%, for *SMB*, *HML*, *RMW* and *CMA*, respectively. When testing the last factor (the sixth principal component), the rejection rates for the null of correct risk premium stand at 1.9%, clearly showing that the test is undersized in this situation (in general, for this set of simulations, the size of the test on the sixth factor is below 2% when a $\theta \approx \theta_{u_t}$). As we move to the bottom right part of the graph, the results converge to the ones discussed in the previous section.

The AM t -statistic seems to behave fairly well when compared to the others. In the extreme case, the power when testing the estimated risk premium on market factor is around 60%, on average, when the value of θ_{u_t} is larger than 0.05. The test on the market factor shows low power (around 10%), when $\theta_{u_t} < 0.15$ and $\theta = 0.5$. When testing the other factors, the test does not seem to perform well in terms of power. As an example, for the observed *HML* and *CMA*, the rejection probabilities of the null of the risk premium being zero when the factor is priced do not exceed 20%. In general, when we are in the top left corner of the charts, the test exhibits no power across all the priced factors.

Lastly, regarding the t -statistic of GOS, its behaviour substantially deteriorates, and it is hard to draw a conclusion that differs from the previous one. The test is strongly oversized for almost all the signal-to-noise ratios. We remind the reader that GOS has been designed especially for the estimation of conditional asset pricing models, and thus, in our context, represents a benchmark regarding estimation strategy.

4.4 Sparse loadings

We now focus on a setting in which the exposures are sparse. In particular, we decide to set to zero (some of) the assets' exposure to the *MKT* factor so that, for those selected assets, their

returns are not driven by changes in the MKT factor (see Section 3.2).

Since the true factor loadings of some assets are zero for one of the factors in the DGP, when the model is estimated, the source of variation in the betas for such assets is driven by measurement errors, thus making the estimated risk premium biased. Under this scenario, the proposed methodology of Giglio et al. (2022) becomes appealing. In fact, among the assets, there exists by construction a set of test asset returns in which the market factor is priced. Therefore, if the screening procedure for the test assets of Giglio et al. (2022) can consistently screen out those assets with zero exposure to the market, we should be able to estimate the risk premium of the factor with higher precision than the one estimated with the other methodologies.

Fig. 5 shows the size of the t -test of the null of the estimated factor risk premium being equal to the true value for the estimators that do not perform a selection on the test assets. The rejection frequencies under the null are plotted for different specifications of the error rate and sparsity rate. The sparsity rate ranges from 5% to 30%, indicating that the proportion of assets in our universe with zero exposure to the market varies from 5% to 30% respectively.

Figure 5 and Figure 6 about here

First of all, regarding the MKT 's t -statistic, the PCA-based methodologies of GX and GXLP produce a t -test that is severely oversized, and the over-rejections do not depend neither on the number of assets with zero exposure to the market factor (i.e., the θ_{sparse}), nor the error term scale (i.e., θ_{ut}), as illustrated by the flat surface. In terms of magnitude, for MKT , the simulated rejection frequencies are around 60% when the number of latent factors included in the estimation of the risk premium is 7. Moreover, there seems to be little to no difference when factors are extracted by the weighted PCA methodology of Lettau and Pelger (2020).

Overall, the analysis regarding the MKT 's t -stat does not change when we employ the AM and GOS estimators. The rejection rates are still over 50% for the market factor even when this factor is spurious for a small portion of the total assets. For the other factors, the over-rejection under the null of correct risk premia is less severe, although still relevant for all four estimators, with the GX and GXLP ones performing relatively better. For these two estimators, in fact, we see that the rejection rates for the t -test on the SMB and HML factors are respectively around 20% and 10%, while when the risk premia are estimated using the GOS and AM methodologies, the

frequencies of the over-rejection are well above 50% for each of the two factors. The same pattern is found when we look at the test results for the *RMW* and *CMA* factors. Under the null of the risk premium being equal to the true value, the test is oversized by 15% and 2% respectively for *RMW* and *CMA* even when the rate on errors or the sparsity rate varies. On the other hand, when we consider the AM and GOS estimators, the test seems unreliable as the empirical rejection frequencies are far from the nominal level of the test.

Even when we employ the methodology that embeds test asset selection, our conclusions remain unchanged. The third row of panel a in Fig. 5 shows the empirical size of the test for the SPCA estimator. We do not see any particular difference from the results of the other PCA-based estimators. This might support the hypothesis that the first step in the SPCA estimator fails to perform consistent test asset selection, and thus screen out all those assets for which the market factor is spurious (i.e., those assets associated with zero exposure on the market factor). This in turn produces distortions when conducting inference for the risk premia estimate of *MKT* that are similar to the other PCA-based estimators.

What we can infer from these results is that if the loadings are sparse, the betas that are recovered through PCA are not, in general, sparse. Thus, when we use the estimated loadings on the unobservable factors to estimate the risk premium, the error-in-variable in our estimated factor loadings renders inference problematic not only for the sparse factor but also for the other priced risk factors.

4.5 Local and strong/semi-strong loadings

Finally, in order to analyze the properties of proposed estimators when factors are not pervasive, we again simulate a 5-factor model from the denoised Fama and French (2015)'s factors, similar to our benchmark scenario. However, we focus now on understanding the behaviour of *t*-test when factors' strength varies across groups of asset returns. (See Section 3.2)

To put this scenario into context, suppose the dataset is heterogeneous in asset classes, in the sense that comprehends returns from portfolios of stocks and bonds. Now, we can imagine that some of the factors that strongly drive bond returns (for example, credit spread) might only weakly drive the cross-section of returns for the equity portfolio. However, in practice, when we extract principal components from the variance-covariance matrix of our test asset returns, our estimated

latent factors might not capture the feature of the low-volatility assets as the principal components are estimated to maximize the variance of the dataset.

To study this scenario, we then split the test assets into two groups: strong factors (or loadings) for low-volatility assets and weak(er) loadings for high-volatile assets. In doing so, we do not take a stance on whether the specification of the volatility dynamics characterizes a particular asset class, as in our example for credit spread, bonds and stocks. Figures 7 and 8 display the size and power properties of the asymptotic t -test of the various estimators when the *SMB* factor is strong for the bottom half of the assets when ranked in descending order in terms of volatility and weaker for the remaining assets.¹⁶

Figure 7 and Figure 8 about here

As for size, we note that the test performs poorly when θ_{u_t} is relatively low and thus the factor structure of the empirical errors is not extremely weak. From the plots, it seems that what drives the performance of the test is not the rate applied to the factor loading, θ , but rather the rate on the errors, θ_{u_t} . In fact, if we fix the θ_{u_t} , the magnitude of the size of the test is roughly constant across different specifications of θ .

For the GX estimator, interestingly, we see a U-shape behaviour for the size of the test across different factors when the θ_{u_t} moves from 0 to 0.5. Only when the error rate is moderate (i.e., $\theta_{u_t} \approx 0.2$), the test has a good size (the test is slightly undersized, with empirical rejection rates around 4%). Surprisingly, as flattening on the errors increases ($\theta_{u_t} \rightarrow 0.5$), the rejection rates under the null of correct risk premia reach at most 30%.

A different pattern arises when we consider the weighted estimator GXLP. In this case, the plots show a similar pattern as for the GX estimator when the rate on the errors is low, but as the rate increases, making the empirical signal-to-noise ratio higher, the empirical size converges to around 6% in magnitude. Despite this compelling convergence toward the nominal level of the test, we note that when the θ_{u_t} is fixed at 0.35, the test is substantially undersized at values lower than 2% across different factors. Since the results are qualitatively similar when using a larger number of PCs ($\hat{p} = 9, 12$) in the estimation of the risk premia, we do not report them. Regarding SPCA,

¹⁶The remaining results for different specifications of the percentile rankings are available upon request.

the patterns are similar to GX and often much worse for the t -test associated with some factors, such as the MKT and RMW factors.

Analogous conclusions are reached when looking at the charts corresponding to AM t -test. In particular, we find that, for values of θ_{u_t} larger than 0.35, the size of the test across simulations and factors is virtually 0, which raises some suspicions about the reliability of this estimator in this extreme scenario. The behaviour of the t -test associated with the GOS estimates is behaving extremely poorly.

The analysis holds when we look at the power property of the test. In general, across estimators and factors, it seems that the size of the test is of the same order of magnitude as the power (this is true in particular for the GOS estimator). In conclusion, inference is not generally reliable because the properties of the test become fairly unstable when the signal-to-noise ratio is low.

5 Dissecting Giglio and Xiu (2021)

Our analysis so far focuses on the behaviour of the asymptotic t -tests for various estimators used in empirical work in asset pricing. This section points our attention specifically to the methodology proposed by Giglio and Xiu (2021).

Since the estimation strategy estimates the risk premia on observable factors, $\hat{\gamma}_g$, as a byproduct of two components (the latent factor risk premia and the projection coefficient of observable factors onto latent factors), the goal is to have a clearer picture on whether identification issues could arise from the matrix of beta exposures to latent factors. If this matrix of risk exposure tends to be column-rank deficient, or nearly column-rank deficient, identification issues arise and jeopardize statistical inference. As principal components are usually ordered according to the amount of variance they represent, higher-ranked PCs might only be weakly correlated with test assets returns (see the sWF model in Uematsu and Yamagata (2022a), Uematsu and Yamagata (2022b)). In other words, this could lead to weakly identified risk premia of the estimated latent factor model, $\hat{\gamma}$ presented in Eq. (1). Consider, for example, the full sample of Giglio and Xiu (2021) asset returns, and estimates the $\hat{\beta}$ of model Eq. (1) including in the estimation 12 estimated principal components. The left chart of Fig.9 displays the frequencies on how many times the $\{|\hat{\beta}_i|\}_{i=1,\dots,N}$ are of order of 10^{-4} or below. The frequency is remarkably low for the loading associated with the

1st principal component, but it is nearly 30% for the 7th PC and even reaches almost 50% for the 12th PC.

From a more formal perspective, we apply the sequential test proposed by Onatski (2009) on their panel of assets. Consider an estimate of the number of latent factors \hat{p} , and (a potential candidate for) the true value p , the null hypothesis of the test states that the estimated number of factors is equal to the true one $\hat{p} = p$. The alternative hypothesis is that the true number of factors is between \hat{p} and $\hat{p} + K$, with K being a positive integer that is chosen by practitioners. The right chart of Fig. 9 indicates that the number of latent factors in Giglio and Xiu (2021)’s dataset is 4 (p -values > 0.1). We remind the readers that most of Giglio and Xiu (2021)’s empirical results are obtained considering $\hat{p} = 7$ in the estimation of the risk premia.

Figure 9 about here

What happens to the asymptotics of $\hat{\gamma}_g$ when the exposures are so weak? In particular, when too many latent factors are estimated, what happens to the Fama and MacBeth (1973)’s estimates of the latent model? To a certain extent, this is discussed in Giglio et al. (2022), but their focus is rather on the number of latent factors.

To study this issue, and the distortions of including too many latent factors in the estimation of the risk premia, we design a Monte Carlo experiment following the procedure outlined in Section 3. We consider returns simulated from a one-factor model: the only factor that drives return is the *MKT* factor. In the same vein as in Section 3.2, we apply different convergence rates to the betas and to the residuals to weaken the original factor structure, so that we can analyze diverse signal-to-noise ratios. We generate a panel of monthly returns from 400 different simulated (portfolios of) assets for 1200 periods. We choose such a large \check{T} framework because spuriousness is more severe when the number of time series observations tends to be large. (Kleibergen (2009)) We then apply the three-step methodology to compute the market risk premium, using, in the time series regression, 7 estimated principal components (i.e., $\hat{p} = 7$) as in Giglio and Xiu (2021). Recall that the factor structure of the artificial returns is generated by a single factor. This suggests that the 2nd to 7th PCs should exhibit progressively weaker correlations (possibly nearing spurious correlations) with test asset returns. Consequently, the risk premia estimates might be affected by distortions similar to those that arise when a spurious factor is included in the estimation of asset

pricing models.

In figures 10 and 11, each chart depicts, respectively, the empirical distribution of the risk premium estimate on the first principal component, $\hat{\gamma}$, and estimated market risk premium, $\hat{\gamma}_g$, as functions of θ on the y -axis and of the convergence rates of the residuals θ_{u_t} on the x -axis.

Figure 10 and figure 11 about here

We provide the empirical distribution of all the γ 's and η 's (Eq. 1-4) corresponding to the different PCs in Appendix ??.

From a first look at the plots, we can clearly see that, confirming our findings on the weak/semi-strong loadings, inference starts to break down for some θ . For instance, when considering $\theta_{u_t} = 0.05$ (top-left) and $\theta > 0.2$, we see that the empirical distribution ceases to be bimodal. The bimodal behaviour in the distribution of the estimates is not necessarily informative regarding the spuriousness of the latent factors as it can be the result of a rotation matrix. However, jointly looking at the unimodality and/or bimodality of the γ 's and η 's might be rather informative in some cases, e.g., $\{\theta_{u_t} = 0.05, \theta > 0.15\}$ (red coloured) for the second PCs (η_2 , γ_2 , and $\gamma_2 \times \eta_2$) in Appendix ??. In this latter example, we see that the empirical distribution of the risk premia estimates of the *MKT* factor onto the second PC (i.e, $\gamma_2 \times \eta_2$) has fatter tails. This could potentially corroborate our suspicions in finite sample. As Giglio and Xiu (2021)'s methodology makes use of the two-step procedure in the latent factor model, it is not immune to the problems of the original methodology, which might translate into having the estimator's properties quite sensitive to the number of (over-)extracted principal components.

6 Conclusions

This paper investigates the properties of the t -tests associated with risk premium estimates in linear beta-pricing models when the number of assets is large. Our goal is to provide practitioners with the first comprehensive comparison of the statistical properties of recently proposed large- N risk premium estimators.

In our Monte Carlo settings, we maintain the cross-sectional structure of the artificial data as close as possible to the one of the real data and we consider various scenarios that can lead to

identification problems. Specifically, we analyze situations in which the test asset returns exhibit i) a strong and pervasive factor structure, ii) a weak and pervasive factor structure, iii) specifications with spurious factors and with local loadings. In our simulations, we also account for measurement error issues (proxy variables) and mispricing.

In terms of statistical properties of the t -tests, we find that no estimator consistently outperforms the others. Thus, a practitioner would need to decide which estimator to use depending on her priors on the factor that is deemed to be priced and the underlying factor structure. When the factors are strong and pervasive, the PCA-based estimators are outperforming. Vice versa, when considering a weak and pervasive factor, Anatolyev and Mikusheva (2021)'s t -test shows good statistical properties. None of the candidate estimators shows satisfactory performance when the loadings are sparse, or are local and semi-strong. This result is quite relevant and indicates that the composition of the panel of asset returns matters. For example, if a practitioner were to work with a panel of heterogeneous asset classes (e.g., bonds and stocks), the resulting approximate factor structure would be probably close to the one described in these latter scenarios with sparse or local factors. Regarding these cases, newly proposed methods may potentially mitigate the documented size and power distortions, such as Bryzgalova et al. (2023), heteroPCA of Zhang et al. (2022) and sparsePCA of Cai et al. (2013) (although the properties of the last ones have not been studied yet in the asset-pricing context). We leave these investigations for future research.

Table 1: Simulation results – Giglio and Xiu (2021) methodology This table presents the performance of the t -statistic in terms of *size* and *power* for different numbers of latent factors in the estimation (1 - 11). Data are simulated according to the 5-factor model of Fama and French (2015). Factors are de-noised using the first 7 principal components. Panel A to D report results for different specifications of N and T .

Panel A: $N = 400, T = 400$										
#PC	Size					Power				
	<i>Mkt</i>	<i>SMB</i>	<i>HML</i>	<i>RMW</i>	<i>CMA</i>	<i>Mkt</i>	<i>SMB</i>	<i>HML</i>	<i>RMW</i>	<i>CMA</i>
1	0.042	0.159	1.000	1.000	1.000	0.727	0.702	0.690	0.681	0.708
3	0.040	0.131	0.063	0.638	0.189	0.627	0.096	0.358	0.105	0.147
5	0.037	0.044	0.052	0.083	0.046	0.648	0.273	0.301	0.603	0.314
7	0.036	0.050	0.044	0.045	0.042	0.645	0.267	0.260	0.627	0.260
9	0.036	0.048	0.044	0.044	0.043	0.646	0.271	0.258	0.619	0.267
11	0.036	0.047	0.047	0.039	0.046	0.647	0.273	0.255	0.618	0.262

Panel B: $N = 400, T = 600$										
#PC	Size					Power				
	<i>Mkt</i>	<i>SMB</i>	<i>HML</i>	<i>RMW</i>	<i>CMA</i>	<i>Mkt</i>	<i>SMB</i>	<i>HML</i>	<i>RMW</i>	<i>CMA</i>
1	0.072	0.180	1.000	1.000	1.000	0.880	0.866	0.866	0.864	0.874
3	0.064	0.141	0.084	0.797	0.224	0.801	0.136	0.476	0.129	0.190
5	0.064	0.044	0.050	0.108	0.052	0.825	0.401	0.406	0.773	0.450
7	0.065	0.041	0.049	0.059	0.067	0.824	0.412	0.330	0.796	0.338
9	0.065	0.040	0.050	0.056	0.066	0.824	0.411	0.329	0.790	0.338
11	0.065	0.041	0.053	0.059	0.063	0.825	0.408	0.330	0.788	0.343

Panel C: $N = 600, T = 600$										
#PC	Size					Power				
	<i>Mkt</i>	<i>SMB</i>	<i>HML</i>	<i>RMW</i>	<i>CMA</i>	<i>Mkt</i>	<i>SMB</i>	<i>HML</i>	<i>RMW</i>	<i>CMA</i>
1	0.068	0.321	1.000	1.000	1.000	0.859	0.840	0.840	0.827	0.846
3	0.052	0.168	0.061	0.909	0.293	0.787	0.099	0.393	0.077	0.123
5	0.051	0.053	0.057	0.196	0.047	0.790	0.417	0.390	0.655	0.448
7	0.051	0.055	0.050	0.060	0.057	0.791	0.393	0.322	0.778	0.426
9	0.052	0.054	0.054	0.054	0.058	0.789	0.395	0.316	0.780	0.417
11	0.052	0.052	0.052	0.051	0.055	0.789	0.394	0.317	0.780	0.407

Panel D: $N = 600, T = 1200$										
#PC	Size					Power				
	<i>Mkt</i>	<i>SMB</i>	<i>HML</i>	<i>RMW</i>	<i>CMA</i>	<i>Mkt</i>	<i>SMB</i>	<i>HML</i>	<i>RMW</i>	<i>CMA</i>
1	0.072	0.521	1.000	1.000	1.000	0.987	0.986	0.987	0.986	0.987
3	0.065	0.344	0.079	0.996	0.473	0.967	0.101	0.690	0.124	0.200
5	0.066	0.048	0.069	0.298	0.059	0.967	0.664	0.699	0.938	0.749
7	0.064	0.050	0.060	0.059	0.051	0.966	0.622	0.582	0.984	0.744
9	0.066	0.051	0.056	0.049	0.050	0.966	0.623	0.580	0.984	0.735
11	0.066	0.050	0.059	0.049	0.050	0.966	0.619	0.580	0.981	0.741

Table 2: Simulation results – Lettau and Pelger (2020) weighted methodology This table presents the performance of the t -statistic in terms of *size* and *power* for different numbers of latent factors in the estimation (1-11). Data are simulated according to the 5-factor model of Fama and French (2015). Factors are de-noised using the first 7 principal components. Panel A to D report results from different specifications of N and T .

Panel A: $N = 400, T = 400$										
#PC	Size					Power				
	<i>Mkt</i>	<i>SMB</i>	<i>HML</i>	<i>RMW</i>	<i>CMA</i>	<i>Mkt</i>	<i>SMB</i>	<i>HML</i>	<i>RMW</i>	<i>CMA</i>
1	0.065	0.253	1.000	1.000	1.000	0.781	0.765	0.752	0.735	0.768
3	0.074	0.170	0.764	0.183	0.601	0.499	0.518	0.958	0.811	0.942
5	0.042	0.105	0.110	0.165	0.124	0.639	0.489	0.418	0.860	0.562
7	0.045	0.086	0.076	0.148	0.091	0.659	0.429	0.376	0.737	0.444
9	0.036	0.066	0.046	0.076	0.049	0.639	0.361	0.317	0.509	0.284
11	0.038	0.062	0.052	0.058	0.045	0.641	0.331	0.284	0.461	0.254

Panel B: $N = 400, T = 600$										
#PC	Size					Power				
	<i>Mkt</i>	<i>SMB</i>	<i>HML</i>	<i>RMW</i>	<i>CMA</i>	<i>Mkt</i>	<i>SMB</i>	<i>HML</i>	<i>RMW</i>	<i>CMA</i>
1	0.093	0.294	1.000	1.000	1.000	0.915	0.909	0.904	0.904	0.911
3	0.090	0.166	0.881	0.208	0.738	0.693	0.642	0.990	0.906	0.979
5	0.064	0.104	0.102	0.191	0.134	0.820	0.643	0.513	0.958	0.692
7	0.062	0.087	0.086	0.168	0.109	0.827	0.594	0.481	0.895	0.603
9	0.063	0.059	0.060	0.080	0.064	0.823	0.508	0.435	0.696	0.407
11	0.062	0.054	0.059	0.061	0.064	0.825	0.471	0.405	0.618	0.342

Panel C: $N = 600, T = 600$										
#PC	Size					Power				
	<i>Mkt</i>	<i>SMB</i>	<i>HML</i>	<i>RMW</i>	<i>CMA</i>	<i>Mkt</i>	<i>SMB</i>	<i>HML</i>	<i>RMW</i>	<i>CMA</i>
1	0.072	0.597	1.000	1.000	1.000	0.889	0.878	0.882	0.868	0.888
3	0.075	0.195	0.955	0.143	0.825	0.657	0.635	0.999	0.781	0.993
5	0.056	0.220	0.093	0.257	0.190	0.783	0.791	0.524	0.926	0.767
7	0.056	0.141	0.079	0.093	0.097	0.771	0.665	0.454	0.671	0.535
9	0.051	0.115	0.069	0.067	0.074	0.766	0.645	0.448	0.535	0.453
11	0.052	0.072	0.065	0.060	0.046	0.790	0.493	0.390	0.530	0.368

Panel D: $N = 600, T = 1200$										
#PC	Size					Power				
	<i>Mkt</i>	<i>SMB</i>	<i>HML</i>	<i>RMW</i>	<i>CMA</i>	<i>Mkt</i>	<i>SMB</i>	<i>HML</i>	<i>RMW</i>	<i>CMA</i>
1	0.095	0.829	1.000	1.000	1.000	0.990	0.990	0.990	0.990	0.990
3	0.108	0.174	0.997	0.141	0.970	0.904	0.810	1.000	0.959	1.000
5	0.066	0.284	0.112	0.456	0.307	0.963	0.962	0.813	1.000	0.967
7	0.069	0.145	0.085	0.110	0.127	0.961	0.891	0.742	0.897	0.819
9	0.066	0.121	0.081	0.057	0.084	0.960	0.891	0.743	0.822	0.775
11	0.064	0.059	0.059	0.067	0.067	0.966	0.735	0.677	0.823	0.655

Table 3: Simulation results – Anatolyev and Mikusheva (2021) and Gagliardini et al. (2016) methodologies. This table presents the performance of the t -statistic in terms of *size* and *power*. Data are simulated according to the 5-factor model of Fama and French (2015). Factors are de-noised using the first 5 principal components. Panel A reports the simulation results for the Anatolyev and Mikusheva (2021) meanwhile panel D reports the simulation results for the unconditional version of the Gagliardini et al. (2016) estimator.

Panel A: Anatolyev and Mikusheva (2021), circular splitting

N	T	Size					Power				
		<i>Mkt</i>	<i>SMB</i>	<i>HML</i>	<i>RMW</i>	<i>CMA</i>	<i>Mkt</i>	<i>SMB</i>	<i>HML</i>	<i>RMW</i>	<i>CMA</i>
400	400	0.055	0.090	0.387	0.422	0.166	0.604	0.433	0.184	0.663	0.230
	600	0.054	0.109	0.505	0.458	0.193	0.776	0.603	0.230	0.669	0.227
	1200	0.046	0.185	0.646	0.420	0.150	0.973	0.829	0.258	0.614	0.178
600	600	0.052	0.107	0.376	0.448	0.271	0.760	0.560	0.188	0.627	0.302
	1200	0.063	0.122	0.532	0.409	0.264	0.955	0.819	0.161	0.541	0.301

Panel B: Gagliardini et al. (2016)

N	T	Size					Power				
		<i>Mkt</i>	<i>SMB</i>	<i>HML</i>	<i>RMW</i>	<i>CMA</i>	<i>Mkt</i>	<i>SMB</i>	<i>HML</i>	<i>RMW</i>	<i>CMA</i>
400	400	0.049	0.193	0.606	0.792	0.802	0.589	0.609	0.375	0.851	0.808
	600	0.049	0.142	0.501	0.788	0.801	0.591	0.523	0.297	0.835	0.813
600	600	0.062	0.205	0.625	0.846	0.882	0.754	0.705	0.368	0.871	0.893
	1200	0.083	0.347	0.743	0.914	0.931	0.952	0.818	0.496	0.915	0.927

Table 4: Simulation results – Giglio et al. (2022) methodology This table presents the performance of the t -statistic in terms of *size* and *power*. Data are simulated according to the 5-factor model of Fama and French (2015). Factors are de-noised using the first 5 principal components. Panel A to E report results for different specifications of N and T . Columns 2 to 5 show respectively the minimum, mean, median, and maximum of the number of assets retained after the supervised asset selection step is performed.

Panel A: $N = 50$														
T	N_0				Size					Power				
	Min	Avg	Med	Max	Mkt	SMB	HML	RMW	CMA	Mkt	SMB	HML	RMW	CMA
100	10	27.42	20	50	0.052	0.051	0.051	0.049	0.048	0.228	0.154	0.115	0.068	0.071
200	10	39.20	50	50	0.043	0.036	0.047	0.039	0.045	0.404	0.301	0.161	0.101	0.074
400	10	42.29	50	50	0.052	0.059	0.056	0.050	0.046	0.632	0.493	0.327	0.136	0.154
600	20	44.05	50	50	0.049	0.055	0.045	0.052	0.047	0.779	0.706	0.461	0.212	0.202
1200	20	46.46	50	50	0.047	0.054	0.052	0.061	0.046	0.981	0.932	0.728	0.355	0.304

Panel B: $N = 100$														
T	N_0				Size					Power				
	Min	Avg	Med	Max	Mkt	SMB	HML	RMW	CMA	Mkt	SMB	HML	RMW	CMA
100	20	67.38	60	100	0.074	0.053	0.059	0.046	0.042	0.213	0.056	0.251	0.264	0.111
200	20	83.46	100	100	0.048	0.054	0.044	0.053	0.046	0.311	0.057	0.425	0.521	0.183
400	40	91.80	100	100	0.055	0.043	0.056	0.055	0.053	0.593	0.066	0.690	0.814	0.275
600	60	93.84	100	100	0.060	0.060	0.036	0.049	0.045	0.739	0.084	0.891	0.937	0.437
1200	60	97.10	100	100	0.050	0.057	0.050	0.053	0.041	0.955	0.090	0.994	0.998	0.711

Panel C: $N = 200$														
T	N_0				Size					Power				
	Min	Avg	Med	Max	Mkt	SMB	HML	RMW	CMA	Mkt	SMB	HML	RMW	CMA
200	40	174.960	200	200	0.065	0.050	0.047	0.062	0.053	0.426	0.180	0.153	0.413	0.173
400	80	179.720	200	200	0.045	0.040	0.039	0.053	0.044	0.663	0.297	0.245	0.714	0.305
600	80	182.800	200	200	0.057	0.051	0.048	0.046	0.059	0.832	0.442	0.357	0.858	0.451
1200	120	187.800	200	200	0.044	0.048	0.063	0.054	0.066	0.984	0.703	0.639	0.991	0.739

Panel D: $N = 400$														
T	N_0				Size					Power				
	Min	Avg	Med	Max	Mkt	SMB	HML	RMW	CMA	Mkt	SMB	HML	RMW	CMA
400	160	356.24	400	400	0.037	0.050	0.043	0.050	0.036	0.645	0.264	0.309	0.612	0.330
600	160	366.16	400	400	0.060	0.039	0.047	0.055	0.062	0.827	0.397	0.412	0.796	0.449
1200	240	385.36	400	400	0.054	0.052	0.041	0.038	0.043	0.979	0.689	0.720	0.980	0.785

Panel E: $N = 600$														
T	N_0				Size					Power				
	Min	Avg	Med	Max	Mkt	SMB	HML	RMW	CMA	Mkt	SMB	HML	RMW	CMA
600	360	547.80	600	600	0.052	0.054	0.047	0.047	0.050	0.783	0.418	0.401	0.665	0.467
1200	360	574.20	600	600	0.063	0.049	0.050	0.039	0.056	0.966	0.665	0.721	0.941	0.762

Figure 2: 3D charts for size (or power)

The figure plots an example of size (power) for figures presented in section 4.

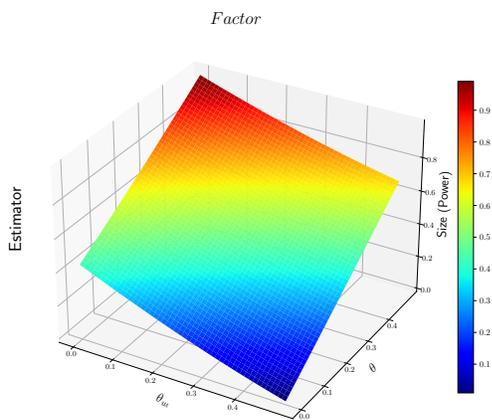
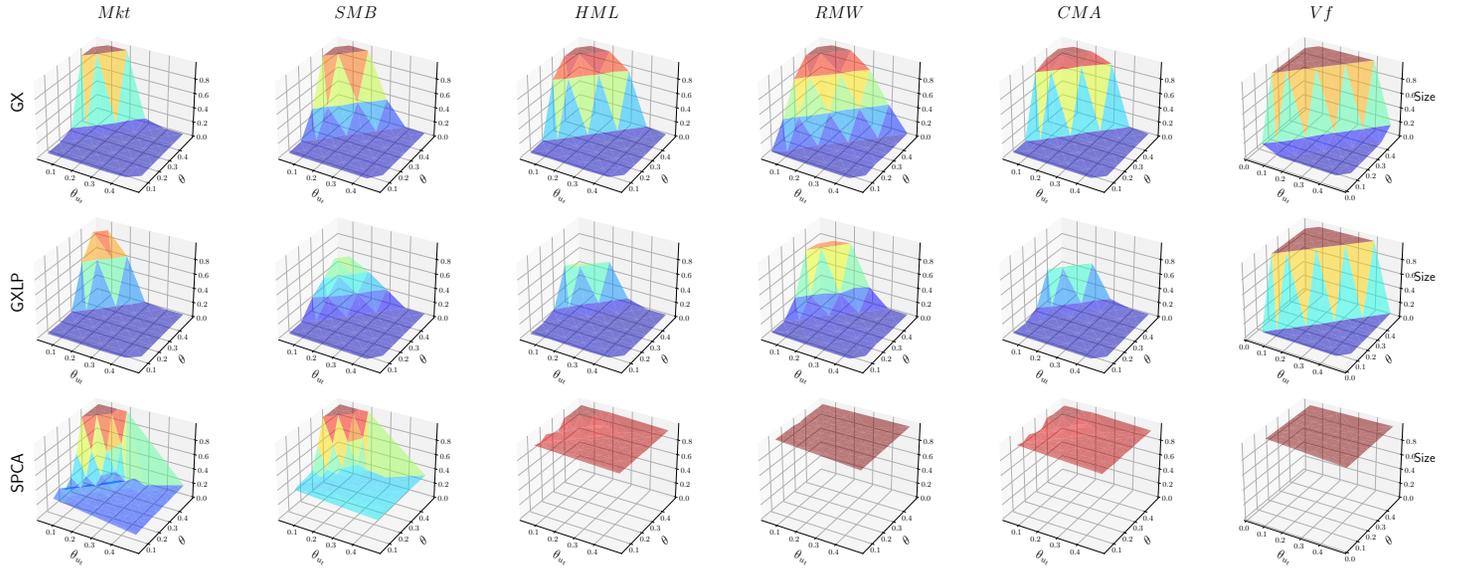
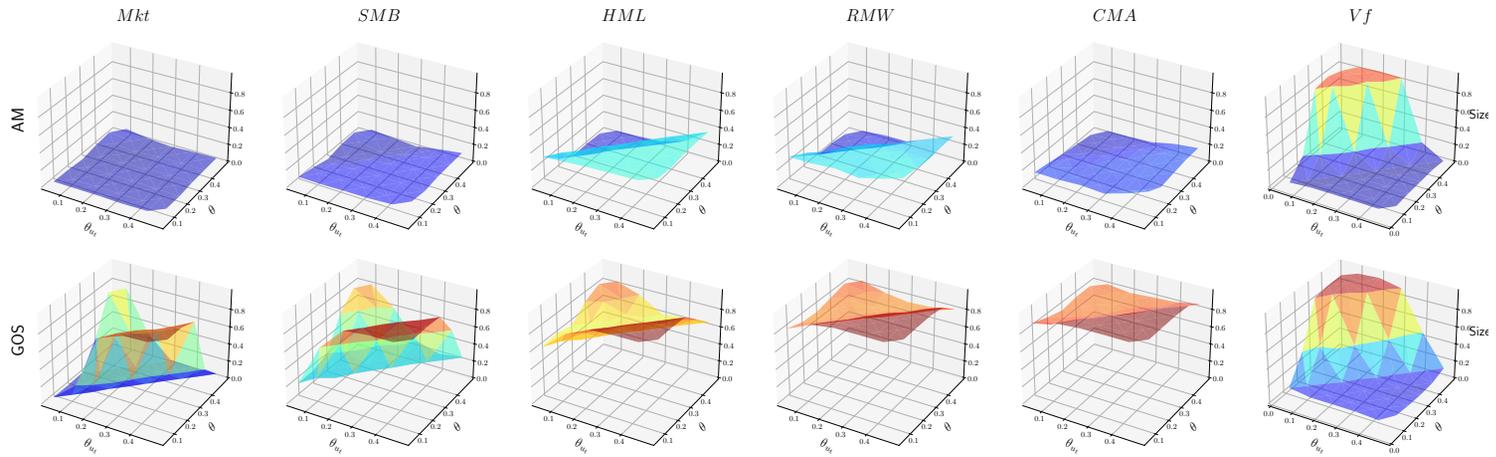


Figure 3: Test size when loadings are weak and pervasive. $N = 400, T = 400$.

The figure plots the size of the t -statistics for PCA-based estimators (Panel A) and estimators related to observable factors (Panel B) when loadings are weak and pervasive.



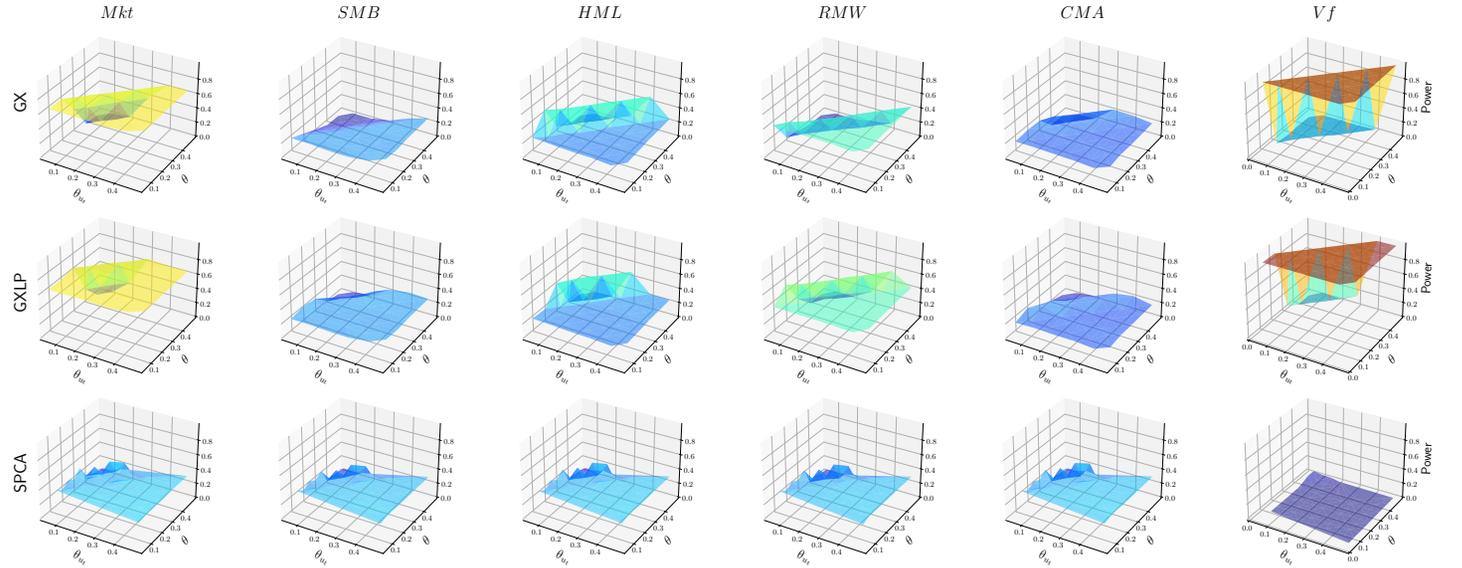
(a) Panel A



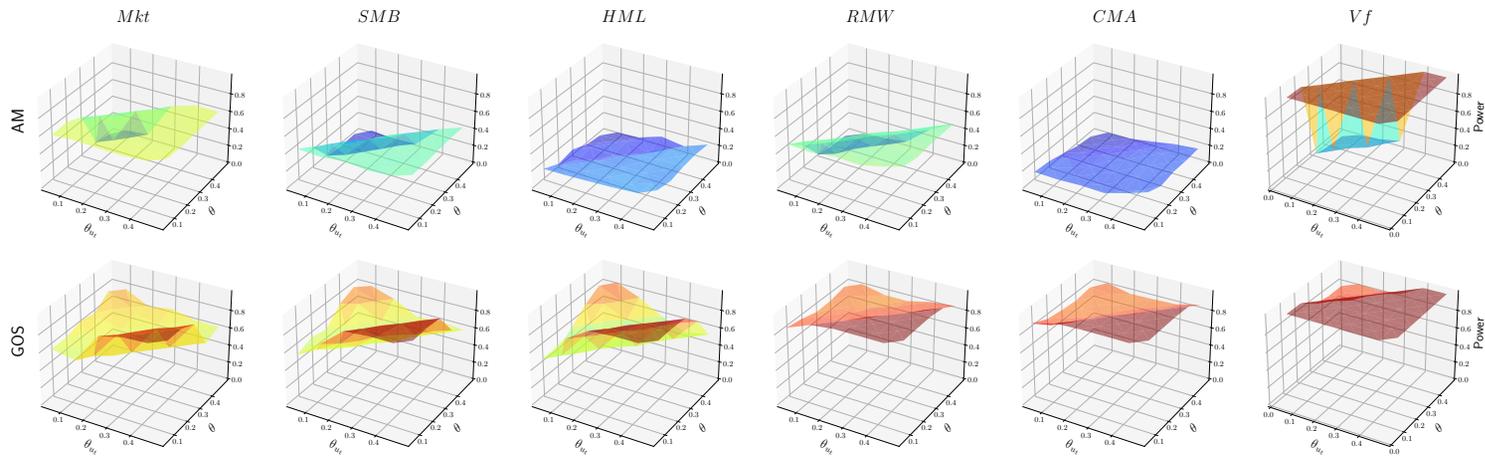
(b) Panel B

Figure 4: Test power when loadings are weak and pervasive. $N = 400, T = 400$.

The figure plots the power of the t -statistics for PCA-based estimators (Panel A) and estimators related to observable factors (Panel B) when loadings are weak and pervasive.



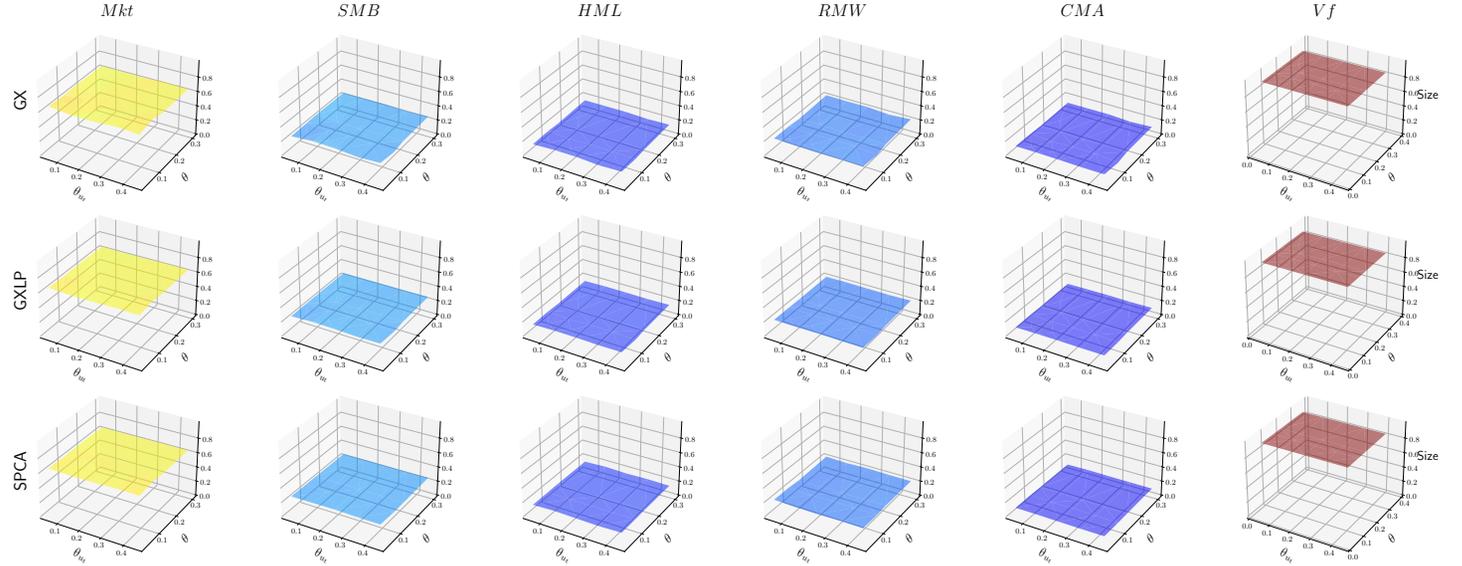
(a) Panel A



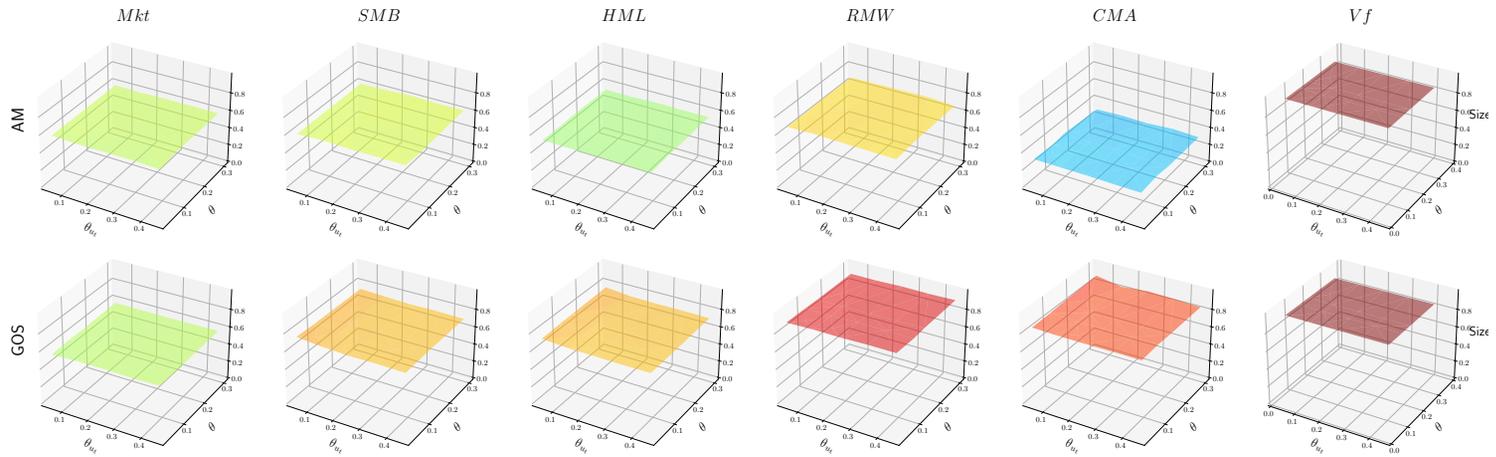
(b) Panel B

Figure 5: Test size when loadings are sparse. $N = 400, T = 400$.

The figure plots the size of the t -statistics for PCA-based estimators (Panel A) and estimators related to observable factors (Panel B) when some assets have zero exposure to the Mkt factor.



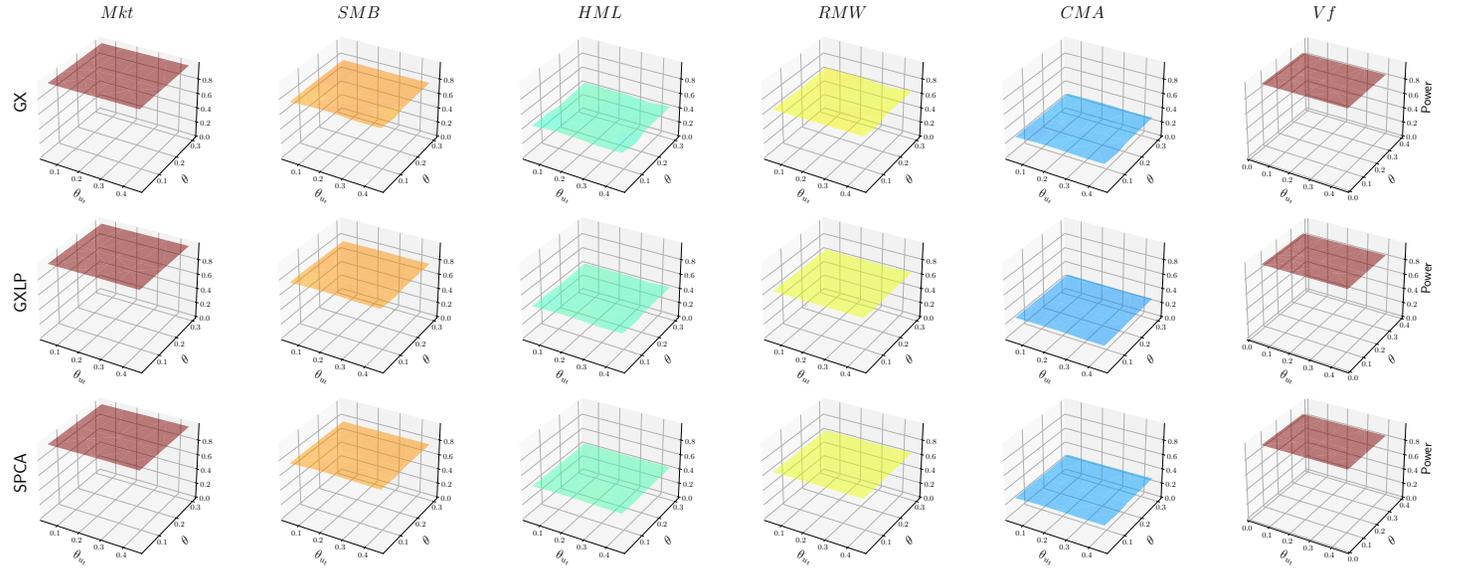
(a) Panel A



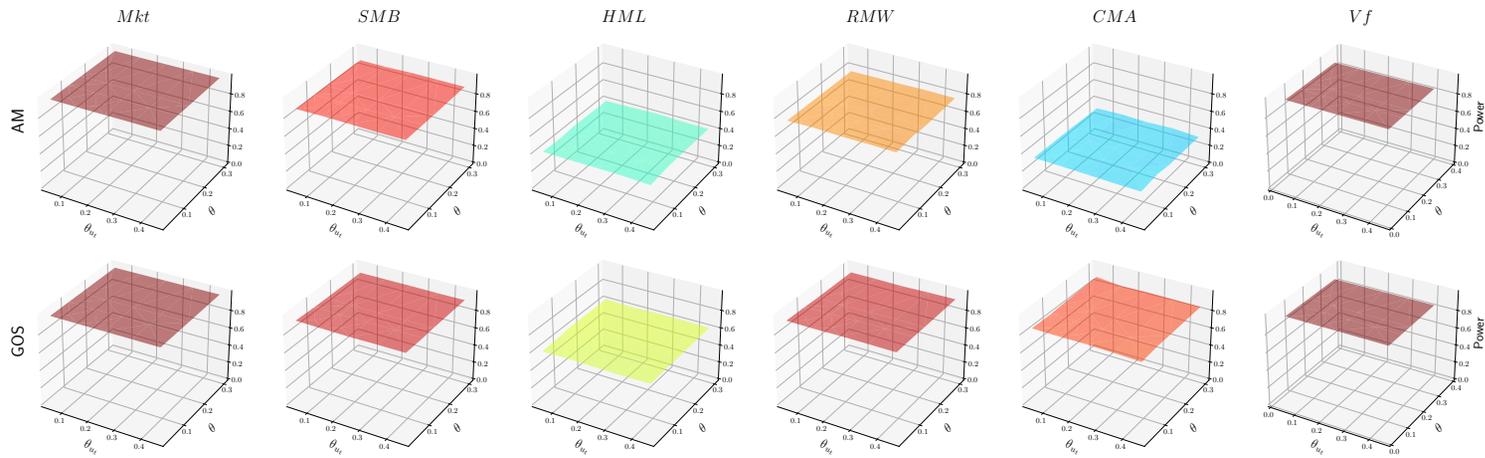
(b) Panel B

Figure 6: Test power when loadings are sparse. $N = 400, T = 400$.

The figure plots the power of the t -statistics for PCA-based estimators (Panel A) and estimators related to observable factors (Panel B) when some assets have zero exposure to the Mkt factor.



(a) Panel A



(b) Panel B

Figure 7: Test size when loadings are local and strong/semi-strong. $N = 400, T = 400$.

The figure plots the size of the t -statistics for PCA-based estimators (Panel A) and estimators related to observable factors (Panel B) when the lowest volatile assets have a strong exposure to the *SMB* factor.

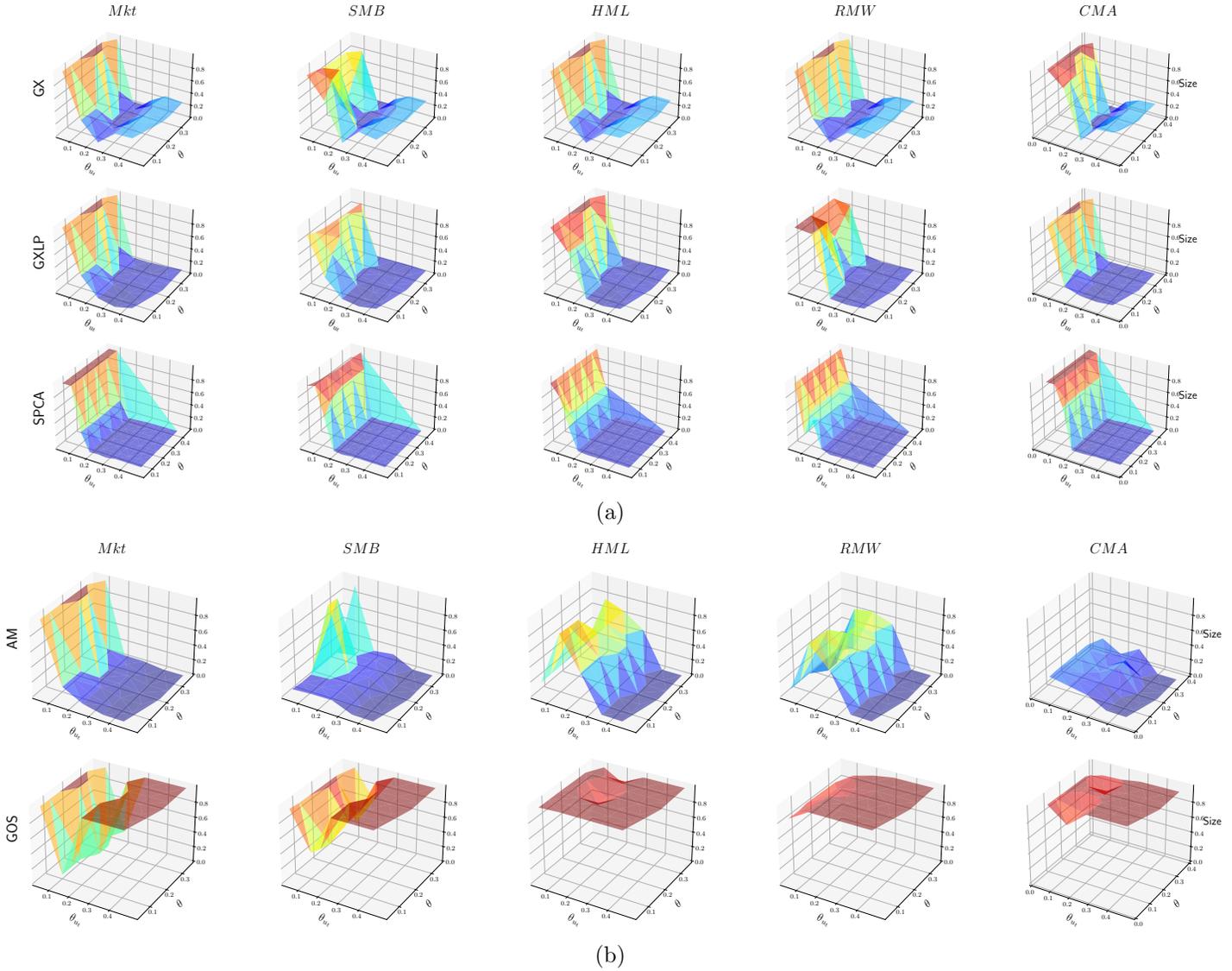
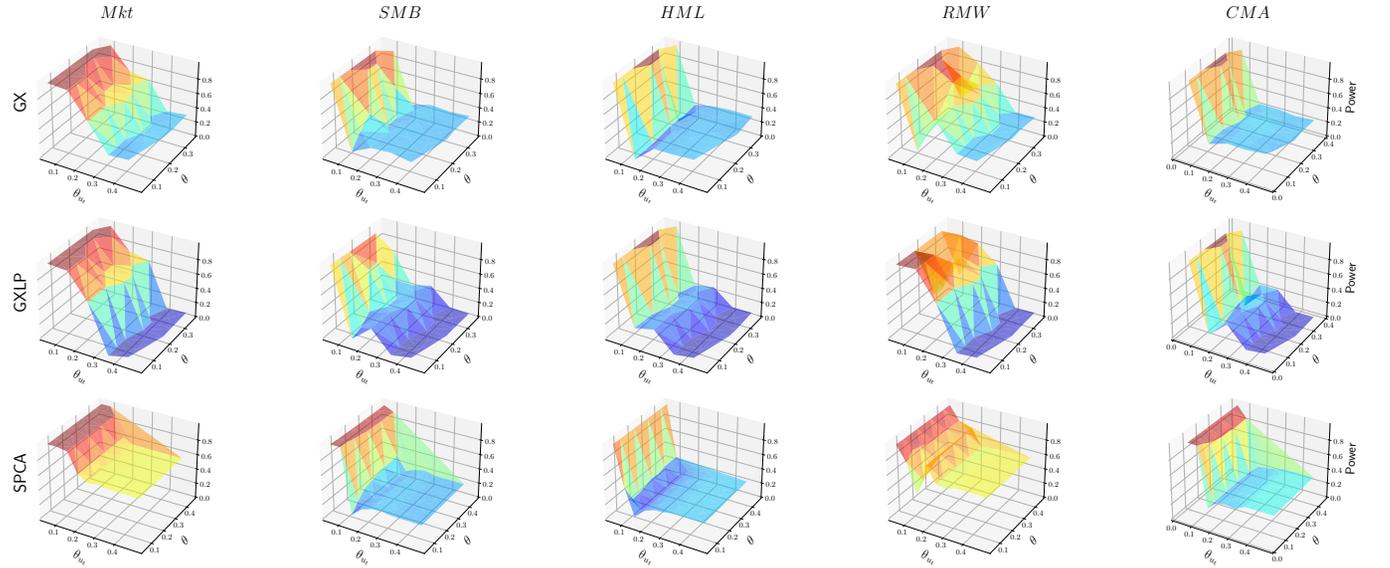
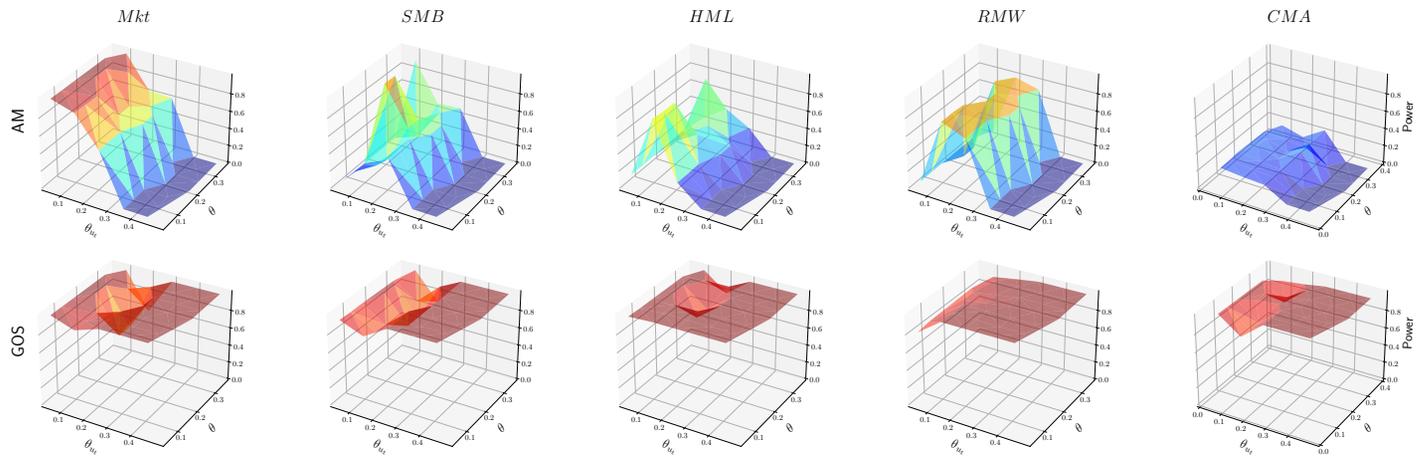


Figure 8: Test power when loadings are local and strong/semi-strong. $N = 400, T = 400$.

The figure plots the power of the t -statistics for PCA-based estimators (Panel A) and estimators related to observable factors (Panel B) when the lowest volatile assets have a strong exposure to the *SMB* factor.



(a)



(b)

Figure 9: Betas and number of latent factors.

The figure plots, on the left, the frequency of $|\beta_i| \lesssim 10^{-4}$. On the right, the sequential testing of Onatski (2009) (on the z-axis, its p-value).

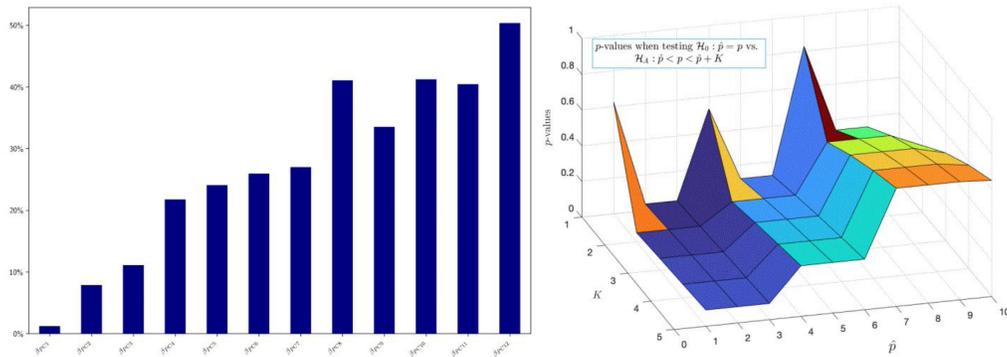


Figure 10: Latent factor risk premium

The figure plots the empirical distribution of the risk premium estimate $\tilde{\gamma}$ on the first principal component as a function of θ and θ_{u_t} .

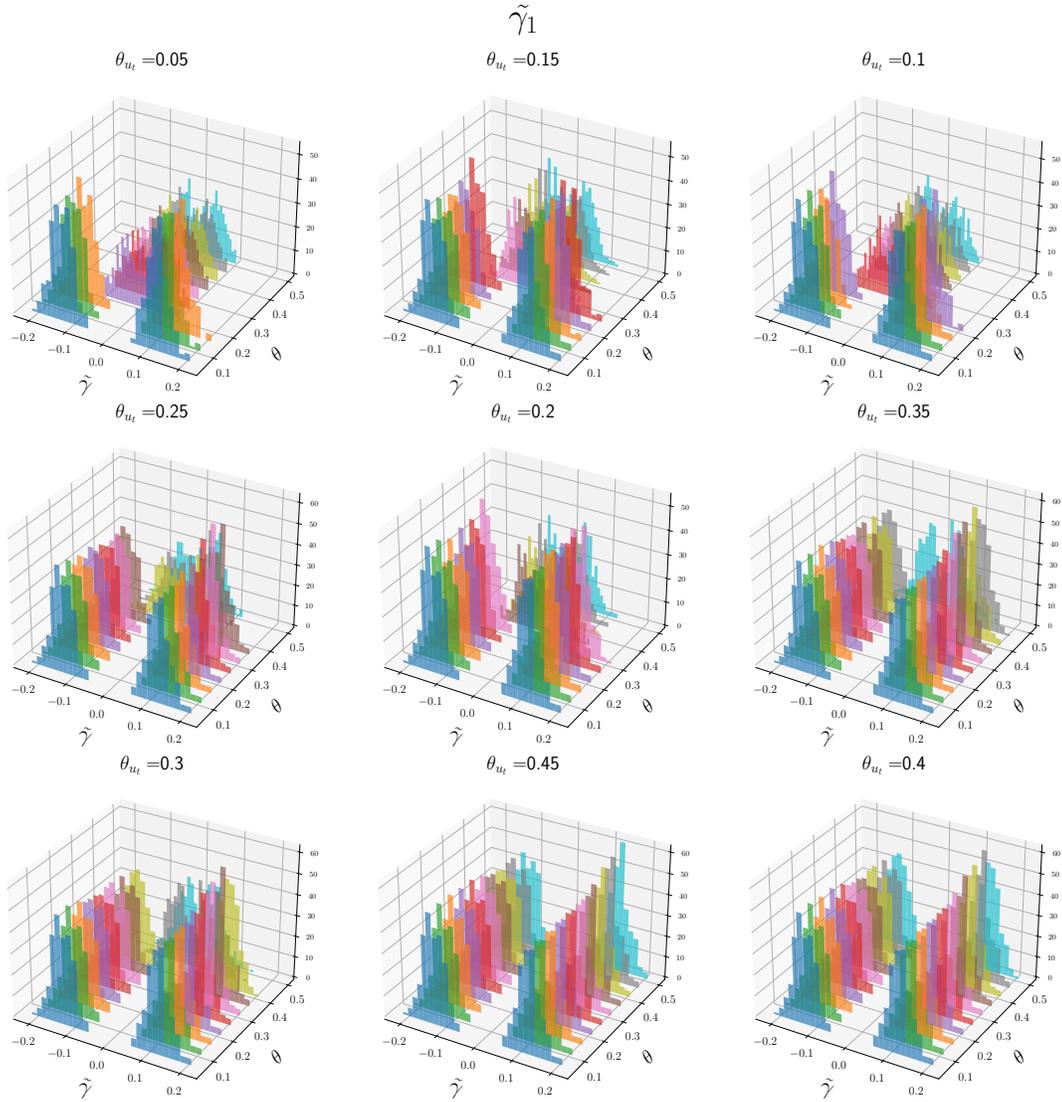
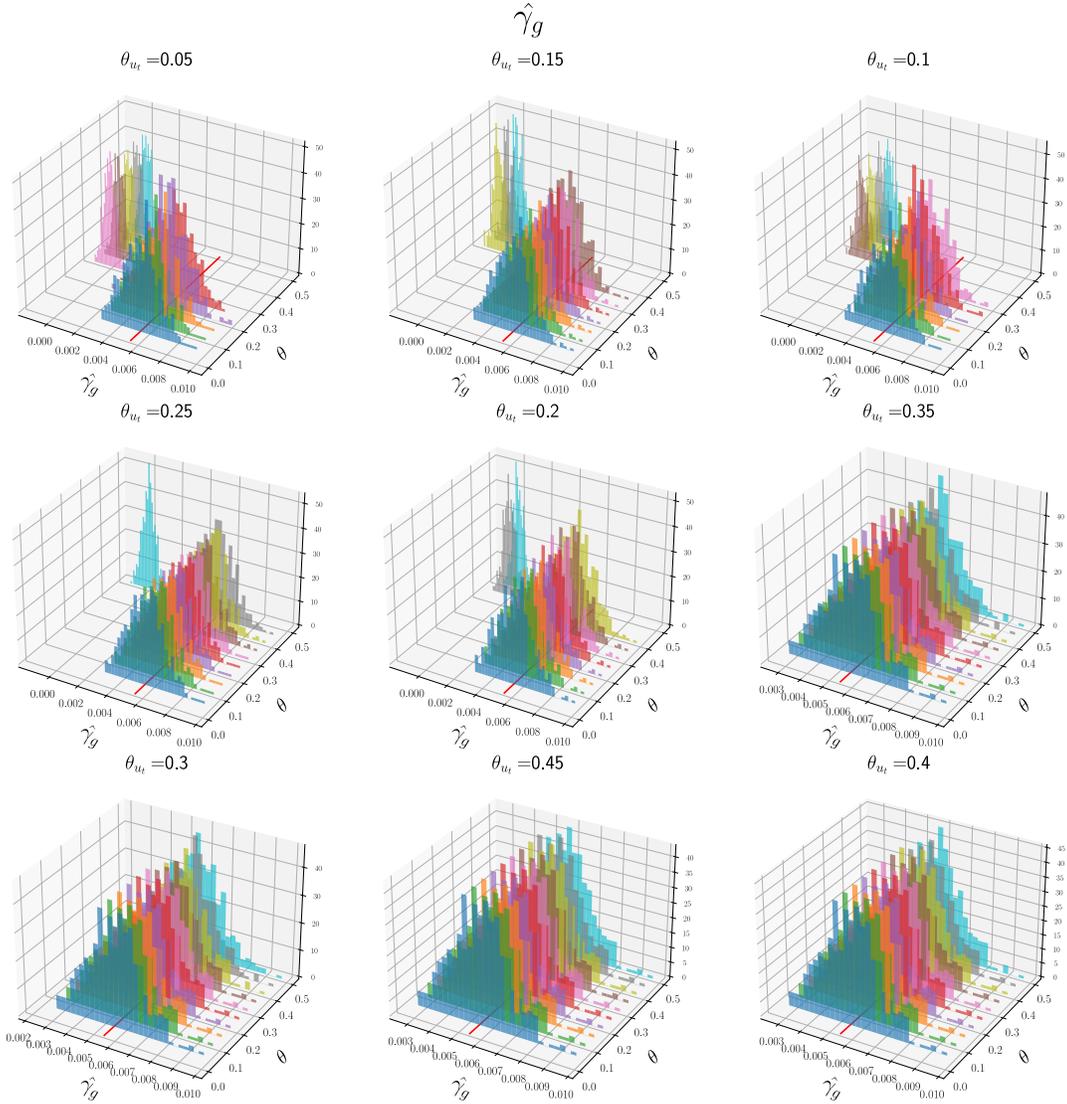


Figure 11: Market risk premium

The figure plots the empirical distribution of the market risk premium $\hat{\gamma}_g$ as a function of θ and θ_{u_t} . The true market risk premium is 0.55%.



References

- Anatolyev, Stanislav, and Anna Mikusheva, 2021, Factor models with many assets: strong factors, weak factors, and the two-pass procedure, *Journal of Econometrics* .
- Antoine, Bertille, Kevin Proulx, and Eric Renault, 2020, Pseudo-true sdfs in conditional asset pricing models, *Journal of Financial Econometrics* 18, 656–714.
- Bai, Jushan, 2003, Inferential theory for factor models of large dimensions, *Econometrica* 71, 135–171.
- Bryzgalova, Svetlana, Victor DeMiguel, Sicong Li, and Markus Pelger, 2023, Asset-pricing factors with economic targets, *Available at SSRN 4344837* .
- Cai, T Tony, Zongming Ma, and Yihong Wu, 2013, Sparse pca: Optimal rates and adaptive estimation, *The Annals of Statistics* 41, 3074–3110.
- Choi, Jungjun, and Xiye Yang, 2022, Asymptotic properties of correlation-based principal component analysis, *Journal of Econometrics* 229, 1–18.
- Connor, Gregory, and Robert A Korajczyk, 2022, Semi-strong factors in asset returns, *Michael J. Brennan Irish Finance Working Paper Series Research Paper* .
- Fama, Eugene F, and Kenneth R French, 2015, A five-factor asset pricing model, *Journal of financial economics* 116, 1–22.
- Fama, Eugene F, and James D MacBeth, 1973, Risk, return, and equilibrium: Empirical tests, *Journal of political economy* 81, 607–636.
- Freyaldenhoven, Simon, 2021, Factor models with local factors—determining the number of relevant factors, *Journal of Econometrics* .
- Gagliardini, Patrick, Elisa Ossola, and Olivier Scaillet, 2016, Time-varying risk premium in large cross-sectional equity data sets, *Econometrica* 84, 985–1046.
- Giglio, Stefano, and Dacheng Xiu, 2021, Asset pricing with omitted factors, *Journal of Political Economy* 129, 1947–1990.
- Giglio, Stefano, Dacheng Xiu, and Dake Zhang, 2022, Test assets and weak factors, Technical report, National Bureau of Economic Research.
- Gospodinov, Nikolay, Raymond Kan, and Cesare Robotti, 2014, Misspecification-robust inference in linear asset-pricing models with irrelevant risk factors, *The Review of Financial Studies* 27, 2139–2170.
- Gospodinov, Nikolay, Raymond Kan, and Cesare Robotti, 2019, Too good to be true? fallacies in evaluating risk factor models, *Journal of Financial Economics* 132, 451–471.
- Hansen, Lars Peter, and Ravi Jagannathan, 1997, Assessing specification errors in stochastic discount factor models, *The Journal of Finance* 52, 557–590.

- Harvey, Campbell R, Yan Liu, and Heqing Zhu, 2016, ... and the cross-section of expected returns, *The Review of Financial Studies* 29, 5–68.
- Huberman, Gur, Shmuel Kandel, and Robert F Stambaugh, 1987, Mimicking portfolios and exact arbitrage pricing, *The Journal of Finance* 42, 1–9.
- Kan, Raymond, Cesare Robotti, and Jay Shanken, 2013, Pricing model performance and the two-pass cross-sectional regression methodology, *The Journal of Finance* 68, 2617–2649.
- Kan, Raymond, and Chu Zhang, 1999, Two-pass tests of asset pricing models with useless factors, *the Journal of Finance* 54, 203–235.
- Kleibergen, Frank, 2009, Tests of risk premia in linear factor models, *Journal of econometrics* 149, 149–173.
- Kleibergen, Frank, and Zhaoguo Zhan, 2015, Unexplained factors and their effects on second pass r-squared's, *Journal of Econometrics* 189, 101–116.
- Lettau, Martin, and Markus Pelger, 2020, Estimating latent asset-pricing factors, *Journal of Econometrics* 218, 1–31.
- Merton, Robert C, 1973, An intertemporal capital asset pricing model, *Econometrica: Journal of the Econometric Society* 867–887.
- Onatski, Alexei, 2009, Testing hypotheses about the number of factors in large factor models, *Econometrica* 77, 1447–1479.
- Onatski, Alexei, 2012, Asymptotics of the principal components estimator of large factor models with weakly influential factors, *Journal of Econometrics* 168, 244–258.
- Pukthuanthong, Kuntara, Richard Roll, and Avanidhar Subrahmanyam, 2019, A protocol for factor identification, *Review of Financial Studies* 32, 1573–1607.
- Ross, Stephen A, 1976, The arbitrage theory of capital asset pricing, *Journal of Economic Theory* 13, 341–360.
- Sharpe, William F, 1964, Capital asset prices: A theory of market equilibrium under conditions of risk, *The journal of finance* 19, 425–442.
- Uematsu, Yoshimasa, and Takashi Yamagata, 2022a, Estimation of sparsity-induced weak factor models, *Journal of Business & Economic Statistics* 41, 213–227.
- Uematsu, Yoshimasa, and Takashi Yamagata, 2022b, Inference in sparsity-induced weak factor models, *Journal of Business & Economic Statistics* 41, 126–139.
- Zhang, Anru R, T Tony Cai, and Yihong Wu, 2022, Heteroskedastic pca: Algorithm, optimality, and applications, *The Annals of Statistics* 50, 53–80.

Appendix A

A Review of estimation strategies

A.1 Giglio and Xiu (2021)

Let us denote the time-demeaned excess returns matrix, $\bar{R} = R - \bar{r}l'_T$, with R as the $N \times T$ matrix of excess returns and $\bar{r} = T^{-1} \sum_{t=1}^T r_t$. Equally, we denote the $d \times T$ matrix of time-demeaned observable factors \bar{G} , which collects the $\{g_t\}$ vectors of observables.

1. PCA STEP. The first step is to extract the PCs from the time-demeaned excess returns matrix, $\bar{R} = R - \bar{r}l'_T$, with $\bar{r} = T^{-1} \sum_{t=1}^T r_t$, from the $T \times T$ matrix $(nT)^{-1} \bar{R}' \bar{R}$. Given a \hat{p} consistent estimator of the number of (latent) factors, we have the following outputs:

$$\hat{V} = T^{1/2}(\xi_1 : \dots : \xi_{\hat{p}}), \quad \hat{\beta} = T^{-1} \bar{R} \hat{V}'$$

where $\xi_{\hat{p}}$ is the normalized (to norm 1) eigenvector corresponding to the \hat{p}^{th} largest eigenvalues of the matrix $(nT)^{-1} \bar{R}' \bar{R}$.

2. CROSS-SECTIONAL REGRESSION STEP. Run a cross-sectional OLS of \bar{R} onto the estimated $\hat{\beta}$. We have the following output:

$$\hat{\gamma} = (\hat{\beta}' \hat{\beta})^{-1} \hat{\beta}' \bar{r}$$

which is the risk premia of the estimated latent factor model.

3. TIME SERIES REGRESSION STEP. Run a time series regression of g_t onto \hat{V} . We have the following output:

$$\hat{\eta} = \bar{G} \hat{V}' (\hat{V} \hat{V}')^{-1}$$

4. RISK PREMIUM OF THE OBSERVED. The risk premium for the observed factor, $\hat{\gamma}_g$ is estimated by:

$$\hat{\gamma}_g = \hat{\eta} \hat{\gamma}$$

A.2 Giglio et al. (2022)

Let us consider the time-demeaned excess returns matrix, $\bar{R} = R - \bar{r}l'_T$, with R as the $N \times T$ matrix of excess returns and $\bar{r} = T^{-1} \sum_{t=1}^T r_t$. Equally, we denote the $d \times T$ matrix of time-demeaned observable factors \bar{G} , which collects the $\{g_t\}$ vectors of observables.

1. ITERATION. The starting values $\{\bar{R}_{(0)}, \bar{r}_{(0)}, \bar{G}_{(0)}\}$ are respectively $\{\bar{R}, \bar{r}, \bar{G}\}$. For an appropriate number of iteration $k = 1, 2, \dots$, using $\{\bar{R}_{(k)}, \bar{r}_{(k)}, \bar{G}_{(k)}\}$:

- (a) **SELECTION.** Select an appropriate choice of subset of test assets $\hat{I}_k \subset [N]$.

The candidate is:

$$\hat{I}_k = \left\{ i | T^{-1} \| (\bar{R}_{(k)})_{[i]} \bar{G}'_{(k)} \|_{MAX} \geq c_q^{(k)} \right\}$$

with $\|\cdot\|_{MAX}$ being the \mathcal{L}_∞ norm of A on the vector space, and $c_q^{(k)}$ the $(1-q)^{th}$ quantile of $(T^{-1} \| (\bar{R}_{(k)})_{[i]} \bar{G}'_{(k)} \|_{MAX})$.

- (b) **GX.** Follow the first three steps of Giglio and Xiu (2021)'s procedure setting $\hat{p} = 1$, with selected return matrix $(\bar{R}_{(k)})_{[\hat{I}_k]}$, and $\bar{G}_{(k)}$. Denote the estimates coming from the estimation: $\{\hat{V}_{(k)}, \hat{\beta}_{(k)}, \hat{\eta}_{(k)}, \hat{\gamma}_{(k)}\}$

- (c) **NEW INPUT.** Obtain $\{\bar{R}_{(k+1)}, \bar{r}_{(k+1)}, \bar{G}_{(k+1)}\}$ from: $\bar{R}_{(k+1)} = \bar{R}_{(k)} - \hat{\beta}_{(k)} \hat{V}_{(k)}$, $\bar{r}_{(k+1)} = \bar{r}_{(k)} - \hat{\beta}_{(k)} \hat{\gamma}_{(k)}$, $\bar{G}_{(k+1)} = \bar{G}_{(k)} - \hat{\eta}_{(k)} \hat{V}_{(k)}$.

- (d) **END.** Stop at $k = \hat{p}$, with \hat{p} based on some proper stopping rule, such as $c_q^{(k)} < c \in \mathbb{R}$.

2. **RISK PREMIA.** The estimate of the risk premia is $\hat{\gamma}^{SPCA} = \sum_{k=1}^{\hat{p}} \hat{\eta}_{(k)} \hat{\gamma}_{(k)}$

A.3 Lettau and Pelger (2020)

Let us denote R as the $N \times T$ matrix of excess returns and $\bar{r} = T^{-1} \sum_{t=1}^T r_t$, and consider the time-demeaned excess returns matrix, $\bar{R} = R - \bar{r} \mathbf{1}'_T$. Equally, we denote the $d \times T$ matrix of time-demeaned observable factors \bar{G} , which collects the $\{g_t\}$ vectors of observables. Let us consider \tilde{Q} a $T \times T$ matrix of given weights.

1. **PCA STEP.** The first step is to extract the PCs from the $T \times T$ matrix Υ :

$$\Upsilon = \tilde{Q}^\top R^\top (I + \frac{\gamma}{T} \mathbf{1} \mathbf{1}^\top) R \tilde{Q}$$

Given a \hat{p} consistent estimator of the number of (latent) factors, we have the following outputs:

$$\hat{V} = T^{1/2} (\xi_1 : \dots : \xi_{\hat{p}}), \quad \hat{\beta} = T^{-1} \bar{R} \hat{V}'$$

where $\xi_{\hat{p}}$ is the normalized (to norm 1) eigenvector corresponding to the \hat{p}^{th} largest eigenvalues of the matrix $(nT)^{-1} \Upsilon$.

2. **CROSS-SECTIONAL REGRESSION STEP.** Run a cross-sectional OLS of \bar{R} onto the estimated $\hat{\beta}$. We have the following output:

$$\hat{\gamma}^{LP} = (\hat{\beta}' \hat{\beta})^{-1} \hat{\beta}' \bar{r}$$

which is the risk premia of the estimated latent factor model.

3. **TIME SERIES REGRESSION STEP.** Run a time series regression of g_t onto \hat{V} . We have the following output:

$$\hat{\eta}^{LP} = \bar{G} \hat{V}' (\hat{V} \hat{V}')^{-1}$$

4. **RISK PREMIUM OF THE OBSERVED.** The risk premium for the observed factor, $\hat{\gamma}_g$ is estimated by:

$$\hat{\gamma}^{LP} = \hat{\eta}^{LP} \hat{\gamma}^{LP}$$

A.4 Anatolyev and Mikusheva (2021)

Let us denote F_t the $K \times 1$ vector of observed factors, r_t the $N \times 1$ vector of excess returns, and β the $N \times K$ vector of exposures. Let us define the demeaned factors as $\tilde{F}_t = F_t - T^{-1} \sum_s F_s$. Let us consider the division the set of time indexes into 4 equal non-intersecting subsets T_j , $j = 1, \dots, 4$.

1. For each asset i and each subset j run a time-series regression to estimate the coefficients of risk exposure:

$$\hat{\beta}_i^{(j)} = \left(\sum_{t \in T_j} \tilde{F}_t^{(j)} \tilde{F}_t^{(j)'} \right) \sum_{t \in T_j} \tilde{F}_t^{(j)} r_{it}$$

2. Run an IV regression of $\bar{r}_i = T^{-1} \sum_{t=1}^T r_{it}$ on the regressors $x_i^{(1)} = (\hat{\beta}_i^{(1)}, (\hat{\beta}_i^{(1)} - \hat{\beta}_i^{(2)})' A_1)'$ with instruments $z_i^{(1)} = (\hat{\beta}_i^{(3)}, (\hat{\beta}_i^{(3)} - \hat{\beta}_i^{(3)})' A_1)'$, with A_1 a non-random $k_v \times K$ matrix of rank k_v , where the last is the supposed dimension of the factor structure in the error terms. Let $\hat{\lambda}^{(1)}$ be the TSLS estimate of the coefficient on regressor $\hat{\beta}_i^{(1)}$.
3. Repeat step (2) three more times exchanging indexes 1 to 4 circularly; that is, the 2nd regression is an IV regression of \bar{r}_i on regressors $x_i^{(2)} = (\hat{\beta}_i^{(2)}, (\hat{\beta}_i^{(2)} - \hat{\beta}_i^{(3)})' A_2)'$ with instruments $z_i^{(2)} = (\hat{\beta}_i^{(4)}, (\hat{\beta}_i^{(4)} - \hat{\beta}_i^{(1)})' A_2)'$; denote the estimate $\hat{\lambda}^{(2)}$, etc.
4. Obtain the four-split estimate as:

$$\hat{\lambda}_{4S} = \frac{1}{4} \sum_{j=1}^4 \hat{\lambda}^{(j)}$$

Appendix B

B Mispricing

Table B.1: Simulation results –Giglio and Xiu (2021) (Panel A) and Lettau and Pelger (2020) (Panel B) methodologies This table presents the performance of the t -statistic in terms of *size* and *power* for different numbers of latent factors in the estimation (1, 3, and 5) for several specifications of N and T when the asset pricing model is misspecified. Data are simulated according to the 5-factor model of Fama and French (2015). Factors are de-noised using the first 5 principal components. We calibrate the mispricing $\hat{\alpha}$ from the empirical pricing errors.

Panel A: Giglio and Xiu (2021) methodology

N	T	#PC	Size					Power				
			MKT	SMB	HML	RMW	CMA	MKT	SMB	HML	RMW	CMA
200	200	1	0.071	0.081	0.966	0.999	0.959	0.417	0.389	0.362	0.345	0.376
		3	0.084	0.099	0.190	0.233	0.138	0.318	0.211	0.439	0.270	0.303
		5	0.070	0.100	0.142	0.134	0.109	0.376	0.232	0.216	0.396	0.258
	400	1	0.073	0.087	0.999	1.000	0.999	0.678	0.668	0.652	0.651	0.652
		3	0.074	0.142	0.321	0.339	0.206	0.505	0.322	0.674	0.440	0.485
		5	0.074	0.147	0.259	0.214	0.187	0.603	0.367	0.370	0.607	0.393
	600	1	0.074	0.103	1.000	1.000	1.000	0.842	0.834	0.825	0.826	0.835
		3	0.097	0.194	0.432	0.373	0.264	0.647	0.422	0.775	0.541	0.607
		5	0.084	0.193	0.301	0.268	0.247	0.754	0.462	0.485	0.738	0.523
400	400	1	0.058	0.258	1.000	1.000	0.999	0.655	0.629	0.629	0.610	0.634
		3	0.046	0.180	0.165	0.501	0.261	0.560	0.144	0.390	0.175	0.205
		5	0.052	0.109	0.135	0.129	0.115	0.582	0.308	0.351	0.578	0.365
	600	1	0.060	0.308	1.000	1.000	1.000	0.811	0.803	0.802	0.791	0.803
		3	0.068	0.252	0.212	0.630	0.336	0.725	0.208	0.490	0.221	0.270
		5	0.068	0.156	0.181	0.169	0.149	0.747	0.416	0.456	0.737	0.517
	1200	1	0.067	0.499	1.000	1.000	1.000	0.981	0.980	0.981	0.979	0.981
		3	0.070	0.360	0.291	0.809	0.475	0.941	0.331	0.639	0.321	0.376
		5	0.072	0.230	0.272	0.246	0.258	0.957	0.628	0.598	0.884	0.685
600	600	1	0.056	0.577	1.000	1.000	1.000	0.816	0.793	0.803	0.787	0.805
		3	0.056	0.276	0.169	0.605	0.371	0.744	0.161	0.420	0.168	0.181
		5	0.051	0.130	0.155	0.150	0.153	0.748	0.448	0.429	0.632	0.481
	1200	1	0.063	0.819	1.000	1.000	1.000	0.976	0.977	0.976	0.976	0.976
		3	0.061	0.455	0.245	0.765	0.502	0.950	0.235	0.586	0.266	0.285
		5	0.062	0.187	0.231	0.228	0.217	0.953	0.636	0.620	0.835	0.694

Panel B: Lettau and Pelger (2020) methodology

N	T	#PC	Size					Power				
			MKT	SMB	HML	RMW	CMA	MKT	SMB	HML	RMW	CMA
200	200	1	0.112	0.115	0.954	0.984	0.957	0.426	0.394	0.368	0.342	0.380
		3	0.080	0.120	0.368	0.138	0.253	0.325	0.312	0.611	0.601	0.567
		5	0.069	0.114	0.118	0.142	0.076	0.381	0.190	0.178	0.407	0.158
	400	1	0.080	0.092	0.991	0.997	0.991	0.696	0.686	0.662	0.652	0.665
		3	0.075	0.175	0.541	0.231	0.422	0.517	0.496	0.767	0.838	0.817
		5	0.071	0.151	0.187	0.249	0.116	0.612	0.311	0.271	0.626	0.261
	600	1	0.076	0.111	0.999	0.998	1.000	0.863	0.854	0.845	0.834	0.849
		3	0.103	0.244	0.636	0.276	0.535	0.653	0.622	0.837	0.910	0.894
		5	0.088	0.207	0.233	0.321	0.158	0.768	0.384	0.375	0.692	0.325
400	400	1	0.068	0.276	0.993	0.996	0.992	0.663	0.643	0.638	0.609	0.641
		3	0.056	0.117	0.502	0.127	0.337	0.515	0.432	0.815	0.776	0.800
		5	0.056	0.114	0.115	0.121	0.094	0.621	0.277	0.257	0.490	0.238
	600	1	0.060	0.330	0.998	1.000	1.000	0.817	0.810	0.805	0.790	0.809
		3	0.085	0.167	0.595	0.192	0.450	0.680	0.565	0.864	0.879	0.898
		5	0.078	0.154	0.146	0.153 _V	0.093	0.782	0.382	0.338	0.620	0.315
	1200	1	0.068	0.507	1.000	1.000	1.000	0.984	0.983	0.983	0.982	0.983
		3	0.080	0.269	0.718	0.286	0.628	0.913	0.775	0.931	0.976	0.955
		5	0.074	0.226	0.218	0.251	0.158	0.969	0.592	0.453	0.809	0.459
600	1	0.056	0.586	0.999	1.000	1.000	0.821	0.801	0.806	0.780	0.809	
	3	0.065	0.173	0.600	0.168	0.458	0.660	0.617	0.882	0.805	0.902	
	5	0.055	0.148	0.150	0.092	0.117	0.770	0.449	0.370	0.414	0.318	

Table B.3: Simulation results –Anatolyev and Mikusheva (2021) (Panel A) and Gagliardini et al. (2016) (Panel B) methodologies This table presents the performance of the t -statistic in terms of *size* and *power* for several specifications of N and T when the asset pricing model is misspecified. Data are simulated according to the 5-factor model of Fama and French (2015). Factors are de-noised using the first 5 principal components. We calibrate the mispricing $\hat{\alpha}$ from the empirical pricing errors.

Panel A: Anatolyev and Mikusheva (2021) methodology											
N	T	Size					Power				
		MKT	SMB	HML	RMW	CMA	MKT	SMB	HML	RMW	CMA
200	200	0.058	0.067	0.109	0.104	0.070	0.298	0.154	0.074	0.172	0.080
	400	0.048	0.061	0.125	0.093	0.056	0.438	0.179	0.076	0.148	0.069
	600	0.044	0.068	0.122	0.093	0.049	0.543	0.195	0.084	0.136	0.051
400	400	0.047	0.064	0.140	0.106	0.077	0.508	0.237	0.071	0.151	0.086
	600	0.058	0.075	0.155	0.101	0.066	0.636	0.274	0.084	0.161	0.078
	1200	0.050	0.073	0.148	0.072	0.048	0.844	0.253	0.073	0.104	0.052
600	600	0.056	0.085	0.168	0.131	0.108	0.676	0.298	0.089	0.166	0.107
	1200	0.058	0.075	0.165	0.114	0.097	0.860	0.314	0.080	0.140	0.094

Panel B: Gagliardini et al. (2016) methodology											
N	T	Size					Power				
		MKT	SMB	HML	RMW	CMA	MKT	SMB	HML	RMW	CMA
200	200	0.112	0.449	0.723	0.859	0.908	0.349	0.535	0.731	0.872	0.908
	400	0.159	0.605	0.819	0.921	0.954	0.549	0.662	0.814	0.925	0.954
	600	0.214	0.695	0.865	0.943	0.953	0.664	0.719	0.862	0.945	0.961
400	400	0.114	0.496	0.728	0.904	0.949	0.513	0.645	0.696	0.897	0.949
	600	0.187	0.624	0.780	0.950	0.962	0.632	0.741	0.754	0.948	0.964
	1200	0.339	0.773	0.874	0.969	0.984	0.774	0.821	0.867	0.964	0.988
600	600	0.105	0.515	0.730	0.949	0.948	0.671	0.684	0.675	0.947	0.950
	1200	0.248	0.707	0.828	0.968	0.988	0.812	0.783	0.811	0.978	0.980

Figure B.1: Test size when loadings are weak/semi-strong and pervasive. $N = 400, T = 400$.

The figure plots the size of the t -statistics for PCA-based estimators and estimators related to observable factors when loadings are weak and pervasive and mispricing is introduced.

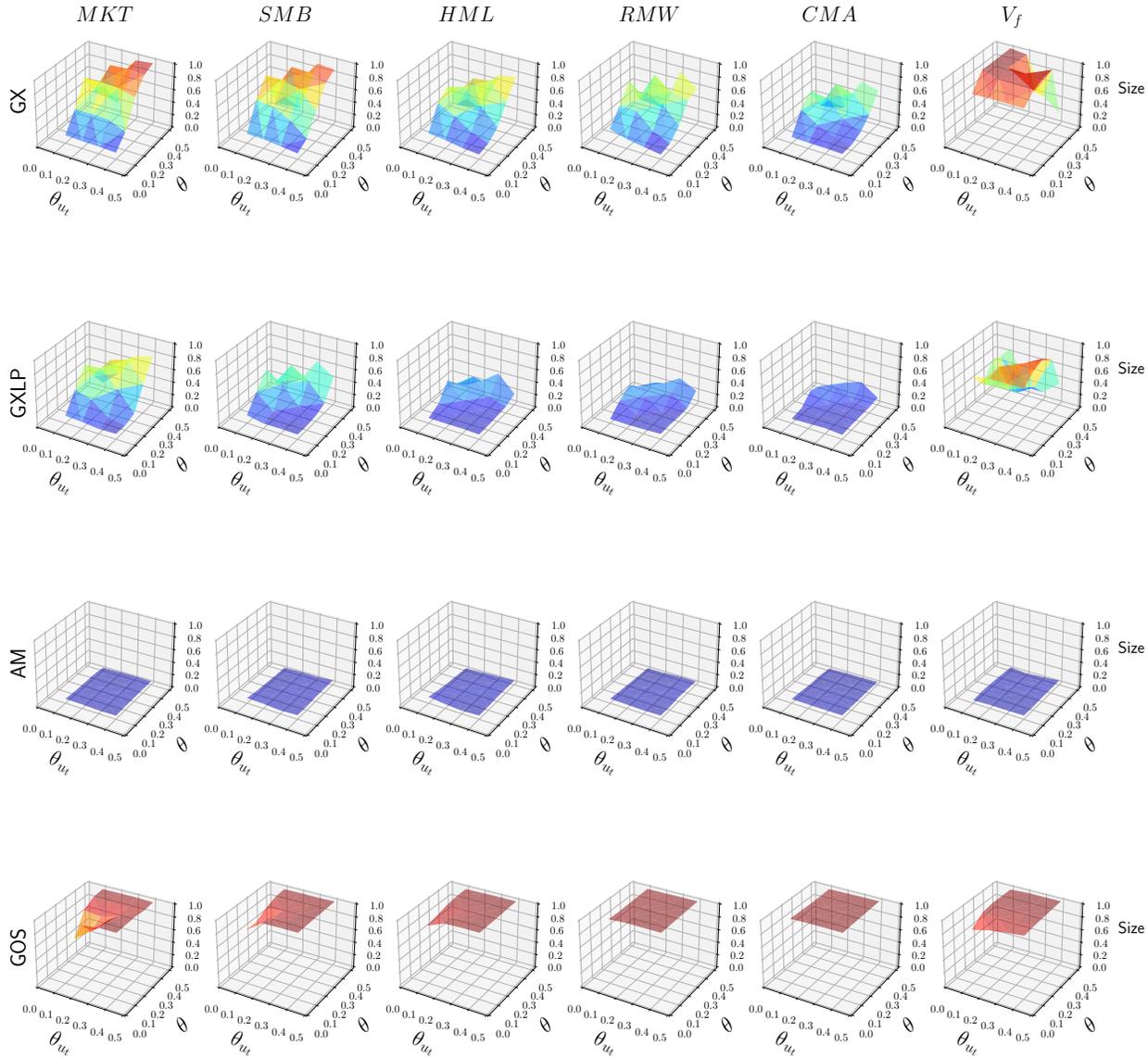


Figure B.2: Test power when loadings are weak/semi-strong and pervasive. $N = 400$, $T = 400$.

The figure plots the power of the t -statistics for PCA-based estimators and estimators related to observable factors when loadings are weak and pervasive and mispricing is introduced.

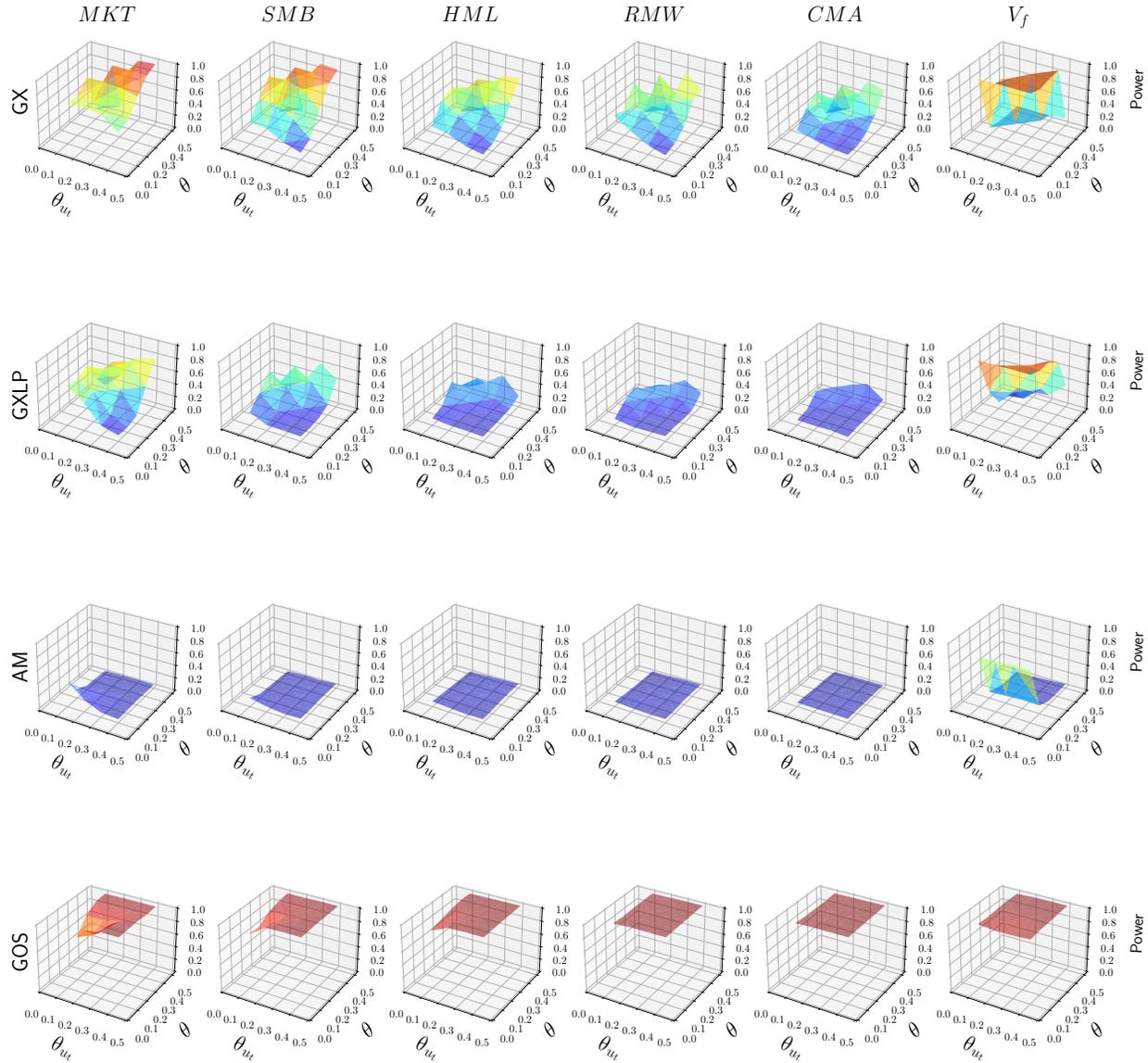


Figure B.3: Test size when loadings are sparse. $N = 400, T = 400$.

The figure plots the size of the t -statistics for PCA-based estimators and estimators related to observable factors when loadings are sparse and mispricing is introduced.

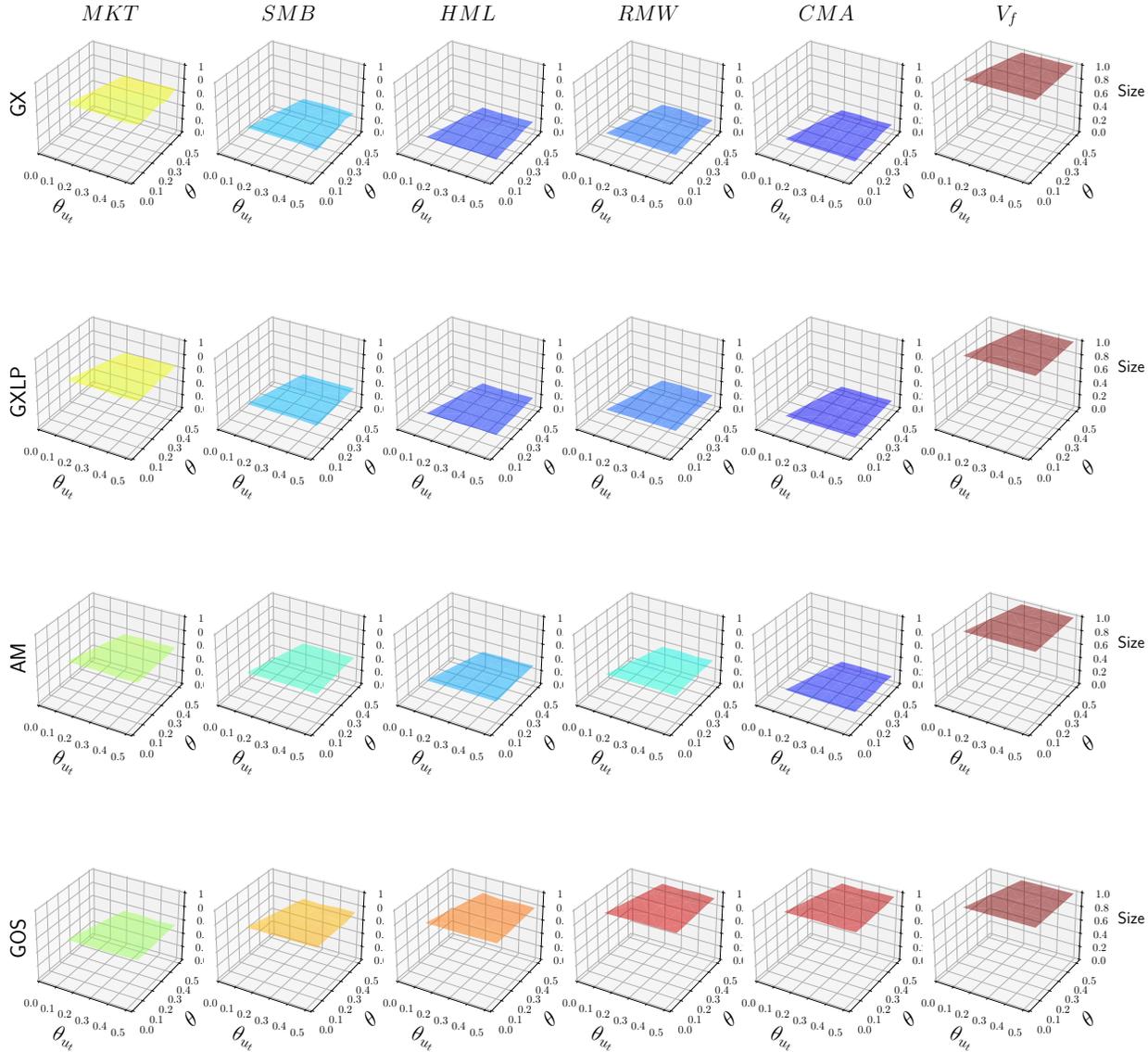
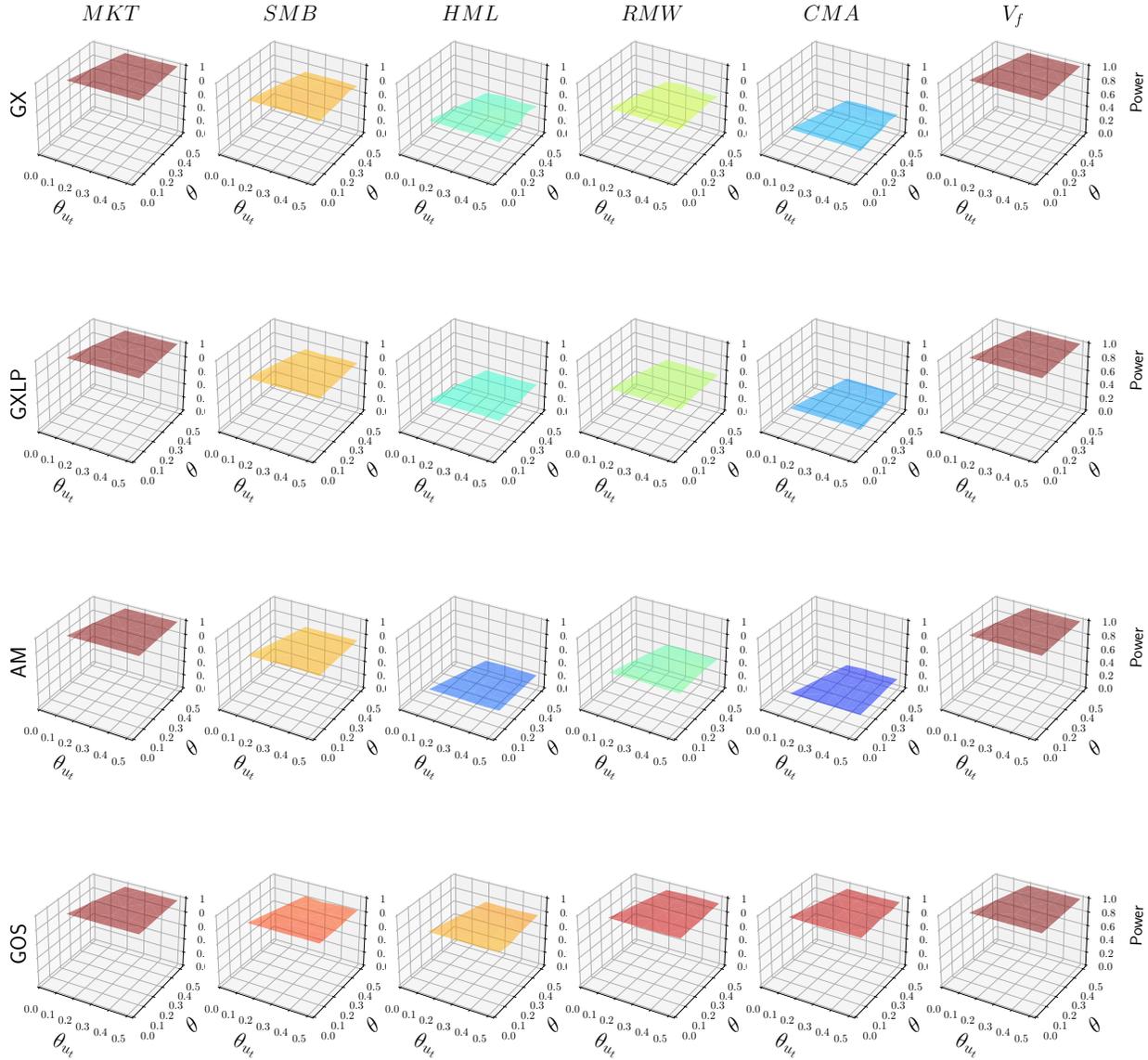


Figure B.4: Test power when loadings are sparse. $N = 400, T = 400$.

The figure plots the size of the t -statistics for PCA-based estimators and estimators related to observable factors when loadings are sparse and mispricing is introduced.



Appendix C

C Anatolyev and Mikusheva (2021)'s calibrations

In this calibration strategy, we generate returns following Eq.(7)-(8).

In detail, we first estimate the risk premia and the average of the four matrices of exposures following their proposed methodology based on observable factors. We then calibrate the population risk premia and exposure to these estimates, respectively $\check{\lambda} = 1/4 \sum_i^4 \hat{\lambda}_i^{AM}$, $\check{\beta} = 1/4 \sum_i^4 \hat{\beta}_i^{AM}$.

Related to the structure error terms (Eq.(8)), after the estimation of the residualized (and demeaned) component, the u 's, we extract two principal components from it. To ensure these PCs are to be considered strong (or rather the loading κ 's), we scale their strength to be proportional to the Frobenius norm of the factor structure of the observable FF5. These PCs constitute the calibrated unobserved factor structure of the error terms.

Finally, after residualizing the two-factor structures (the observed and unobserved strong ones), the remaining component dictates the idiosyncratic errors, the e 's, which we used to calibrate the variance (imposing to be homoskedastic).

Given a number \check{N} of observed returns $r_t^{(\check{N})}$ and the estimates coming from Anatolyev and Mikusheva (2021)'s methodology, the 5×1 simulated FF5 factors F_t^\diamond , the 2×1 simulated latent factors v_t , and the 1×1 idiosyncratic errors e_t^\diamond , are generated from a multivariate normal as follows:

$$\begin{pmatrix} F_t^\diamond \\ v_t^\diamond \\ e_t^\diamond \end{pmatrix} \sim \mathcal{N} \left(\begin{pmatrix} \bar{F} \\ 0_{(2+1) \times 1} \end{pmatrix}, \begin{pmatrix} \widehat{\Sigma}^F & 0 & 0 \\ 0 & \widehat{\Sigma}^v & 0 \\ 0 & 0 & \widehat{\sigma}^e \end{pmatrix} \right), \quad (12)$$

where \bar{F} is the 5×1 vector of sample mean FF5, $\widehat{\Sigma}^F$ is the 5×5 sample covariance matrix of the FF5, $\widehat{\Sigma}^v$ is the 2×2 sample covariance matrix of the v_t and $\widehat{\sigma}^e$ is the sample variance of the idiosyncratic error.

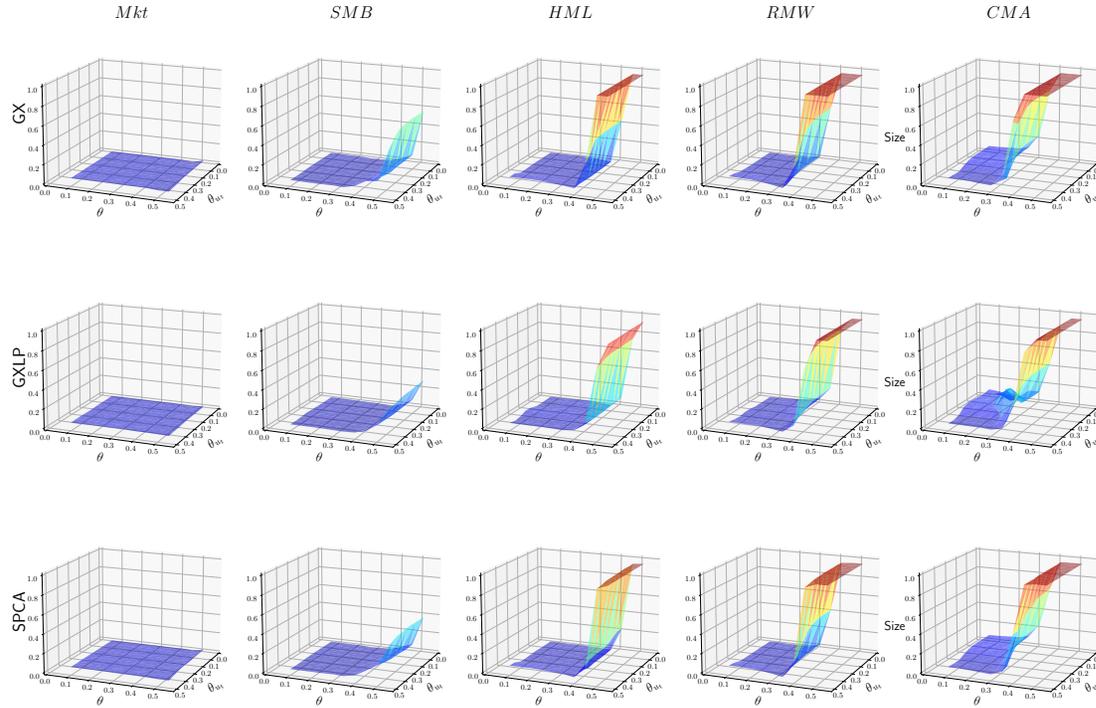
Through Eq.(7)-(8), the realizations of the i of the \check{N} simulated returns $r_{i,t}^\diamond$ is:

$$r_{i,t}^\diamond = \check{\beta}_i \check{\lambda} + \check{\beta}_i (F_t^\diamond - \bar{F}_t^\diamond) + \hat{\kappa}_i v_t^\diamond + e_t^\diamond$$

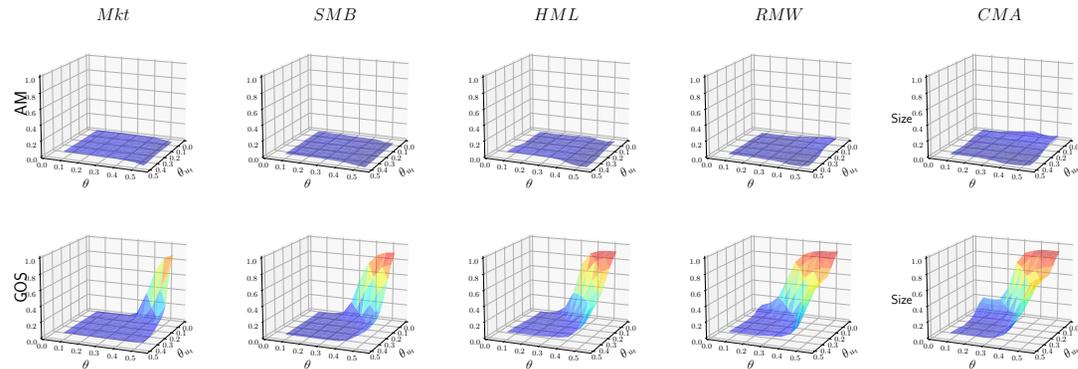
This description regards the case dubbed as Strong Factors. For the other scenarios, we follow analogous configurations to Section 3.2. The unique change is in the Local and Strong/Semi-Strong Loadings scenario, where we change the exposure related to the *HML* factor, rather than *SMB*, for a broader variety of results. Here below, we present the Weak/Semi-Strong, the Sparse and the Local and Strong/Semi-Strong Loadings scenarios.

Figure C.5: Test size when loadings are weak/semi-strong and pervasive. $N = 400, T = 400$.

The figure plots the size of the t -statistics for PCA-based estimators (Panel A) and estimators related to observable factors (Panel B) when loadings are weak and pervasive.



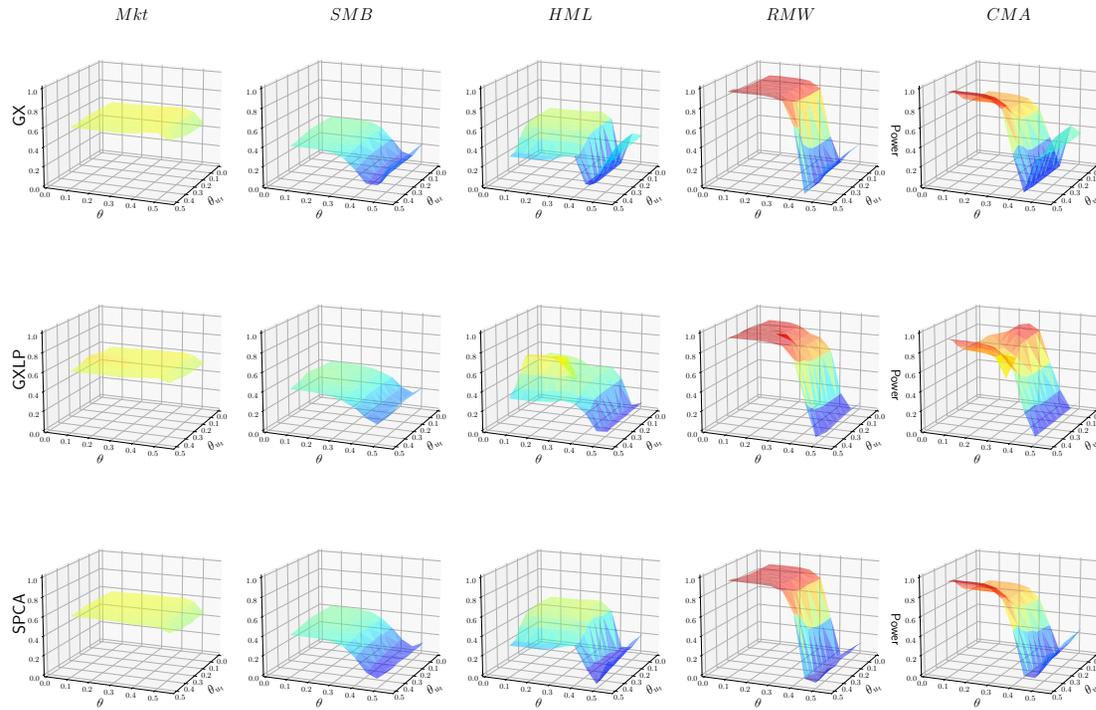
(a) Panel A



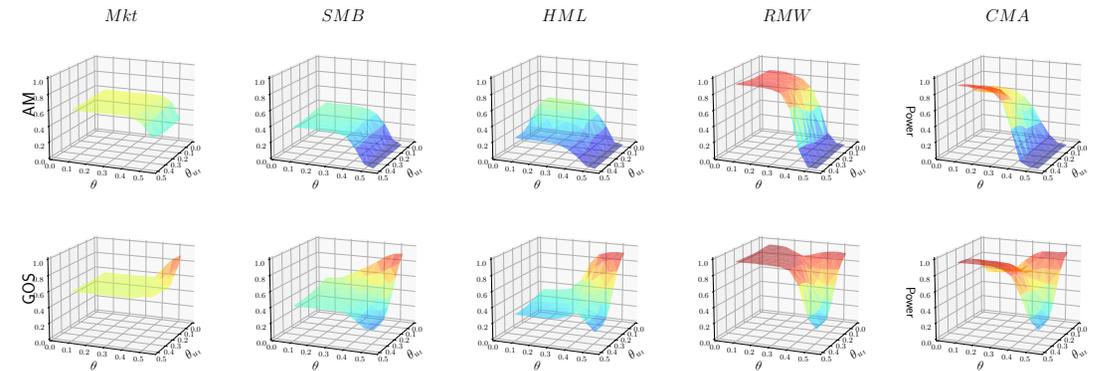
(b) Panel B

Figure C.6: Test power when loadings are weak/semi-strong and pervasive. $N = 400$, $T = 400$.

The figure plots the power of the t -statistics for PCA-based estimators (Panel A) and estimators related to observable factors (Panel B) when loadings are weak and pervasive.



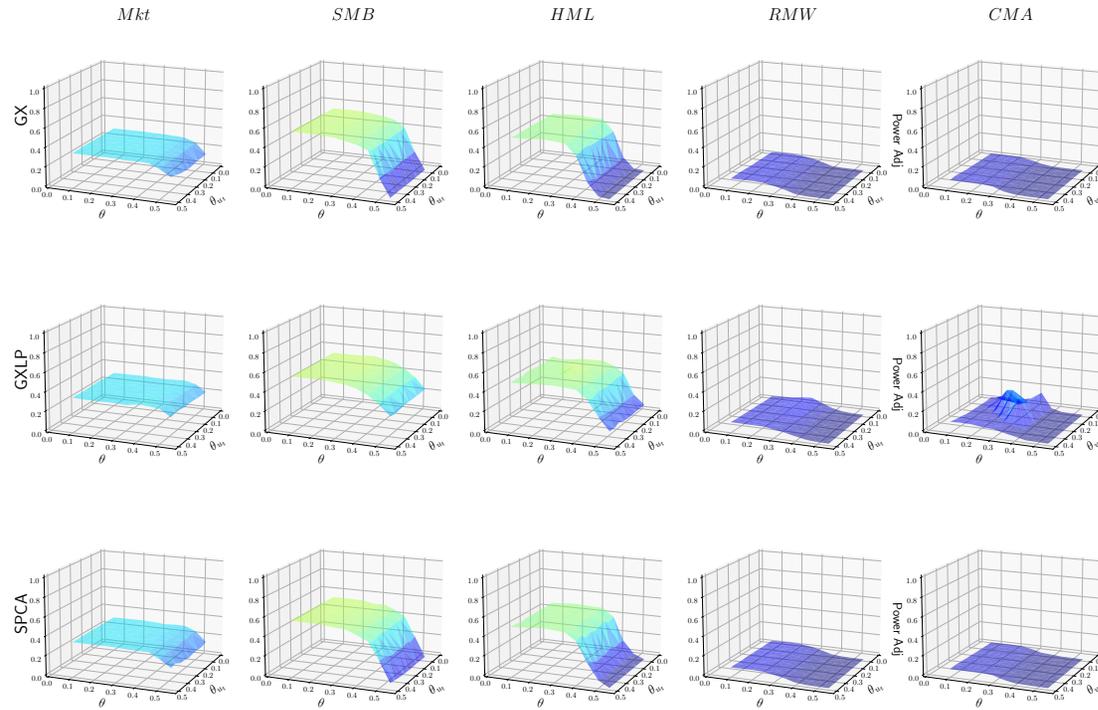
(a) Panel A



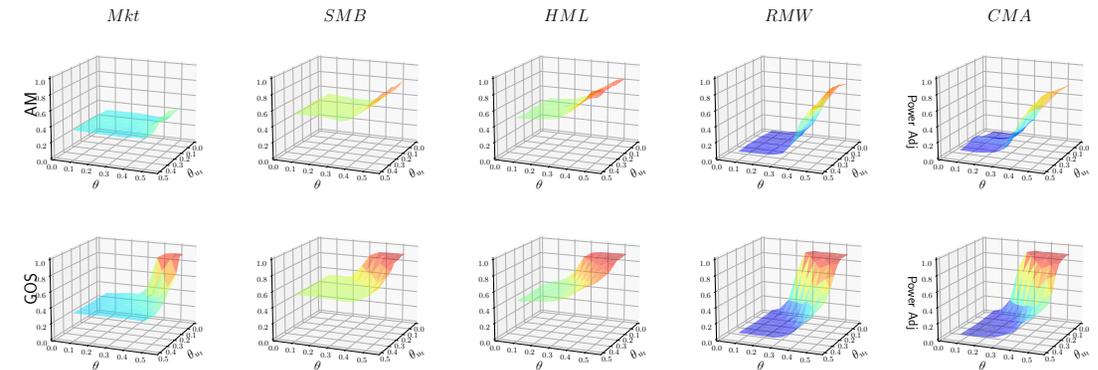
(b) Panel B

Figure C.7: Test size-adjusted power when loadings are weak/semi-strong and pervasive. $N = 400, T = 400$.

The figure plots the size-adjusted power of the t -statistics for PCA-based estimators (Panel A) and estimators related to observable factors (Panel B) when loadings are weak and pervasive.



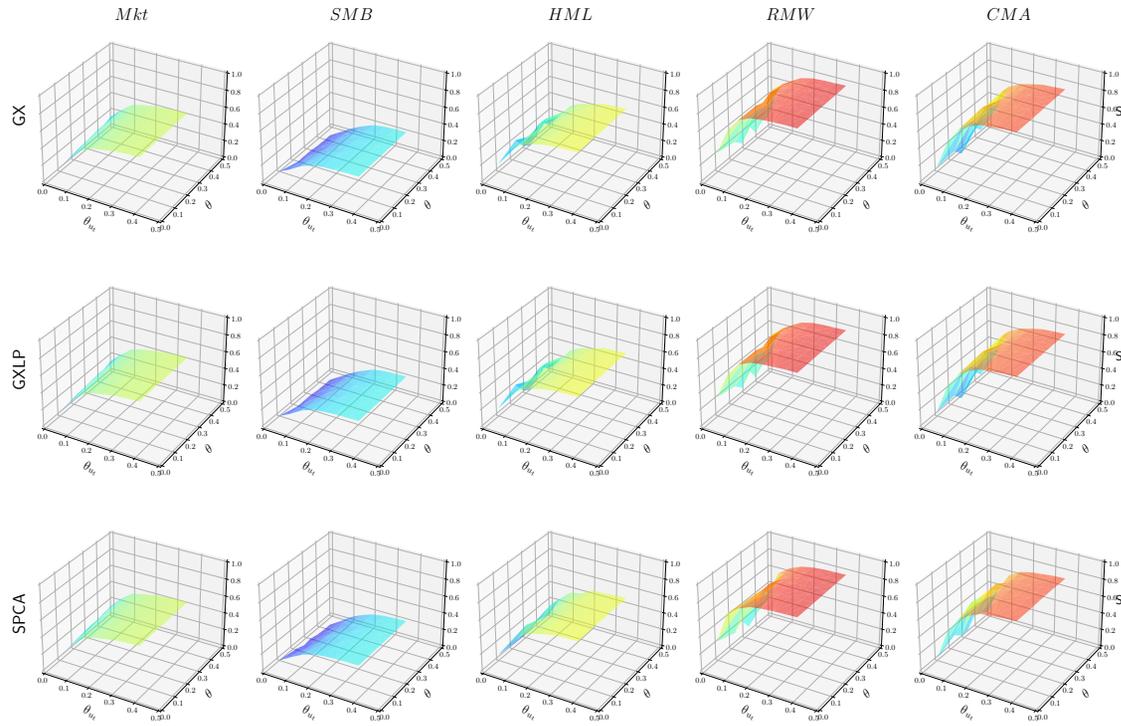
(a) Panel A



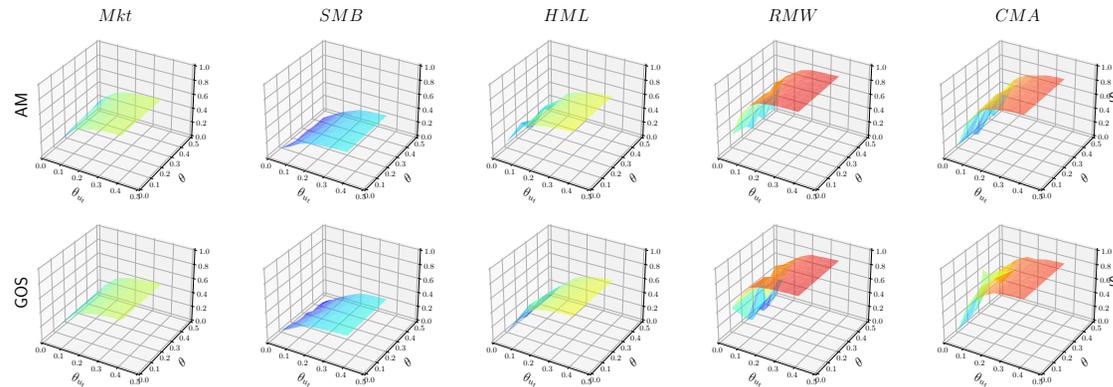
(b) Panel B

Figure C.8: Test size when loadings are sparse. $N = 400, T = 400$.

The figure plots the size of the t -statistics for PCA-based estimators (Panel A) and estimators related to observable factors (Panel B) when loadings are sparse.



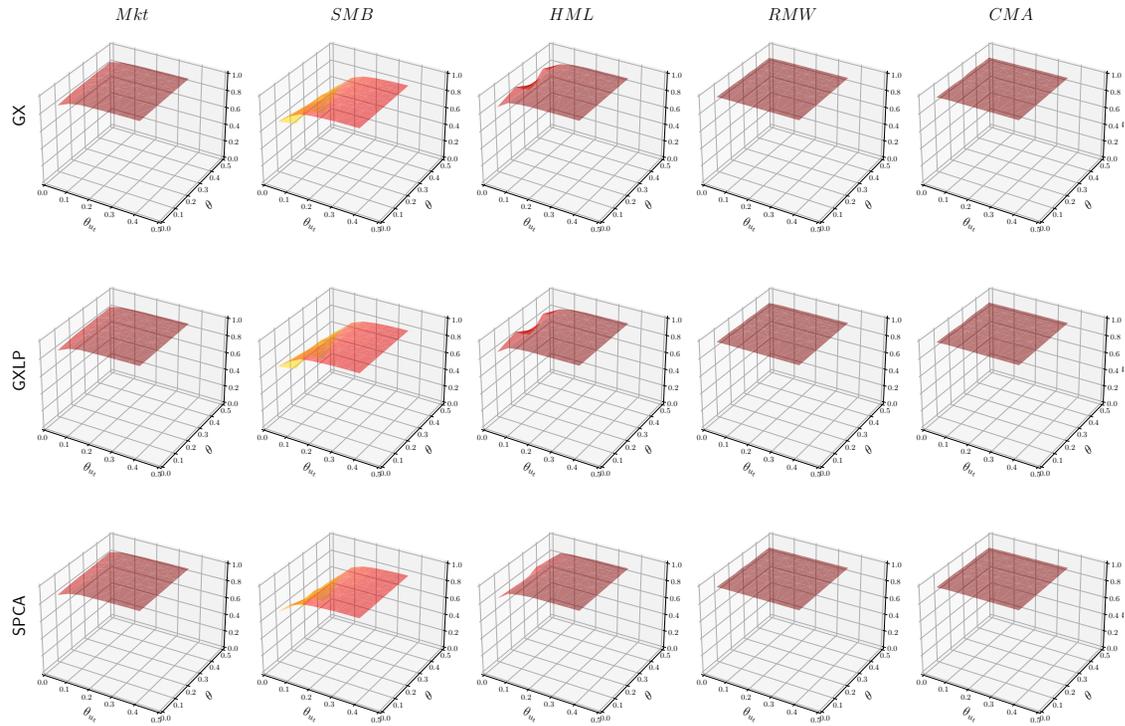
(a) Panel A



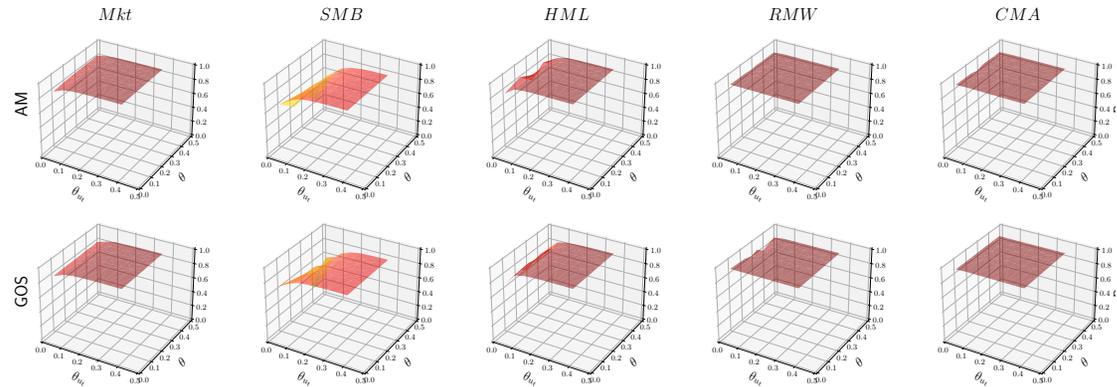
(b) Panel B

Figure C.9: Test power when loadings are sparse. $N = 400$, $T = 400$.

The figure plots the power of the t -statistics for PCA-based estimators (Panel A) and estimators related to observable factors (Panel B) when loadings are sparse.



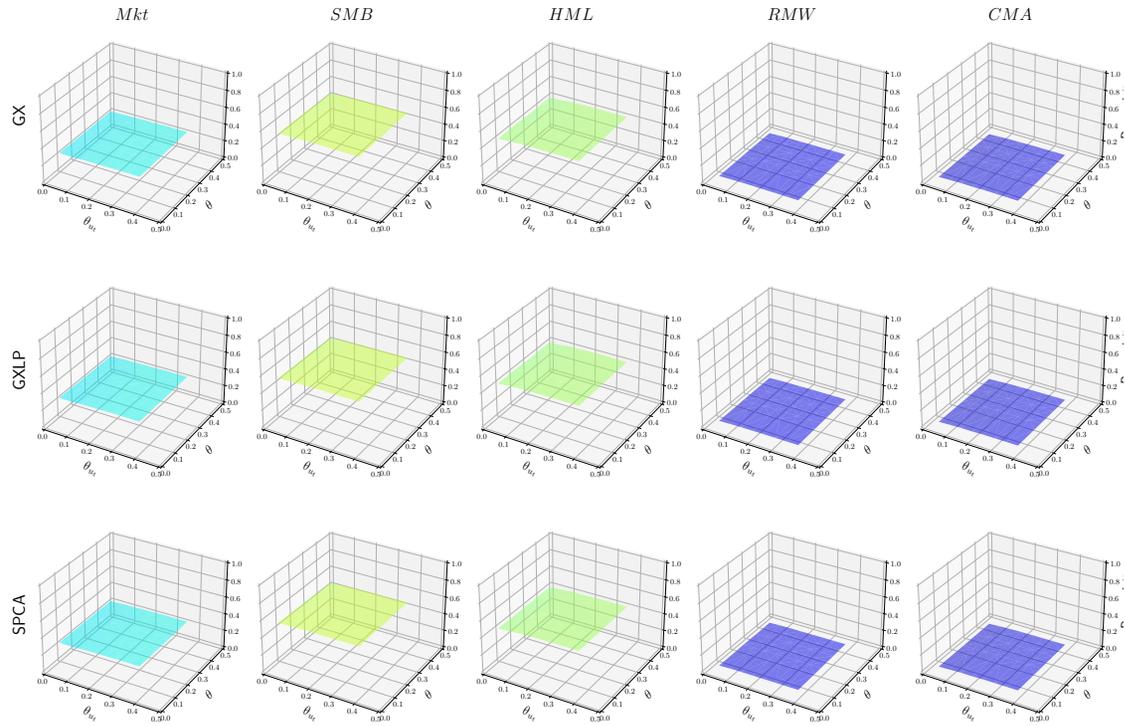
(a) Panel A



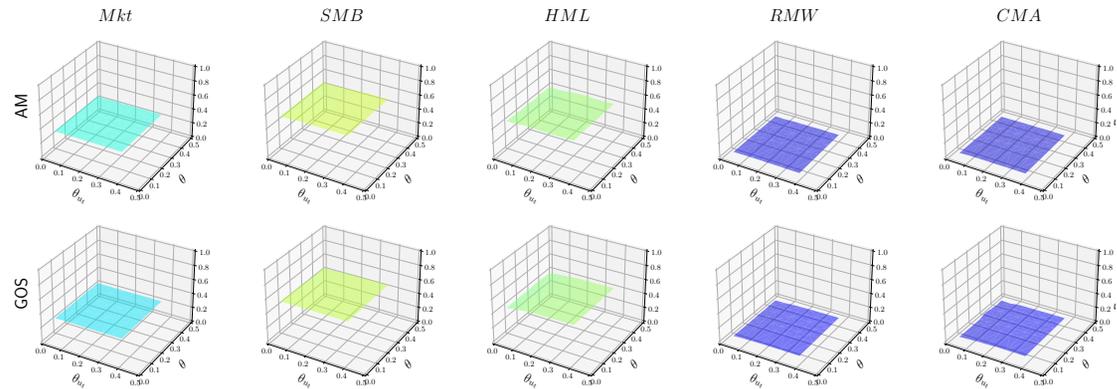
(b) Panel B

Figure C.10: Test size-adjusted power when loadings are sparse. $N = 400, T = 400$.

The figure plots the size-adjusted power of the t -statistics for PCA-based estimators (Panel A) and estimators related to observable factors (Panel B) when loadings are sparse.



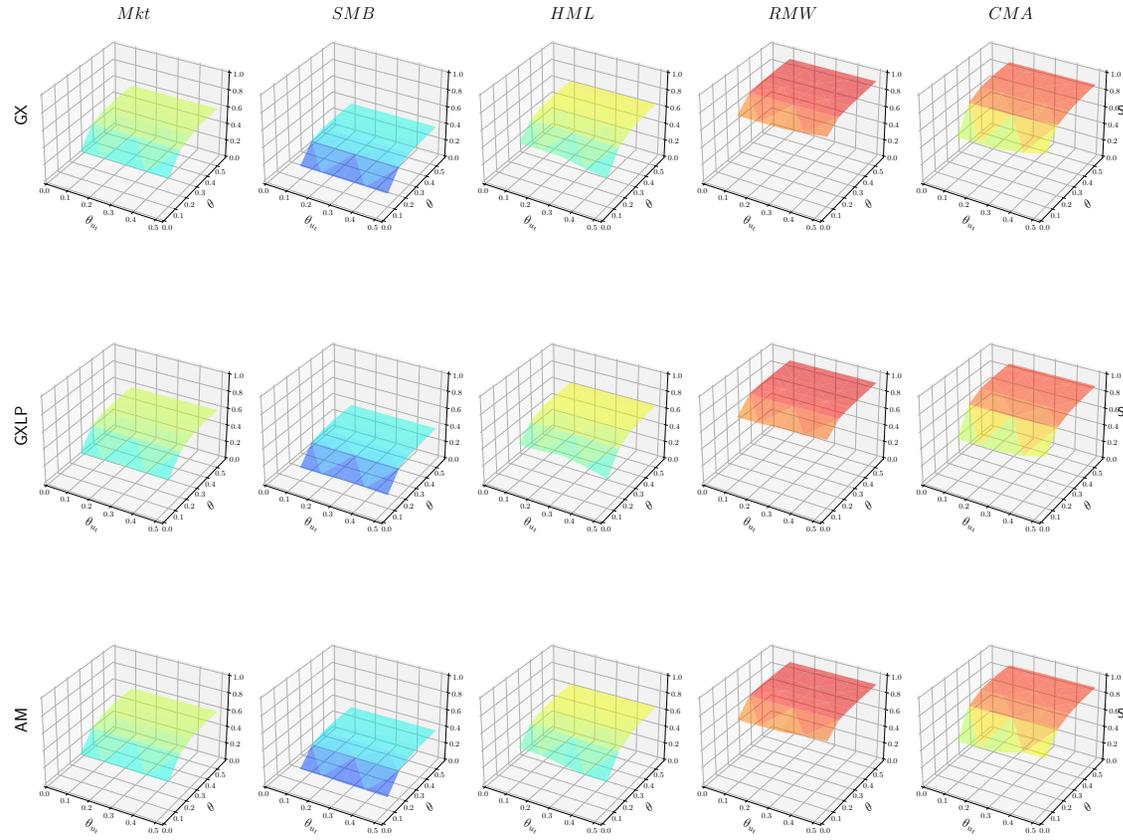
(a) Panel A



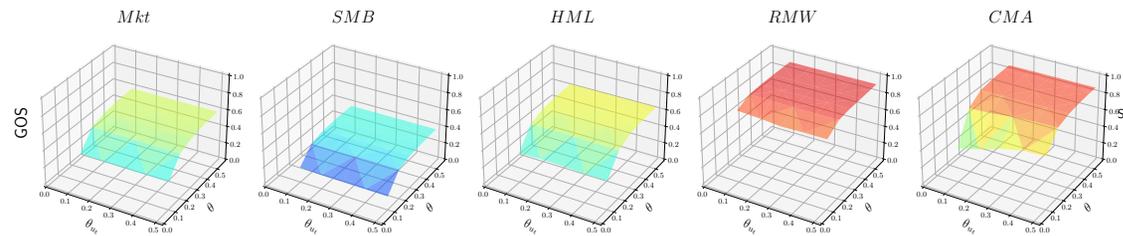
(b) Panel B

Figure C.11: Test size when loadings are local and strong/semi-strong. $N = 400$, $T = 400$.

The figure plots the size of the t -statistics for PCA-based estimators and Anatolyev and Mikusheva (2021)'s (Panel A) and Gagliardini et al. (2016) estimator (Panel B) when loadings are local and strong/semi-strong.



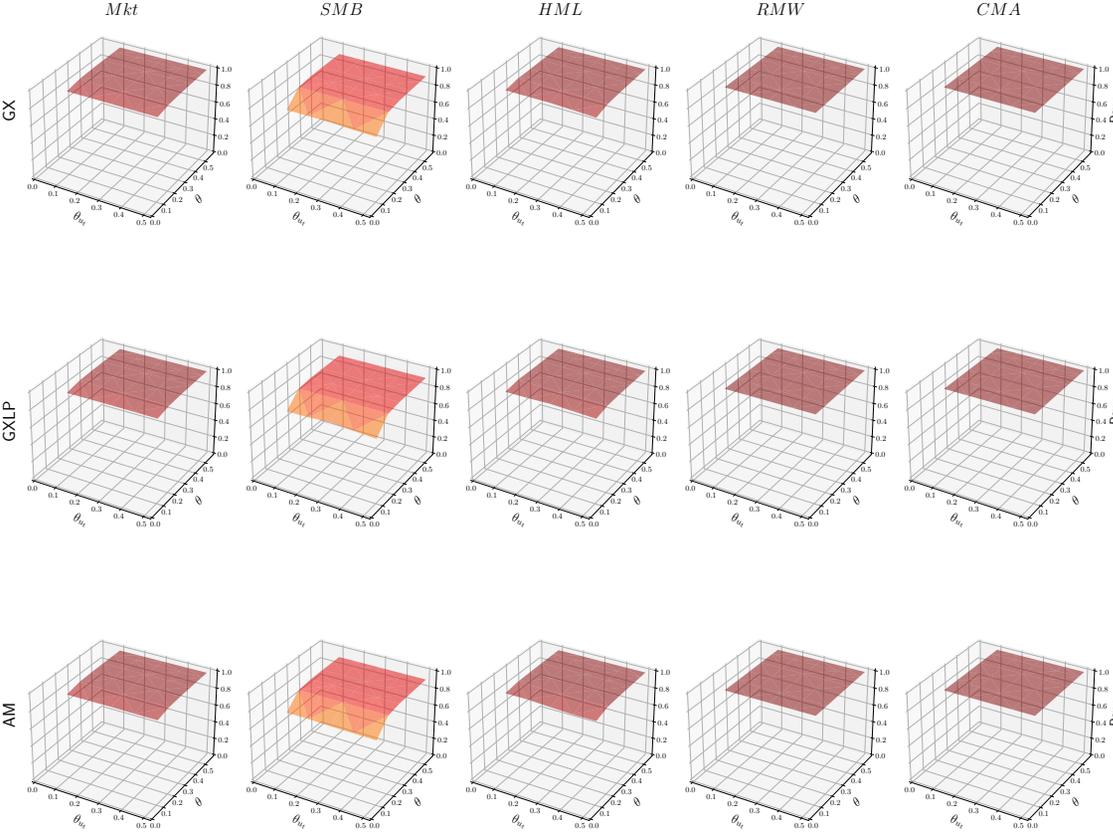
(a) Panel A



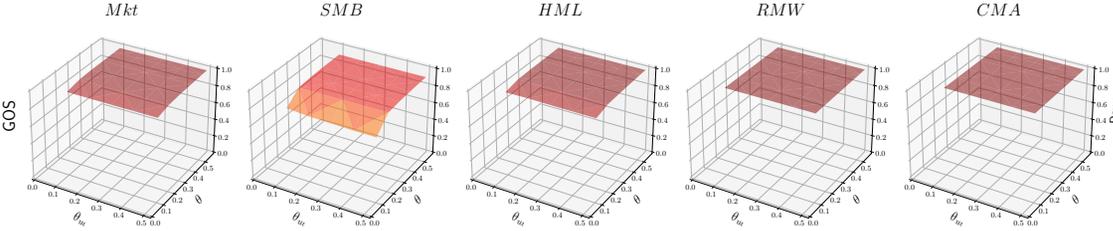
(b) Panel B

Figure C.12: Test power when loadings are local and strong/semi-strong. $N = 400, T = 400$.

The figure plots the power of the t -statistics for PCA-based estimators and Anatolyev and Mikusheva (2021)'s (Panel A) and Gagliardini et al. (2016) estimator (Panel B) when loadings are local and strong/semi-strong.



(a) Panel A



(b) Panel B

Figure C.13: Test size-adjusted power when loadings are local and strong/semi-strong. $N = 400$, $T = 400$.

The figure plots the size-adjusted power of the t -statistics for PCA-based estimators and Anatolyev and Mikusheva (2021)'s (Panel A) and Gagliardini et al. (2016) estimator (Panel B) when loadings are local and strong/semi-strong.

