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

Amedeo Andriollo, Cesare Robotti, Giulio Rossetti *

June 7, 2023

Preliminary draft, comments welcome!

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 Codes: C12, C38, C53, G12

*E-mails: amedeo.andriollo@warwick.ac.uk; cesare.robotti@wbs.ac.uk; giulio.rossetti.19@mail.wbs.ac.uk; Robotti and Rossetti are at Warwick Business School. Andriollo is in the Economics Department of the University of Warwick. We thank Stefano Giglio and Dacheng Xiu for sharing their data and codes. We would also like to thank Anna Mikusheva, Stanislav Anatolyev, Elisa Ossola, and Olivier Scaillet for sharing their codes. Amedeo Andriollo is grateful to Eric Renault for helpful discussions and comments.

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 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 cross-sectional structure mimics the one observed in the data. In different 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 (e.g., Giglio and Xiu (2021) and Giglio et al. (2022)).

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 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 Appendix and an online Appendix.

2 A Review of Recent Methods

This section provides an overview of the available methods in a large-N setting. In the first subsection, we focus on three methodologies that involve both latent and observable factors. Next, we discuss two approaches that only involve observable factors.

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) (GX) and Giglio et al. (2022) (SPCA) share the same setup. They presume an $N \times 1$ vector of test asset excess returns, r_t, to follow the latent linear factor model

$$r_t = \alpha + \beta\gamma + \beta v_t + u_t, \quad \mathbb{E}[v_t] = \mathbb{E}[u_t] = \mathbb{E}[v_t u_t'] = 0, \alpha = 0, \;\; \forall t \in \mathbb{Z}, \tag{1}$$

where β is a $N \times p$ matrix of factor exposures, v_t is a $p \times 1$ vector of factors' innovations, and u_t is an $N \times 1$ vector of idiosyncratic errors. The model is latent since we do not observe the innovations $\{v_{k,t}\}_{k=1,..,p}$. GX assume that expected excess returns are proportional to the factor exposures, which means that there is no mispricing[1] in the latent model since

$$\mathbb{E}[r_t] = 0 + \beta\gamma. \tag{2}$$

[1]In their Online Appendix, GX provide an alternative set of results 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 will briefly discuss the case of mispricing.

GX relate observable and latent factors as follows:

$$g_t = \delta + \eta v_t + w_t, \quad \mathbb{E}[w_t] = \mathbb{E}[w_t v_t'] = \mathbb{E}[w_t u_t'] = 0, \ \ \forall t \in \mathbb{Z}, \tag{3}$$

which is essentially a projection of the observable factor, whose price of risk we are interested in, onto the latent factors. GX refer to the residual of such regression, w_t, as the measurement error in the factor.[2] The risk premium implied by Eqs. (1)-(3) is given by $\gamma_g = \eta\gamma$.

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 exploit well-known results in the econometric literature for recovering the latent factor space, which consists in estimating v_t via PCA from the time covariance matrix of the test assets (or equivalently one can estimate β from the cross-section covariance matrix). Under their assumptions and double asymptotics, the estimated latent factors are consistent for their population counterparts and converge at an appropriate rate. The other parameters in Eq. (1) are then estimated using the Fama-Macbeth procedure, while the parameters in Eq. (3) are estimated via time-series regression. Please, refer to Appendix A.1 for a detailed description of GX's procedure. Up to a rotation, such latent factors or PCs are the unique factors in population that price the test assets r_t or, to cite Anatolyev and Mikusheva (2021), "*[the method] successfully eliminates strong missing factors, but assumes from the outset that all important pricing factors can be uncovered by PCA*". Moreover, their " *results hold under similar or weaker assumptions compared to those in Bai (2003) [...] One notable assumption is the so-called pervasive condition for a factor model*". Given 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 (which allow to separate signal, i.e., the factors, from noise). The high-level assumptions 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.

Indeed, one of the most pivotal assumptions is on the strength of the latent factors.

[2]In fact, GX label the fitted part coming from the PCs space, $\hat{g}_t = \hat{\delta} + \hat{\eta}\hat{v}_t$, as the denoised factor.

This assumption is discussed in Giglio et al. (2022), where an alternative methodology is proposed to allow for weak (or rather sparse, asymptotically) factors to enter the latent asset pricing formulation in Eq. (1). They show that if $N/\lambda_{min}(\beta'\beta)T \not\to 0$, most of the existing estimators proposed in the literature might not yield a consistent estimate of the risk premium, as the latent factor structure might not be correctly recovered.

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. Namely, 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 initial step, 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, looking at the most recent version of their codes, the implementation proposed by Giglio et al. (2022) is based on a grid search of the optimal tuning parameters through cross-validation by considering the out-of-sample R^2 of the mimicking portfolio Eq. (3): "*[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*". We also underline that hedging might constrain the estimates to a distinct goal which is different from pricing: "*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.* "(See Antoine et al. (2020))

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) this is not the case, as 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 being above a certain threshold. 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

[3]Please, see Appendix A.2 for a thorough description of the SPCA method.

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, we argue that, differently from Giglio and Xiu (2021), SPCA may suffer of the problem of factor omission in the risk premium analysis. In particular, for the asymptotically normal behavior of the estimator, Giglio et al. (2022) "*need g_t to have at least equal number of variables as the true number of 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 nonnested asset pricing models[4], with corresponding risk premia that are sensitive to the chosen formulation of the model.

Since Giglio et al. (2022) assume the existence of a subset of assets where the factors are strong, we also would like to stress that they are ruling out other possibly weak factors with varying rates of strength, as discussed for instance in Freyaldenhoven (2021). On a more technical side, the assumptions in place for guaranteeing the good asymptotic behaviour of the proposed estimator are not necessarily mild. First, in order to easily reuse the three-step methodology (Appendix A.2), they assume that the subset of test assets for which the latent factor is strong is non-random. Second, to guarantee consistency in the mechanical selection with respect to quantiles, they assume that their proposed selection procedure is consistently recovering, at each step, the uniquely identified subsets (and thus the factors). Last, the assumptions in place bound the trade-off in the asymptotic rate between the N total assets and the number of selected ones. The latter needs to explode fast enough to guarantee consistency of the PCA, but slow enough to ensure the correct asset selection.

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

Following Anatolyev and Mikusheva (2021) and Giglio et al. (2022), we modify the first step of the procedure of Giglio and Xiu (2021) by incorporating the latent factor estimator proposed by Lettau and Pelger (2020) (GXLP). The authors argue that (latent) factors with large Sharpe-ratios but with a relatively small variance-covariance structure are poten-

[4]In particular, their "*Assumption A.3 implies that there exists a subset of test assets, within which all latent factors are strong* "(Giglio et al. (2022)). In population, they model such subset of test assets as a function of the L_∞ norm of a peculiar linear combination between β's and η's. In such way, the observed factors shape the latent factor model.

tially not recovered by conventional PCA.[5] Given the importance of having explanatory power in both co-movements and pricing error terms, their main intuition is to combine the first and second moments in a unique objective function in order to estimate the latent factors.

Denote by R the $N \times T$ matrix of excess returns, whose $(i,t)-$entries are $\{r_{i,t}; i = 1,\dots,N, t = 1,\dots,T\}$ with $\bar{r} = T^{-1}\sum_{t=1}^{T} r_t$. The estimation strategy consists in recovering the latent factors, $\{f_t\}$, by applying PCA to the covariance matrix of the returns with overweighted mean:

$$\frac{1}{T}R^\top R + \mu\bar{r}\bar{r}^\top, \tag{4}$$

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

$$\min_{\{\beta_i\},f_t} \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\tilde{r}_{i,t} - \beta_i v_t\right)^2 + (1+\mu)\frac{1}{N}\sum_{i=1}^{N}\left(\bar{r}_i - \beta_i \bar{f}\right)^2, \tag{5}$$

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 authors generalize the above objective function by considering

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

where the case in Eq. (4) obtains when $\tilde{Q} = I_T$. In our simulations, we will focus on the $\tilde{Q} = I_T$ case because Lettau and Pelger (2020) provide asymptotic results for this scenario. However, following Lettau and Pelger (2020), we also consider the case where $\tilde{Q}$ is the inverse of the diagonal matrix of standard deviations of the test asset returns[6], motivated by the fact that "*usually, estimation based on the correlation matrix is more robust than based on the covariance matrix as it is less affected by a few outliers with very large variances*."(See Lettau and Pelger (2020) and Choi and Yang (2022))

[5]"*Conventional PCA [...] only uses information contained in the second moment but ignores all information in the first moment*."(Lettau and Pelger (2020))

[6]"*The statistical rationale is that certain cross-sectional observations contain more information about systematic risk than others and hence should obtain larger weight in the statistical analysis. The standard deviation of each cross-sectional observation serves as a proxy for how large the noise is and therefore down-weighs very noisy observations*."(Lettau and Pelger (2020))

Finally, to estimate the risk premia, once the latent factors have been estimated, we adopt the second and third steps of Giglio and Xiu (2021) as outlined in the previous section.

Since the procedure borrows part of the strategy (and the asymptotic theory) of Giglio and Xiu (2021), the previous comments on Giglio and Xiu (2021) apply. Moreover, we would like to emphasize several issues.

First, 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 achieving the largest correlation for a specific factor*."(Lettau and Pelger (2020)) 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 $\mu \to \infty$, as reported in Giglio et al. (2022) (Proposition 4), the estimator "*can be robust to weak factors if the information about β from the expected return $\beta\gamma$ dominates the information from return covariances*."

Second, 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., Fig. 8 in Lettau and Pelger (2020) and Figs. 1-2 in Anatolyev and Mikusheva (2021))

Last, the (high-level) assumptions set in place for achieving desirable asymptotic properties are stronger than the standard ones.[7]

[7]See, for example, their Assumption 2.C on the exposures, β, and Assumptions 1.C.3 and 2.D on the time dependence of the residuals.

2.2 Methods with observable factors

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

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

$$r_t = \beta\lambda + \beta(F_t - \mathbb{E}[F_t]) + u_t, \quad \forall t \in \mathbb{Z}, \tag{6}$$

where the $p \times 1$ vector of factors, F_t, are observable and no longer latent. The model is assumed to be correctly specified, and the asset pricing restriction (Eq.(2)) holds.

To estimate risk premia from a large cross-section, the authors propose solutions that are robust to two scenarios: (i) the observable factors are weak; (ii) (sufficiently) strong cross-sectional dependence among the error terms, u_t. Namely, they assume that some of the factors may enter the model with a local-to-zero (i.e., of order $O(T^{-1/2})$) exposure coefficient, and they impose a latent factor structure on the (unobserved) error terms:

$$u_{i,t} = v_t'\kappa_i + e_{i,t}, \quad \forall i, t \tag{7}$$

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 latter assumption is that many papers show that the post-estimation residuals usually have a non-trivial factor structure which, however, is thought to carry no risk premia.[8]

To estimate the parameter of interest, the authors propose a sample-splitting IV methodology, based on the intuition that, if the sources of potential bias coming from the exposures estimation (i.e., the first step of Fama and MacBeth (1973)) 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 (i.e., the second stage of Fama and MacBeth (1973)). In other words, they propose to first split the sample along the time series dimension into four non-overlapping subsets, and then estimate the subsample exposures via a (standard) time-series regression. After estimating the subsample exposures, they suggest to run an IV cross-sectional regression, using the estimated exposures as instrument, by repeating this stage in a circular fashion, to get four estimates

[8]*"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))

of the risk premia. Lastly, the average of those four risk premia estimates represents the four-split estimate of the risk premia. (Please, refer to Appendix A.4 for more details on the algorithm)

Several remarks are in order. First, 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, which means that the observable factors are the unique drivers of the risk premia, thus rendering the procedure subject to an omitted variable problem. Second, 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 at most local-to-zero, but not zero in population. Last, 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.

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

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

$$r_t(\gamma) = \beta_t(\gamma)'\lambda_t + \beta_t(\gamma)'(F_t - \mathbb{E}[F_t|\mathcal{F}_{t-1}]) + \varepsilon_t(\gamma), \quad \forall t \in \mathbb{Z}, \tag{8}$$

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[9].

To estimate the parameters of interest, the authors further specialize 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 pass consists in an OLS time series regression of excess returns on the factors. The second pass consists in a cross-sectional weighted least squares (WLS)

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

regression (augmented with the factor mean) with respect to a trimming device: an indicator function that guarantees the avoidance of ill-conditioning, which depends on a function of the eigenvalues corresponding to the covariance matrix of the first pass estimates.[10] (Please, refer to their work for further details)

A couple of remarks are needed. First, 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, 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 competitor of the various methods described above. Second, the properties of this estimation strategy when spurious factors are present in the analysis have not been investigated. Thus, a careful examination is needed.

3 Data Generating Processes

This section presents the Data Generating Processes we will 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 risk premium analysis is then carried out for the weak, spurious, and semi-strong factor cases. In these two sections, we assume that the underlying model is correctly specified. Finally, in Section 3.3, we briefly discuss scenarios in which the asset pricing restriction does not hold.

Likewise, in Appendix B, we study all the aforementioned scenarios tuning the DGPs in agreement with 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, a scenario where all estimators should do fairly well at least for reasonably large sample sizes.

Specifically, we simulate a five-factor DGP based on the Fama and French (2015) (FF5:

[10] For the definition of the objects, please refer to Section 3.2 of Gagliardini et al. (2016).

Mkt, SMB, HML, RMW, CMA), with respect to two different calibration strategies: Giglio and Xiu (2021) (latent AP model), and Anatolyev and Mikusheva (2021) (observable AP model). Regarding the latter calibration strategy, please refer to Appendix B.

In the first calibration strategy, the returns are generated according to Eq. (1), following Giglio and Xiu (2021). In other words, the factors that generate the dynamics in the returns are calibrated to match the denoised[11] version of the five Fama and French (2015)'s. The factors (v_t^{FF5}) are latent and we observe a noisy version of them (as specified in Eq. (3)). We calibrate the η and the measurement error w_t as in Giglio and Xiu (2021).

Given a number $\breve{N}$ of observed returns $r_t^{(\breve{N})}$, the realizations of the $\breve{N} \times 1$ simulated returns $r_t^{\diamond}$, together with the 5×1 simulated denoised factors $v_t^{\diamond}$, and the 5×1 simulated measurement errors $w_t^{\diamond}$, are generated from a multivariate normal as follows:

$$\begin{pmatrix} r_t^{\diamond} \\ v_t^{\diamond} \\ w_t^{\diamond} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \bar{r} \\ 0_{(2\times 5)\times 1} \end{pmatrix}, \begin{pmatrix} \widehat{\Sigma}^r & \widehat{\Sigma}^{r,v} & 0 \\ \widehat{\Sigma}^{v,r} & \widehat{\Sigma}^v & 0 \\ 0 & 0 & \widehat{\Sigma}^w \end{pmatrix} \right), \tag{9}$$

where $\bar{r}$ is the $\breve{N} \times 1$ vector of sample meanreturns, $\widehat{\Sigma}^r$ is the $\breve{N} \times \breve{N}$ sample covariance matrices 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, while $\widehat{\Sigma}^{r,v}$ is the $\breve{N} \times 5$ 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)-(3), i.e., $\{\breve{\beta}, \breve{\gamma}, \hat{u}_t^{(\breve{N})}\}$, are then obtained as

$$\breve{\beta} = (\widehat{\Sigma}^v)^{-1}(\widehat{\Sigma}^{r,v})', \qquad \breve{\gamma} = (\breve{\beta}'\breve{\beta})^{-1}\breve{\beta}'\bar{r}$$
$$\hat{u}_t^{(\breve{N})} = r_t - \breve{\beta}\breve{\gamma} - \breve{\beta} v_t^{FF5}.$$

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

This design allows us to differentiate our analysis from the simulation studies that

[11]See Giglio and Xiu (2021) for details about the denoising procedure. We extract the first p PCs, i.e., the $\{\widehat{v}_{i,t}\}_{i=1,..,p}$ (Step 1 of Giglio and Xiu (2021); see Appendix A.1). Then, after estimating the parameters in Eq. (3), we define the de-noised and demeaned five Fama and French (2015)'s factors as the particular rotations of the latent factors $v_t^{FF5} = \widehat{\eta}\widehat{v}_t$. We run our simulations with respect to different degrees of denoising, i.e., $p = \{5, 7, 9\}$. However, we report the results only for the case when the number of factors used to denoise is five. The results for the complete set of simulations are available upon request.

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 an artificial factor structure (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)) In the interest of space, the simulations results for the small-N cases are relegated to the Online Appendix. To preserve the heterogeneous factor structure of the original dataset, we always include bonds and currencies portfolios in the selected returns when fixing $\breve{N}$ and calibrating the parameters. Finally, in our simulations, we maintain the condition $\breve{N} \leq \breve{T}$ when generating our artificial returns. For instance, when $\breve{N} = 400$, we then generate samples corresponding to $\breve{T} = \{400, 600, 1200\}$.[12]

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., $\beta_k = cov(g_{k,t}, r_t)/var(g_{k,t}) = 0_{n\times 1}$. As for *weak* factors, we need to consider a variety of cases (Freyaldenhoven (2021)):

- Pervasive/global weak or semi-strong loadings (Onatski (2012)), $\beta = O_p(N^{-1/2})$. We refer the readers to Bai and Ng (2021) for a comprehensive discussion on why we

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

should refer to loadings rather than factors. 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).[13]

- Locally strong loadings (Freyaldenhoven (2021) and Giglio et al. (2022)): $\beta_i = \{\beta_{i,0}, \beta_{i,\not{g}}\}$, with $[\beta_{i,0}]_i = 0$ and $[\beta_{i,\not{g}}]_i = O_p(1)$. The betas are sparse, since $[\beta_{i,0}]_i = 0$, but, if the model is restricted to the assets corresponding to the exposure $\beta_{\not{g}}$, such factors are strong.

Clearly, in such cases, the two-pass procedure might not correctly pin down the risk premia on the proposed factors because the covariance variance matrix of the betas may be singular (spurious factors) or nearly singular up to some rate (weak factors). In summary, the problem of identifying risk premia could be traced back to three scenarios: *spurious factors*, *weak factors* (or loadings), *sparse* loadings, and *semi-strong* factors (or loadings). In our simulations, we consider the following scenarios when returns are simulated with 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 $\breve{g}_t$ with a spurious factor that is uncorrelated with the test asset returns. In practice, we consider $g^{\diamond}_{6,t}$ such that

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

   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 is not entering 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, 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.[14] This factor is observed with precision

[13] See Connor and Korajczyk (2022) and Antoine et al. (2020).

[14] We 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.

(i.e., the measurement error w_t is set equal to zero to avoid additional distortions), is orthogonal to the other observable factors, and is relatively weaker than the others, 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, 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 (opposite) rate of convergence.[15] Since $\breve{\beta}\breve{\beta}'/\breve{N}^{1-2\theta_1} = I_{\breve{N}}$, we have the following relationship between the rate and the signal-to-noise ratio:

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

where the signal-to-noise $= \text{tr}(\Sigma_{\hat{v}})/\text{tr}(\Sigma_{\hat{u}})$. Thus, if we set the different rates of convergence, it corresponds to different signal-to-noise ratios. In particular,

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

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 $\breve{N} \times T$ matrix of residuals $\hat{u}_t^{(\breve{N})}$ (and its $(\breve{N} \times \breve{N})$ variance-covariance matrix). Given its relationship, we scale both $\breve{\beta}$ and $\hat{u}_t^{(\breve{N})}$, which means we will consider

$$\breve{\beta}^{\star} = \breve{\beta} \times \breve{N}^{-\theta}$$
$$\hat{u}_t^{\star} = \hat{u}_t^{(\breve{N})} \times \breve{N}^{-\theta_{u_t}}.$$

where $\breve{\beta}^{\star}$ (and $\hat{u}_t^{\star}$) 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 weak loading behaviour.[16] Subsequently, we generate the new returns, $\breve{r}_t^{\star}$, through the asset pricing model in Eq. (1) with the rescaled exposures, denoised six factors, and rescaled error terms,

[15] If 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), $\breve{\beta}\breve{\beta}' \to I_{\breve{N}}$, i.e., $\breve{\beta} = O_p(\breve{N}^{-\theta})$.

[16] An implicit assumption is that the initial betas are strong loadings.

while maintaining the (unscaled) risk premia $\breve{\gamma}$:

$$\breve{r}_t^\star = \breve{\beta}^\star \breve{\gamma} + \breve{\beta}^\star \hat{v}_t + \hat{u}_t^\star. \tag{10}$$

The simulations are then based on Eq. (11) and $\{\breve{r}_t^\star\}$.

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.[17] In practice, we multiply entry-wise the $\breve{N} \times 1$ loading $\breve{\beta}_{Mkt}$ 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

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

with θ_{u_t} ranging as before. We generate the new returns, $\breve{r}_t^\star$, accordingly to Eq. (10). The simulations are then based on Eq. (11) with respect to $\{\breve{r}_t^\star\}$.

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.[18] 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 $\breve{N} \times 1$ loadings $\breve{\beta}_{SMB}$ 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 $\breve{N}^{-\theta}$ (i.e.,

[17] We 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.

[18] We refer to Section 4.5, where the reasons of why this scenario is relevant are explained.

SMB is a semi-strong factor for such assets). We consider ranges of $q \in \{50\}$ and $\theta \in 0.05 \times \{1 : 10\}$.[19] We then have

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

with θ_{u_t} ranging as before. We then generate the new returns, $\breve{r}_t^{\star}$, according to Eq. (10). The simulations are then performed as before, via Eq. (11), with respect to $\{\breve{r}_t^{\star}\}$.

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 milestone work of Hansen and Jagannathan (1997) or the discussion in Antoine et al. (2020)).

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 (we refer to Eq. (11)) is translated by a $(\breve{N} + 2 \times K) \times 1$ vector $\breve{\alpha} = (\hat{\alpha}', 0_{2\times K}\prime)'$ with variance $\sigma_{\breve{\alpha}}^2 = \breve{N}^{-1}\breve{\alpha}'\breve{\alpha}$, with respect to $K = \{5, 6\}$, the number of factor (depending on the scenario), and $\hat{\alpha}$ is the estimated pricing errors, i.e., $\widehat{\alpha} = \widehat{\mu} - \breve{\beta}\breve{\gamma}$. As these results are similar to, and often worse than, those of the benchmark scenarios, we relegate these simulation results to the Online Appendix.

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

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

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.

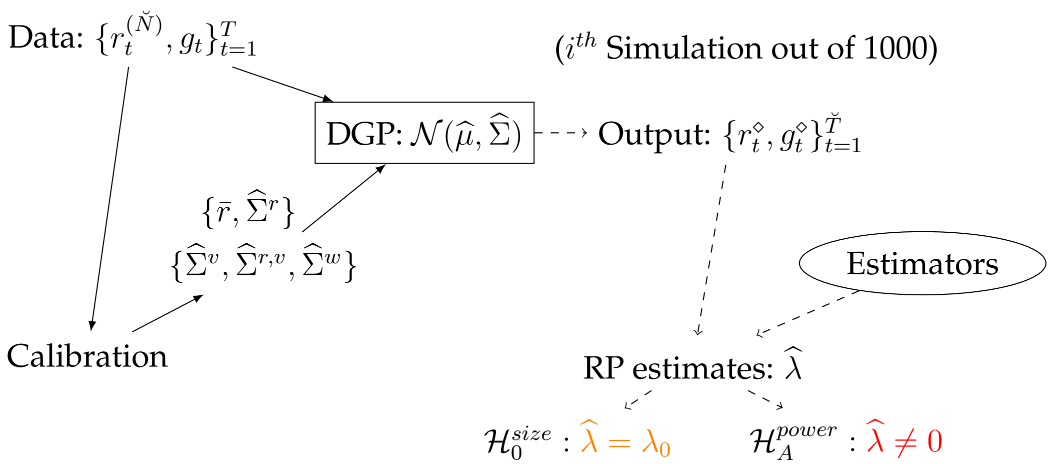


Figure 1: A visual representation of the Monte Carlo simulations for size and power.

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 Online Appendix. In our simulations, once drawn a sample of asset returns and observable factors (i.e., noisy proxies of the true factor structures) that are deemed to be priced, $\{r_t^{\diamond}, g_t^{\diamond}\}_{t=1}^{\breve{T}}$, we can employ one of the estimation strategy presented in Section 2 to get an estimate of the risk premia. Having such estimate, say $\widehat{\lambda}$, when studying size and power, we are testing two different hypotheses:

$$\mathcal{H}_0^{size} : \widehat{\lambda} = \breve{\lambda}$$
$$\mathcal{H}_A^{power} : \widehat{\lambda} \neq 0$$

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 $\breve{\lambda}$), 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 in comparing the t-statistics to the critical values of a standard normal distribution with respect to a certain significance level. Following the literature, we consider hypothesis testing at a 5% significance level. Drawing and repeating the simulation experiments will give us then an empirical distribution of rejection rates: if the asymptotic normality of the estimators hold correctly, we shall expect the empirical size to be around 5%, while having power as large as possible. For ensuring good Monte Carlo properties, all the results are based on 1000 Monte Carlo replications. 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[20].

Three remarks are in order. First and most important, we adopt an *"ubi maior minor cessat"* principle: if the test is not reliable in terms of size, then its power becomes an immaterial feature to evaluate. Thus, when the test is severely oversized, we will not spend many words when describing our results related to power.

Second, if it is not explicitly clear when discussing the PCA-based estimators, we will implicitly estimate the risk premia when the number[21] 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 is requiring to split the sample in four non-overlapping subsets, we consider two possible choices: sequential or non-random splitting (e.g., the first subsample corresponds to the one where $t = (1 : \breve{T}/4)$), and random splitting (e.g., the first subsample is extracted via random draws without replacement).

4.1 Strong and pervasive factors

Tables 1 to Table 6 contain 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

[20] All other simulations are available upon request.

[21] We 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 will not report this here, and it is available upon request.

these five Fama French (2015) factors. When we estimate the model with the methodologies of Giglio and Xiu (2021) and Lettau and Pelger (2020), we evaluate different tuning parameters, i.e., the numbers of latent factors $\hat{p} = \{1 : 12\}$.

The first five columns of each table report 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). We refer to these latter as displaying the empirical power of the test.

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 $\breve{N}$ and $\breve{T}$ 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).

Across the different specifications, we note that the three-step procedure breaks down when the number of principal components used to recover the factor space is less than four (i.e. $\hat{p} < 4$), in particular when we estimate the risk premium of relatively weaker factors. In fact, when $\hat{p} < 4$, the only satisfactory behavior we observe is for the t-test corresponding to the Mkt factor, which is considered to be a strong and pervasive factor. Instead, when the strength (and/or the pervasiveness) of the factors decreases (in relative terms), inference on the estimated risk premia is jeopardized as the empirical size of the t-test is well above the nominal level of the test. For instance, the size of the HML, RMW, CMA factors is, on average, above 90% when $\hat{p} < 4$ and converges to 5% when $\hat{p}$ increases.

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 to the market factor is in line with the nominal level of the test, with power increasing as the time series sample size increases, converging to 1 when $\breve{T}$ is greater than 1200. 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 a similar behaviour as in the case of the market factor (increasing with $\breve{T}$).

Table 1: Size and Power of the test for the risk premia estimated with the GX methodology. The true DGP is the 5-factor model of Fama French (2015). The factors are denoised using the first 7 PCs.

			$N = 400$					$T = 400$		
			Size					Power		
#PC	Mkt	SMB	HML	RMW	CMA	Mkt	SMB	HML	RMW	CMA
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

			$N = 400$					$T = 600$		
			Size					Power		
#PC	Mkt	SMB	HML	RMW	CMA	Mkt	SMB	HML	RMW	CMA
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

			$N = 600$					$T = 600$		
			Size					Power		
#PC	Mkt	SMB	HML	RMW	CMA	Mkt	SMB	HML	RMW	CMA
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

			$N = 600$					$T = 1200$		
			Size					Power		
#PC	Mkt	SMB	HML	RMW	CMA	Mkt	SMB	HML	RMW	CMA
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

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

Table 2: Size and Power of the test for the risk premia estimated with the weighted GXLP methodology ($\mu = 10$). The true DGP is the 5-factor model of Fama French (2015). The factors are denoised using the first 5 PCs.

	$\mu = 10$			$N = 400$			$T = 400$			
			Size					Power		
#PC	Mkt	SMB	HML	RMW	CMA	Mkt	SMB	HML	RMW	CMA
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

	$\mu = 10$			$N = 400$			$T = 600$			
			Size					Power		
#PC	Mkt	SMB	HML	RMW	CMA	Mkt	SMB	HML	RMW	CMA
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

	$\mu = 10$			$N = 600$			$T = 600$			
			Size					Power		
#PC	Mkt	SMB	HML	RMW	CMA	Mkt	SMB	HML	RMW	CMA
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

	$\mu = 10$			$N = 600$			$T = 1200$			
			Size					Power		
#PC	Mkt	SMB	HML	RMW	CMA	Mkt	SMB	HML	RMW	CMA
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

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 to this estimator tend to be not ideal, leading to misleading inference on priced factors when used to price such factor structure.

Table 3: Size and Power of the test for the risk premia estimated with the AM methodology and one potentially missing factor. The true DGP is the 5-factor model of Fama French (2015). The factors are denoised using the first 5 PCs. [circular splitting]

		Size					Power				
N	T	Mkt	SMB	HML	RMW	CMA	Mkt	SMB	HML	RMW	CMA
400	400	0.055	0.090	0.387	0.422	0.166	0.604	0.433	0.184	0.663	0.230
400	600	0.054	0.109	0.505	0.458	0.193	0.776	0.603	0.230	0.669	0.227
400	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
600	1200	0.063	0.122	0.532	0.409	0.264	0.955	0.819	0.161	0.541	0.301

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: Size and Power of the test for the risk premia estimated with the GOS methodology. The true DGP is the 5-factor model of Fama French (2015). The factors are denoised using the first 5 PCs. [circular splitting]

		Size					Power				
N	T	Mkt	SMB	HML	RMW	CMA	Mkt	SMB	HML	RMW	CMA
400	400	0.049	0.193	0.606	0.792	0.802	0.589	0.609	0.375	0.851	0.808
400	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
600	1200	0.083	0.347	0.743	0.914	0.931	0.952	0.818	0.496	0.915	0.927

Table 5: Size and Power of the test for the risk premia estimated with the SPCA methodology. The true DGP is the 5-factor model of Fama French (2015). The factors are denoised using the first 5 PCs.

			N_0			Size					Power				
N	T	Min	Avg	Median	Max	Mkt	SMB	HML	RMW	CMA	Mkt	SMB	HML	RMW	CMA
400	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
400	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
400	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
600	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
600	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

Table 6: Size and Power of the test for the risk premia estimated with the SPCA methodology. The true DGP is the 5-factor model of Fama French (2015). The factors are denoised using the first 5 PCs. [Small N]

			N_0			Size					Power				
N	T	Min	Avg	Median	Max	Mkt	SMB	HML	RMW	CMA	Mkt	SMB	HML	RMW	CMA
50	100	10	27.420	20	50	0.052	0.051	0.051	0.049	0.048	0.228	0.154	0.115	0.068	0.071
50	200	10	39.200	50	50	0.043	0.036	0.047	0.039	0.045	0.404	0.301	0.161	0.101	0.074
50	400	10	42.290	50	50	0.052	0.059	0.056	0.050	0.046	0.632	0.493	0.327	0.136	0.154
50	600	20	44.050	50	50	0.049	0.055	0.045	0.052	0.047	0.779	0.706	0.461	0.212	0.202
50	1200	20	46.460	50	50	0.047	0.054	0.052	0.061	0.046	0.981	0.932	0.728	0.355	0.304
100	100	20	67.380	60	100	0.074	0.053	0.059	0.046	0.042	0.213	0.056	0.251	0.264	0.111
100	200	20	83.460	100	100	0.048	0.054	0.044	0.053	0.046	0.311	0.057	0.425	0.521	0.183
100	400	40	91.800	100	100	0.055	0.043	0.056	0.055	0.053	0.593	0.066	0.690	0.814	0.275
100	600	60	93.840	100	100	0.060	0.060	0.036	0.049	0.045	0.739	0.084	0.891	0.937	0.437
100	1200	60	97.100	100	100	0.050	0.057	0.050	0.053	0.041	0.955	0.090	0.994	0.998	0.711
200	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
200	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
200	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
200	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

4.2 Spurious factors

We also examine the statistical properties of the test when a spurious factor (i.e. a factor that is uncorrelated with test asset returns) is included in observed factors which we deem to price. 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 Online Appendix.

Not surprisingly, when we assess the statistical properties of the PCA-based estimators, we find that for 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 to estimate risk premia for any observable factor, 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. (3)), 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 evidences 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 $\breve{N}$ increases (higher than 26% for $\breve{N} = 600$, while around 18% for $\breve{N} = 400$). 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 $\breve{N} = 600$ (and $\breve{N} = 400$).

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

In general, for all the estimators covered in this work, the standard deviation of the spurious factor does not influence much the size and power of the t-tests, as the size and power 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 consist 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. We remind 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).[22]

We apply a convergence rate on the empirical error term to provide results in terms of a pletora 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 ($\theta \to 0.5$) and the error terms are not scaled ($\theta_{u_t} \to 0$). Vice versa, towards the bottom center of the $(x, y)-$grid, we have the opposite ($\theta \to 0$, $\theta_{u_t} \to 0.5$), meaning that we should tends towards the previous scenario of strong and pervasive factor structure. Regarding the z-axis, warmer colors should signal values that are closer to 1. Fig. 2 below provides a graphical example of how the charts are set up.

[22]The results about the various Sharpe ratios are available upon request.

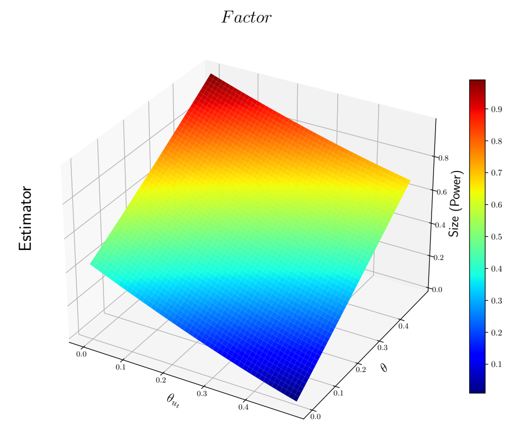


Figure 2: How 3D charts for size (or power) are set up

Figs. 3 and Figure 4 plot respectively the size and power (the z-axis) of the t-statistics for the different estimators as a function of θ on the y-axis and θ_{u_t} on the x-axis, that is our 10×10 grid. (see Section 3.2)

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 artificial returns, the test becomes fairly reliable.

No sign of improvement is shown when test asset selection is performed and the SPCA estimator is adopted.[23] In general, when the factor structure is extremely weak ($\theta \to 0.5$),

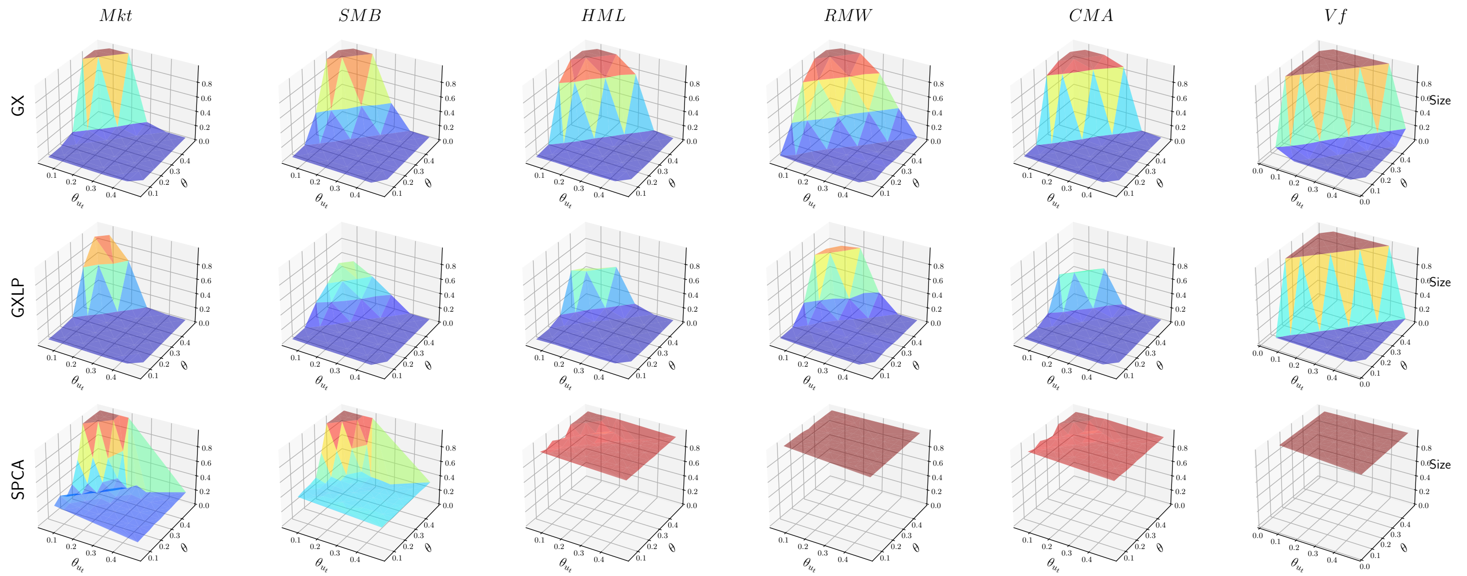
[23]This comes with little to no surprise as the artificial returns in this scenario share the same factor structure. Hence, ideally, the subset of assets that is selected in the first step of the procedure would have the same factor structure as the starting dataset.

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 sign that highly suggest that none of these estimators can provide reliable inference when the factors that drive the cross section of returns are extremely weak.

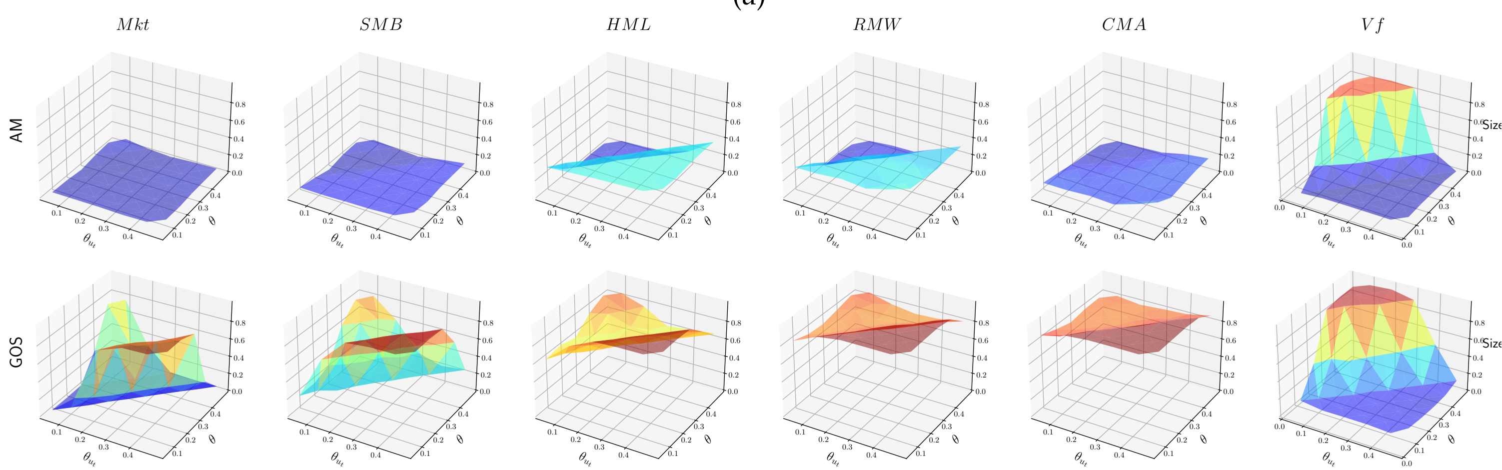
When we look at the performance of AM, we do not see a common pattern across factors when the signal-to-noise ratio is very low. In fact, when testing the estimated risk premium on the market factor, the size of the test is 5.3% while for the other Fama and French (2015)'s factors the empirical sizes 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 stands 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 one discussed in the previous section.

In this case, 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 all 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.



(a)



(b)

Figure 3: Size of the test when loadings are weak and pervasive. $N = 400$, $T = 400$. PCA-based estimators (panel a) and estimators related to observable factors (panel b)

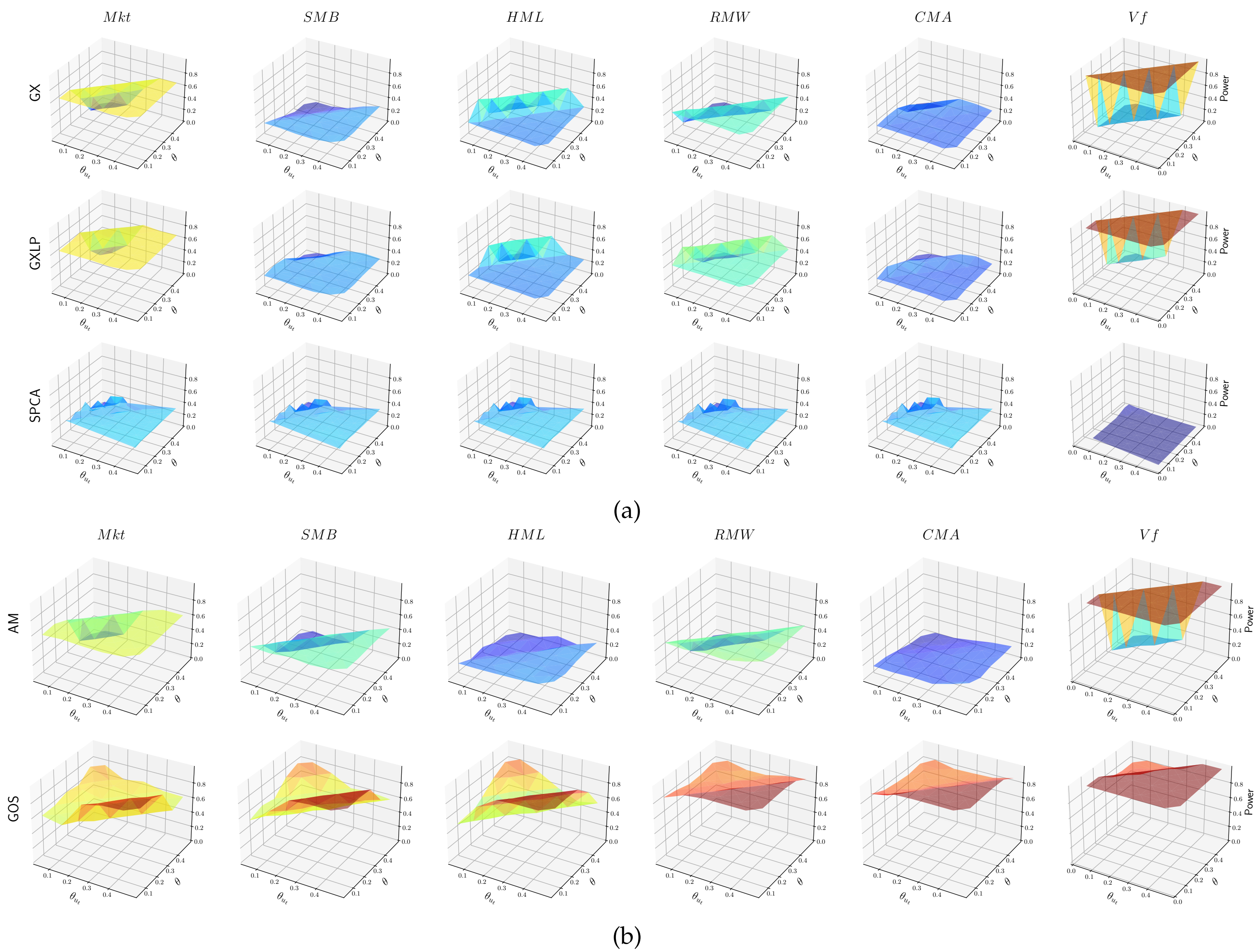


Figure 4: Power of the test when loadings are weak and pervasive. $N = 400$, $T = 400$. PCA-based estimators (panel a) and estimators related to observable factors (panel b)

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).

This scenario is related to the weak factor problem first acknowledged by Kan and Zhang (1999). 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.

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_{u_t}), as illustrated by the flat surface. In terms of magnitude, for the market factor, 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 the market 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 not reliable 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 remains 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 consistently 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.

Fig. 6 describes the power of the tests. One caveat is needed. If the test is not reliable in terms of size, then the power becomes an irrelevant feature to evaluate. For this reason, we do not discuss the power when testing under the null of correct risk premium shows poor results and analyse the power of the test for SMB, HML, RMW and CMA for the GX and GXLP estimators. Again, the results for the GXLP estimation are in line with the ones for the three step procedure of Giglio and Xiu (2021). Across simulations, we find that the power of the test hovers around 20% for CMA, while it is approximately 40% and 60%, respectively, for RMW and HML, across different specifications of the error and sparsity rates.

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 will render inference problematic not only for the sparse factor but also for the other priced risk factors.

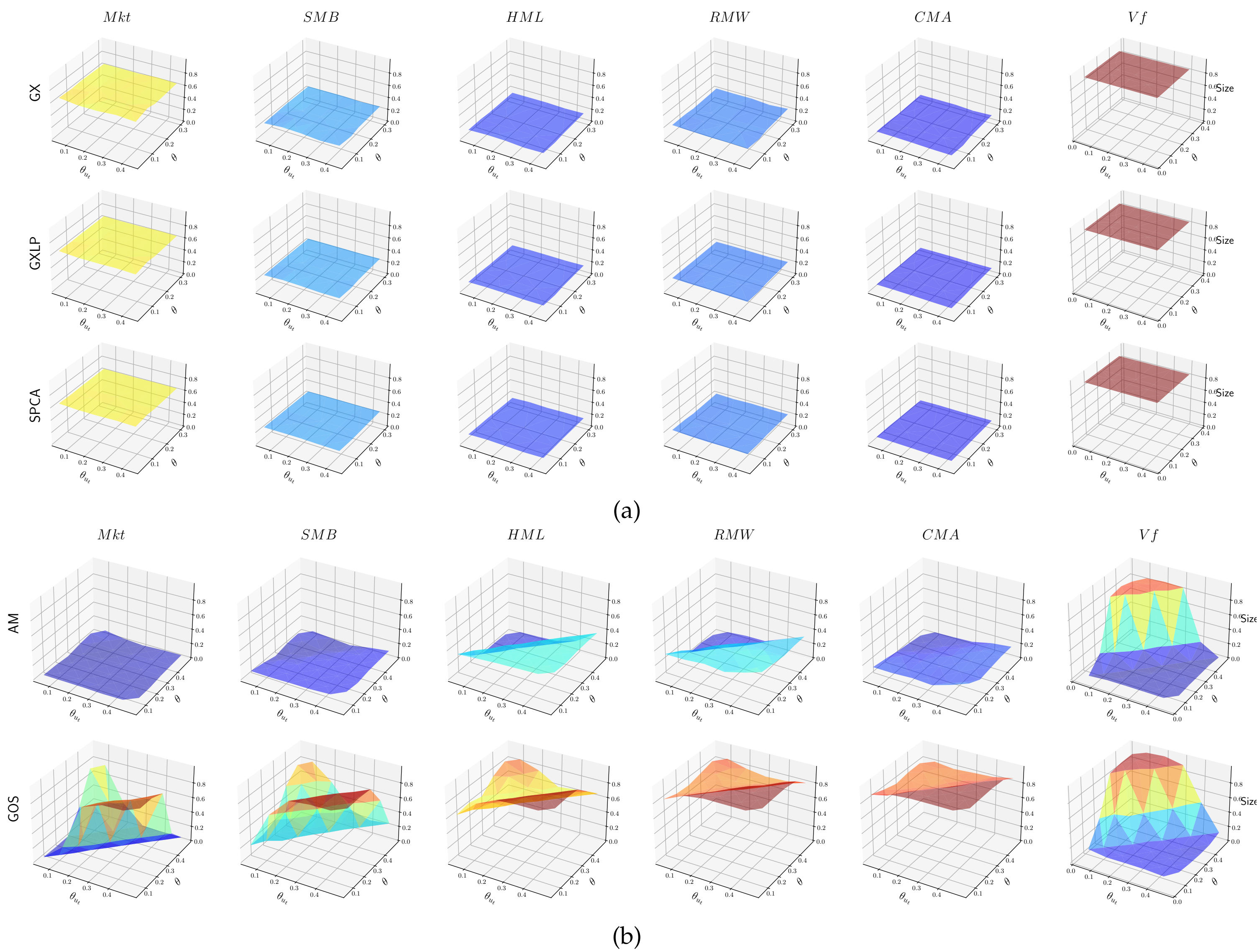
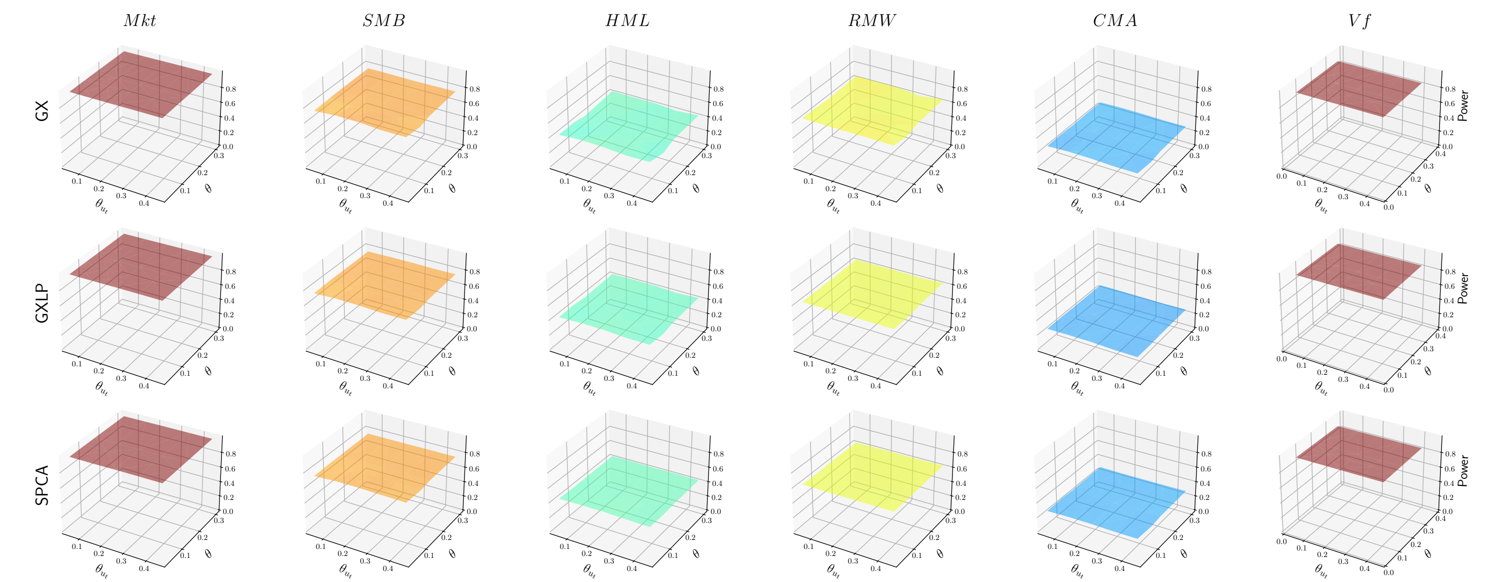
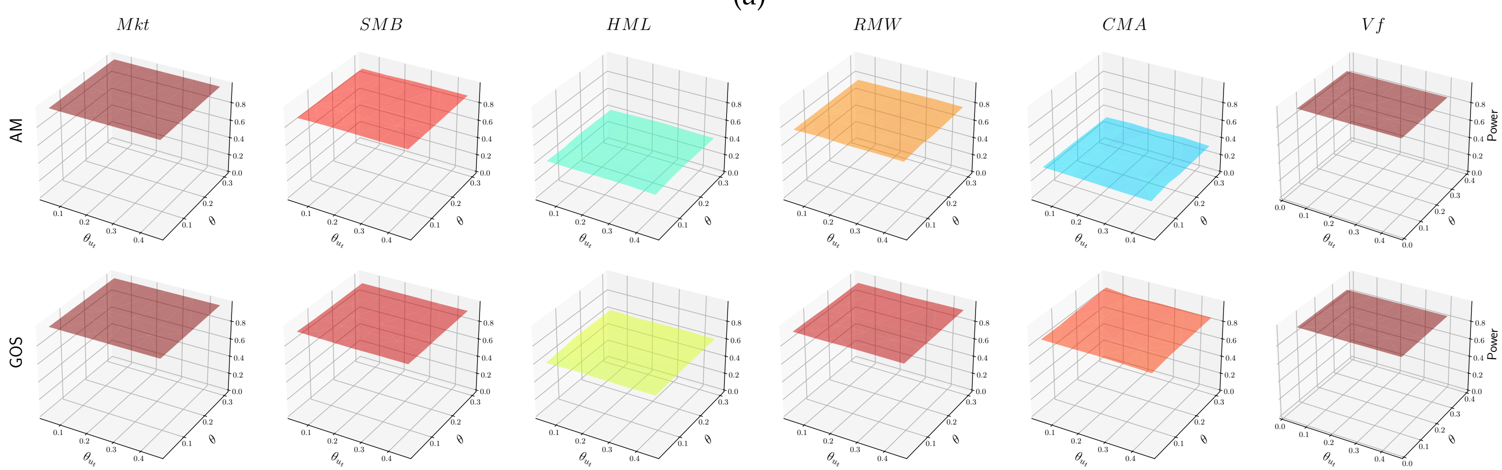


Figure 5: Size of the test when some of the assets have zero exposure to the Mkt factor. $N = 400$, $T = 400$. PCA-based estimators (panel a) and estimators related to observable factors (panel b)



(a)



(b)

Figure 6: Power of the test when some of the assets have zero exposure to the Mkt factor. $N = 400$, $T = 400$. PCA-based estimators (panel a) and estimators related to observable factors (panel b)

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, as we are focusing now in understanding the behaviour of t-test when factors' strength (i.e., the rate of the loadings) 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 in 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. Figs. 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.[24]

As for size, we note that the test performs poorly when θ_{u_t} is relatively low (i.e. the factor structure of the empirical errors is not extremely weak). From the charts, it seems that what drives the performance of the test is not the rate applied to the factor loading (i.e., 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.

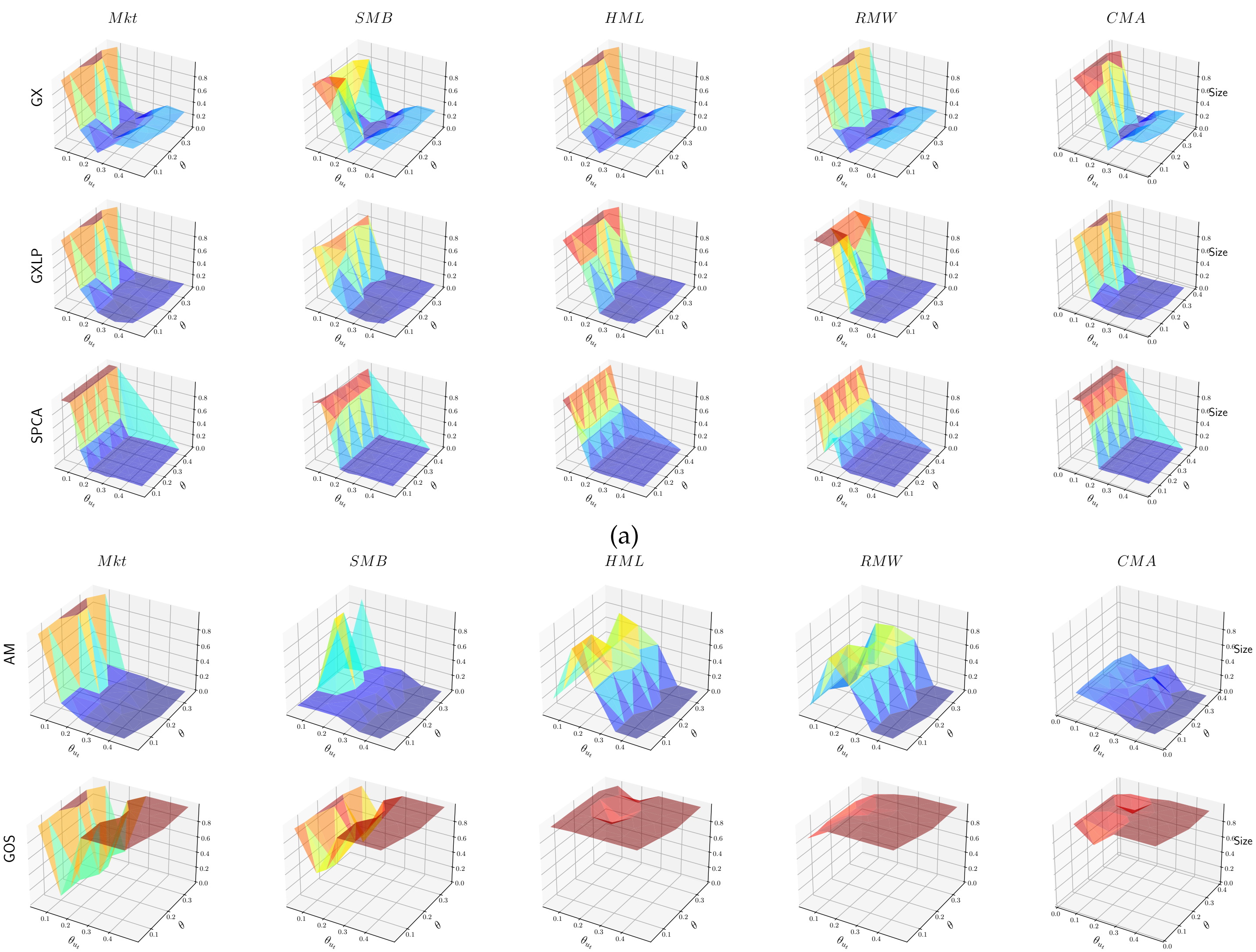
For the GX estimator, interestingly, we see a U-shape behaviour for the size of the test across different factors when the θ_{u_t} moves from 0 to 0.5. Only when the rate on error is moderate (i.e., $\theta_{u_t} \approx 0.2$), the test has a good size (the test is slightly undersized, with empirical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

[24]The remaining results for diverse percentile rankings are available upon request.

crease it, 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, 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.

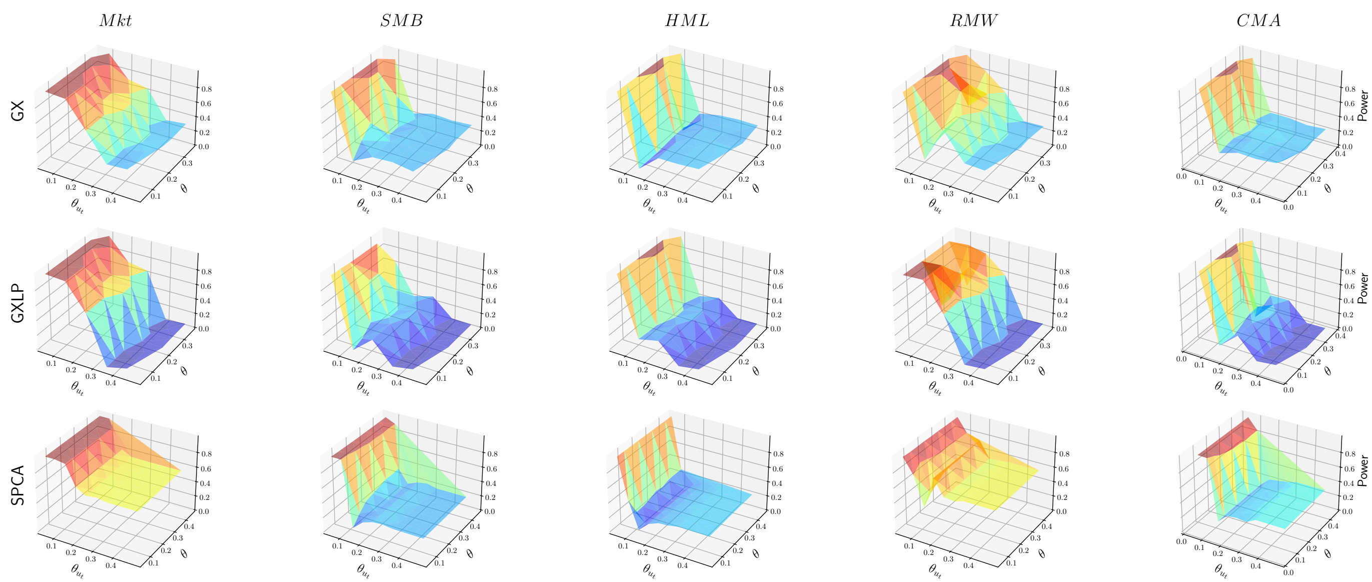
The analysis holds when we look at the power property of the test. In general, across estimators and factors, it seems that the size of the test is of the same order of magnitude as the power (this is true in particular for the GOS estimator). In conclusions, inference is not generally reliable because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



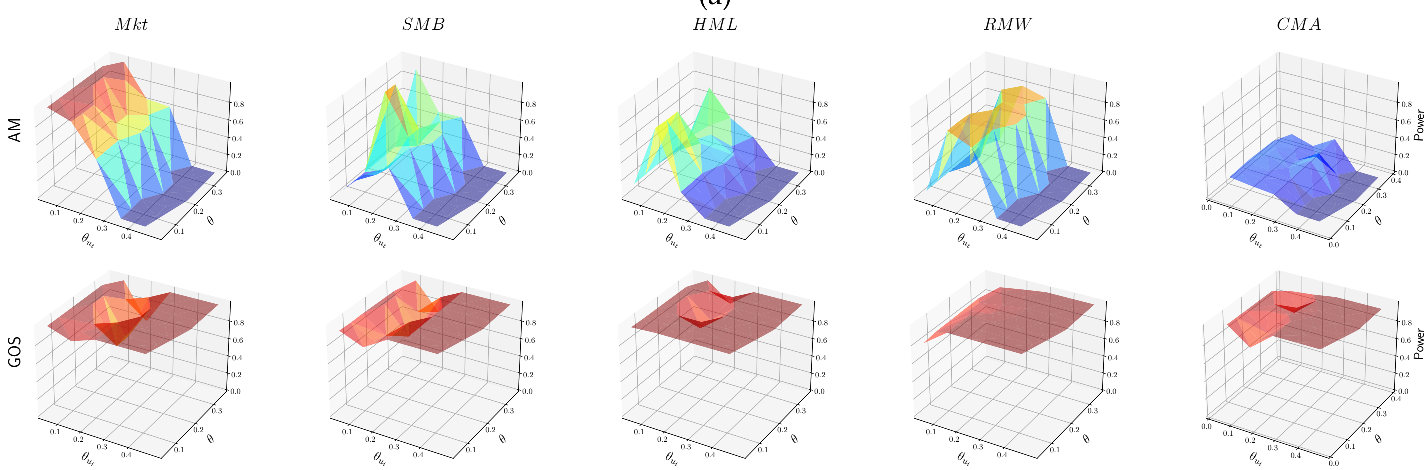
(a)

(b)

Figure 7: Size of the test when the lowest volatile assets have a strong exposure to SMB. $N = 400$, $T = 400$. PCA-based estimators (panel a) and estimators related to observable factors (panel b)



(a)



(b)

Figure 8: Power of the test when the lowest volatile assets have a strong exposure to SMB. N = 400, $T = 400$. PCA-based estimators (panel a) and estimators related to observable factors (panel b)

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 of Giglio and Xiu (2021).

Since the methodology estimates the risk premia on observable factors, $\hat{\gamma}_g$, as a byproduct of two components (i.e., the latent factor risk premia and the projection coefficient), the goal is to have a clearer picture on whether identification issues could arise from the matrix of beta exposures to latent factors tending to be column rank deficient. 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, $\hat{\gamma}$ with respect to Eq. (1). The following example shows why this is issue might be actual. Suppose one considers the full sample of Giglio and Xiu (2021) asset returns, and estimates the $\widehat{\beta}$ of model Eq. (1) with respect to $\hat{p} = 12$ (i.e., extracting 12 PCs). The left chart of Fig.9 displays the frequencies on how many times the $\{|\hat{\beta}_i|\}_{i=1,..,N}$ are of order of 10^{-4} or below. While for the loading associated to the 1^{st} PC, the frequency is extremely low, for the loading associated to the 7^{th} PC, the frequency is nearly 30% (up to almost 50% for the 12^{th} PC).

From a more formal angle, we apply the sequential test of Onatski (2009) to their panel of assets.[25] The right chart of Fig. 9 indicates that the number of latent factors in Giglio and Xiu (2021)'s panel 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.

[25] Consider an estimate of the number of latent factors $\hat{p}$, and (a potential candidate for) the true value p. The null hypothesis is that the estimated number of factor is equal to the true one $\hat{p} = p$. The alternative hypothesis is that true number of factors is between $\hat{p}$ and $\hat{p} + K$, with K being a positive integer that is chosen by practitioners.

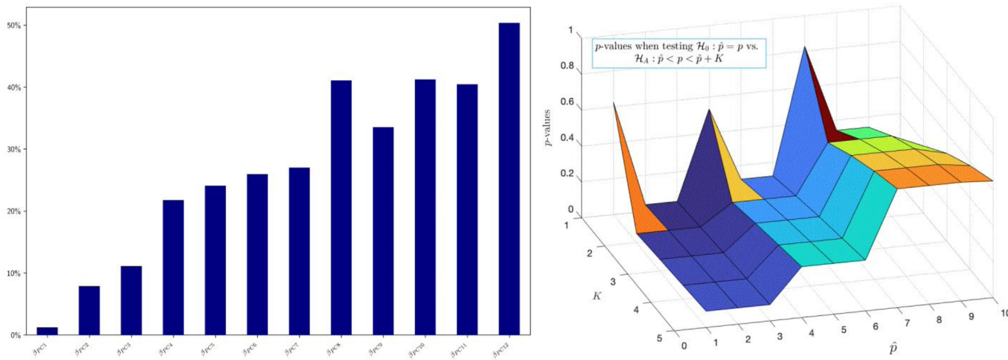


Figure 9: On the left, the frequency of $|\beta_i| \lesssim 10^{-4}$. On the right, the sequential testing of Onatski (2009) (on z-axis, its p-value)

What happens to the asymptotics of $\hat{\gamma}_g$ when the exposures are so weak? In particular, when over-extracting, 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. However, the study of this issue and its distortions in the risk premia estimates translates, in our context, into generating artificial returns following the procedure outlined in Section 3. For the sake of simplicity, we consider 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 large $\breve{T}$ framework because spuriousness is more severe when the number of time series observations tends to be large (see 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 2^{nd} to 7^{th} PCs should tend to be increasingly weakly correlated (if not nearly spurious) with test assets returns, and therefore we suspect the risk premia estimates to suffer distortions similar to those that arise in the spurious factor case. In Figs.10 and 11, each charts 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.

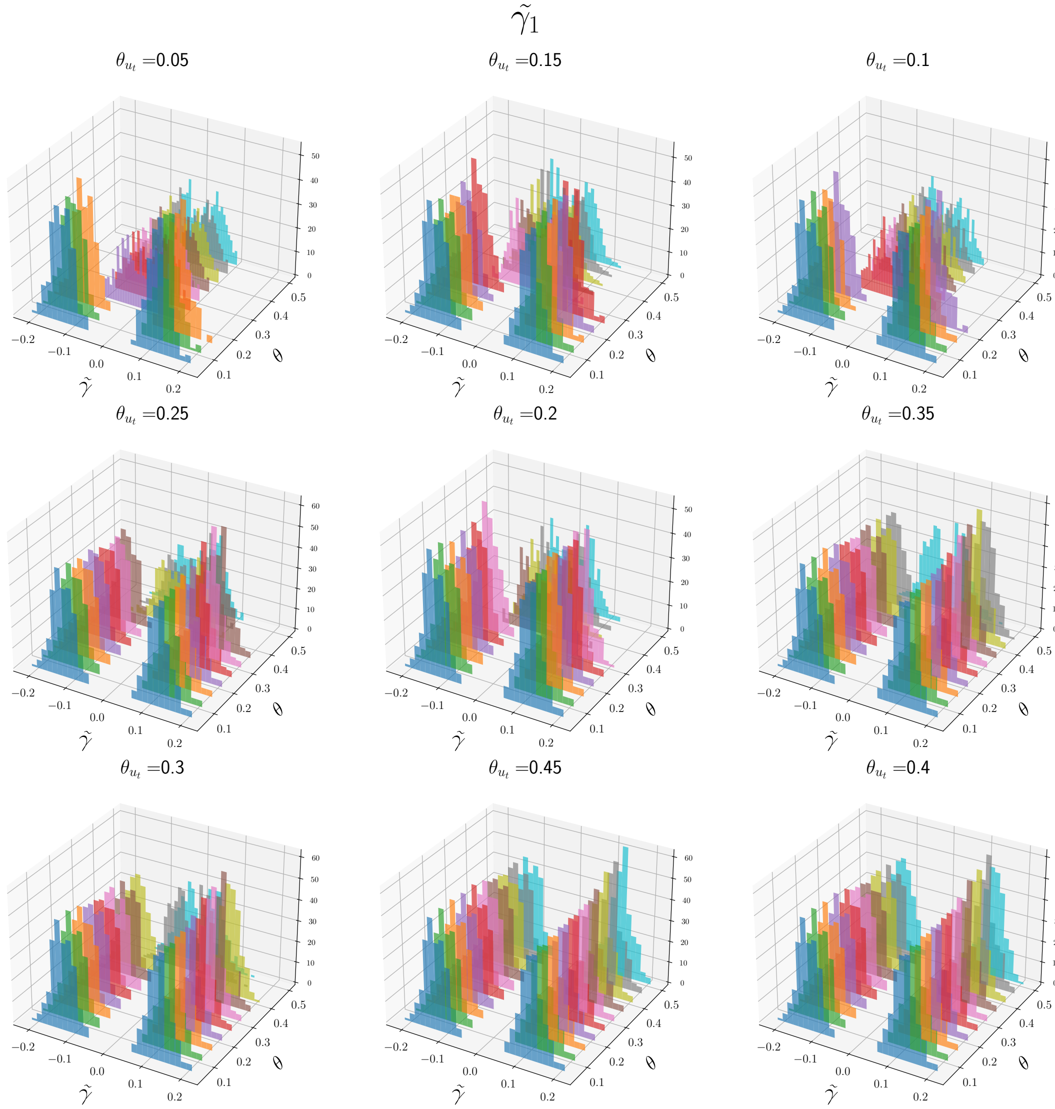


Figure 10: Emp. distribution of the risk premium estimate $\tilde{\gamma}$ on the first principal component in function of θ and θ_{u_t}

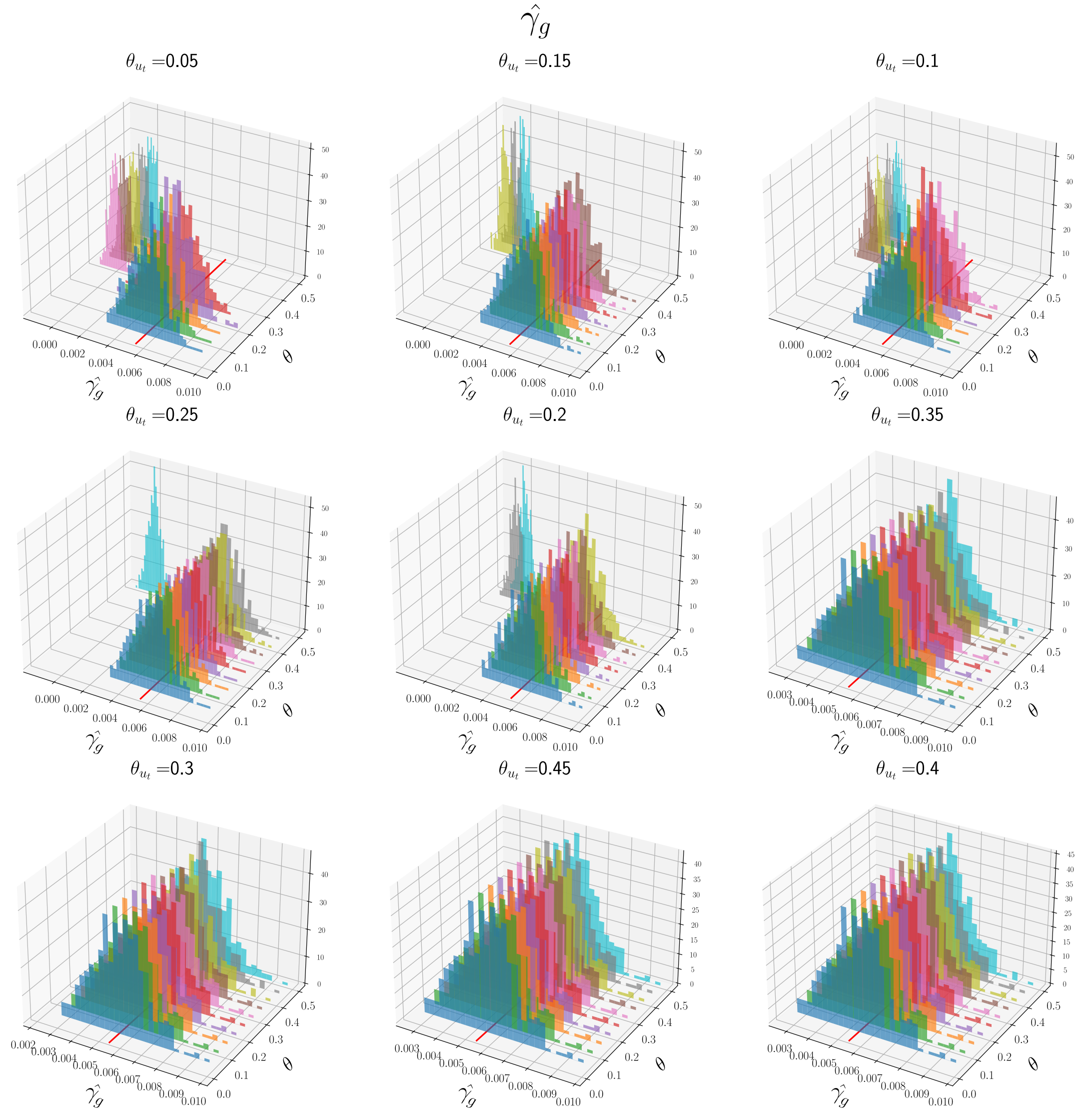


Figure 11: Emp. distribution of the market risk premium $\hat{\gamma}_g$ in function of θ and θ_{u_t}. The true market risk premium is 0.55%.

We provide the empirical distribution of all the γ's and η's (Eq. 1-3) corresponding to the different PCs in Appendix C.

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 C. 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 in 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 aim at maintaining the cross-sectional structure of the artificial data as close as possible to the one of the real data. We consider various scenarios that can lead to identification problems. Specifically, we analyze situations in which the test asset returns exhibit a strong and pervasive factor structure, cases where the factor structure is weak and pervasive, 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 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, in particular Bakalli et al. (2023) or other PCA-based methods 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 important investigations for future research.

References

ANATOLYEV, S. AND A. MIKUSHEVA (2021): "Factor models with many assets: strong factors, weak factors, and the two-pass procedure," *Journal of Econometrics*.

ANTOINE, B., K. PROULX, AND E. RENAULT (2020): "Pseudo-true SDFs in conditional asset pricing models," *Journal of Financial Econometrics*, 18, 656–714.

BAI, J. (2003): "Inferential theory for factor models of large dimensions," *Econometrica*, 71, 135–171.

BAI, J. AND S. NG (2021): "Approximate Factor Models with Weaker Loadings," *arXiv preprint arXiv:2109.03773*.

BAKALLI, G., S. GUERRIER, AND O. SCAILLET (2023): "A penalized two-pass regression to predict stock returns with time-varying risk premia," *Journal of Econometrics*.

BRYZGALOVA, S., V. DEMIGUEL, S. LI, AND M. PELGER (2023): "Asset-Pricing Factors with Economic Targets," *Available at SSRN 4344837*.

CAI, T. T., Z. MA, AND Y. WU (2013): "Sparse PCA: Optimal rates and adaptive estimation," *The Annals of Statistics*, 41, 3074–3110.

CHOI, J. AND X. YANG (2022): "Asymptotic properties of correlation-based principal component analysis," *Journal of Econometrics*, 229, 1–18.

CONNOR, G. AND R. A. KORAJCZYK (2022): "Semi-strong factors in asset returns," *Michael J. Brennan Irish Finance Working Paper Series Research Paper*.

FAMA, E. F. AND K. R. FRENCH (2015): "A five-factor asset pricing model," *Journal of financial economics*, 116, 1–22.

FAMA, E. F. AND J. D. MACBETH (1973): "Risk, return, and equilibrium: Empirical tests," *Journal of political economy*, 81, 607–636.

FREYALDENHOVEN, S. (2021): "Factor models with local factors—Determining the number of relevant factors," *Journal of Econometrics*.

GAGLIARDINI, P., E. OSSOLA, AND O. SCAILLET (2016): "Time-varying risk premium in large cross-sectional equity data sets," *Econometrica*, 84, 985–1046.

GIGLIO, S. AND D. XIU (2021): "Asset pricing with omitted factors," *Journal of Political Economy*, 129, 1947–1990.

GIGLIO, S., D. XIU, AND D. ZHANG (2022): "Test assets and weak factors," Tech. rep., National Bureau of Economic Research.

GOSPODINOV, N., R. KAN, AND C. ROBOTTI (2014): "Misspecification-robust inference in linear asset-pricing models with irrelevant risk factors," *The Review of Financial Studies*, 27, 2139–2170.

——— (2019): "Too good to be true? Fallacies in evaluating risk factor models," *Journal of Financial Economics*, 132, 451–471.

HANSEN, L. P. AND R. JAGANNATHAN (1997): "Assessing specification errors in stochastic discount factor models," *The Journal of Finance*, 52, 557–590.

HARVEY, C. R., Y. LIU, AND H. ZHU (2016): "... and the cross-section of expected returns," *The Review of Financial Studies*, 29, 5–68.

KAN, R., C. ROBOTTI, AND J. SHANKEN (2013): "Pricing model performance and the two-pass cross-sectional regression methodology," *The Journal of Finance*, 68, 2617–2649.

KAN, R. AND C. ZHANG (1999): "Two-pass tests of asset pricing models with useless factors," *the Journal of Finance*, 54, 203–235.

KLEIBERGEN, F. (2009): "Tests of risk premia in linear factor models," *Journal of econometrics*, 149, 149–173.

KLEIBERGEN, F. AND Z. ZHAN (2015): "Unexplained factors and their effects on second pass R-squared's," *Journal of Econometrics*, 189, 101–116.

LETTAU, M. AND M. PELGER (2020): "Estimating latent asset-pricing factors," *Journal of Econometrics*, 218, 1–31.

MERTON, R. C. (1973): "An intertemporal capital asset pricing model," *Econometrica: Journal of the Econometric Society*, 867–887.

ONATSKI, A. (2009): "Testing hypotheses about the number of factors in large factor models," *Econometrica*, 77, 1447–1479.

——— (2012): "Asymptotics of the principal components estimator of large factor models with weakly influential factors," *Journal of Econometrics*, 168, 244–258.

ROSS, S. A. (1976): "The arbitrage theory of capital asset pricing," *Journal of Economic Theory*, 13, 341–360.

SHARPE, W. F. (1964): "Capital asset prices: A theory of market equilibrium under conditions of risk," *The journal of finance*, 19, 425–442.

UEMATSU, Y. AND T. YAMAGATA (2022a): "Estimation of sparsity-induced weak factor models," *Journal of Business & Economic Statistics*, 41, 213–227.

——— (2022b): "Inference in sparsity-induced weak factor models," *Journal of Business & Economic Statistics*, 41, 126–139.

ZHANG, A. R., T. T. CAI, AND Y. WU (2022): "Heteroskedastic PCA: Algorithm, optimality, and applications," *The Annals of Statistics*, 50, 53–80.

A Appendix A

A.1 Giglio and Xiu (2021)

Let us denote the time-demeaned excess returns matrix, $\bar{R} = R - \bar{r}\iota_T'$, 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, $\bar{R} = R - \bar{r}\iota_T'$, 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:

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

   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 $\widehat{\beta}$. We have the following output:

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

   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 $\widehat{V}$. We have the following output:

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

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

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

A.2 Giglio et al. (2022)

Let us consider the time-demeaned excess returns matrix, $\bar{R} = R - \bar{r}\iota_T'$, 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 $k = 1, 2...$, 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 $\mathcal{L}_\infty$ 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: $\{\widehat{V}_{(k)}, \widehat{\beta}_{(k)}, \widehat{\eta}_{(k)}, \widehat{\gamma}_{(k)}\}$

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

   (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 $\hat{\gamma}^{SPCA} = \sum_{k=1}^{\hat{p}} \widehat{\eta}_{(k)} \widehat{\gamma}_{(k)}$

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, $\bar{R} = R - \bar{r}\iota'_T$. 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:

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

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

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

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 $\widehat{\beta}$. We have the following output:

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

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 $\widehat{V}$. We have the following output:

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

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

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

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 $\tilde{F}_t = F_t - T^{-1}\sum_s^T F_s$.

Let us consider the division the set of time indexes into 4 equal non-intersecting subsets T_j, $j = 1, ..., 4$.

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

$$\hat{\beta}_i^{(j)} = \left(\sum_{t \in T_j} \tilde{F}_t^{(j)} \tilde{F}_t^{(j)\prime}\right) \sum_{t \in T_j} \tilde{F}_t^{(j)} r_{it}$$

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 $z_i^{(1)} = (\hat{\beta}_i^{(3)}, (\hat{\beta}_i^{(3)} - \hat{\beta}_i^{(3)})')'$, 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 $z_i^{(2)} = (\hat{\beta}_i^{(4)}, (\hat{\beta}_i^{(4)} - \hat{\beta}_i^{(1)})')'$; denote the estimate $\hat{\lambda}^{(2)}$, etc.

4. Obtain the four-split estimate as:

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

B Appendix: Anatolyev and Mikusheva (2021)'s calibrations

In this calibration strategy, we generate returns following Eq.(6)-(7).

In details, 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 $\breve{\lambda} = 1/4\sum_i^4 \hat{\lambda}_i^{AM}$, $\breve{\beta} = 1/4\sum_i^4 \hat{\beta}_i^{AM}$.

Related to the structure error terms (Eq.(7)), 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 constitue 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 $\breve{N}$ of observed returns $r_t^{(\breve{N})}$ 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:

$$\begin{pmatrix} F_t^{\diamond} \\ v_t^{\diamond} \\ e_t^{\diamond} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \bar{F} \\ 0_{(2+1)\times 1} \end{pmatrix}, \begin{pmatrix} \widehat{\Sigma}^F & 0 & 0 \\ 0 & \widehat{\Sigma}^v & 0 \\ 0 & 0 & \widehat{\sigma}^e \end{pmatrix} \right), \tag{11}$$

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.(6)-(7), the realizations of the i of the $\breve{N}$ simulated returns $r_{i,t}^{\diamond}$ is:

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

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.

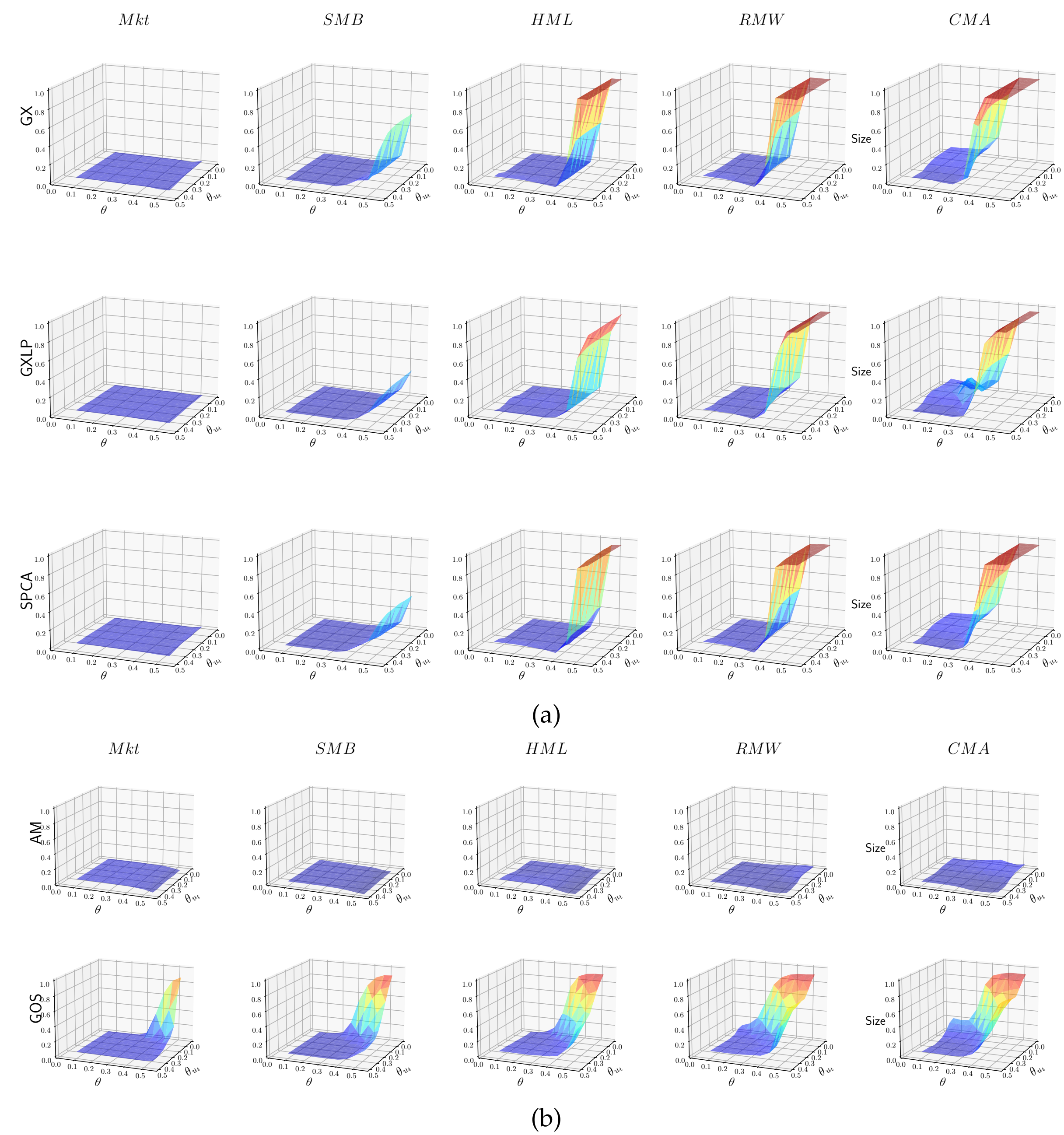


Figure 12: Size of the test when loadings are weak/semi-strong and pervasive. $N = 400, T = 400$. PCA-based estimators (panel a) and estimators related to observable factors (panel b).

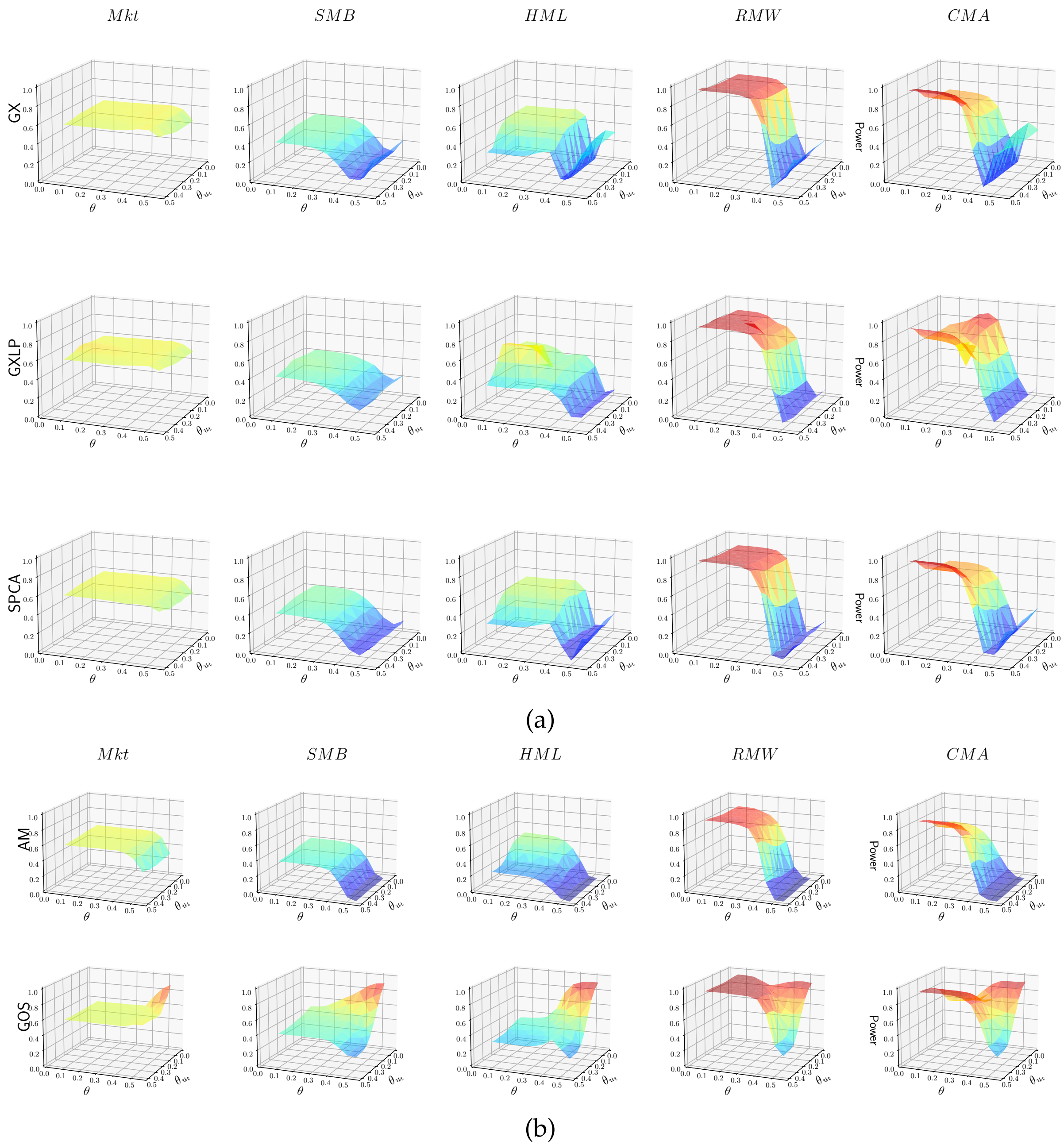


Figure 13: Power of the test when loadings are weak/semi-strong and pervasive. $N = 400, T = 400$. PCA-based estimators (panel a) and estimators related to observable factors (panel b).

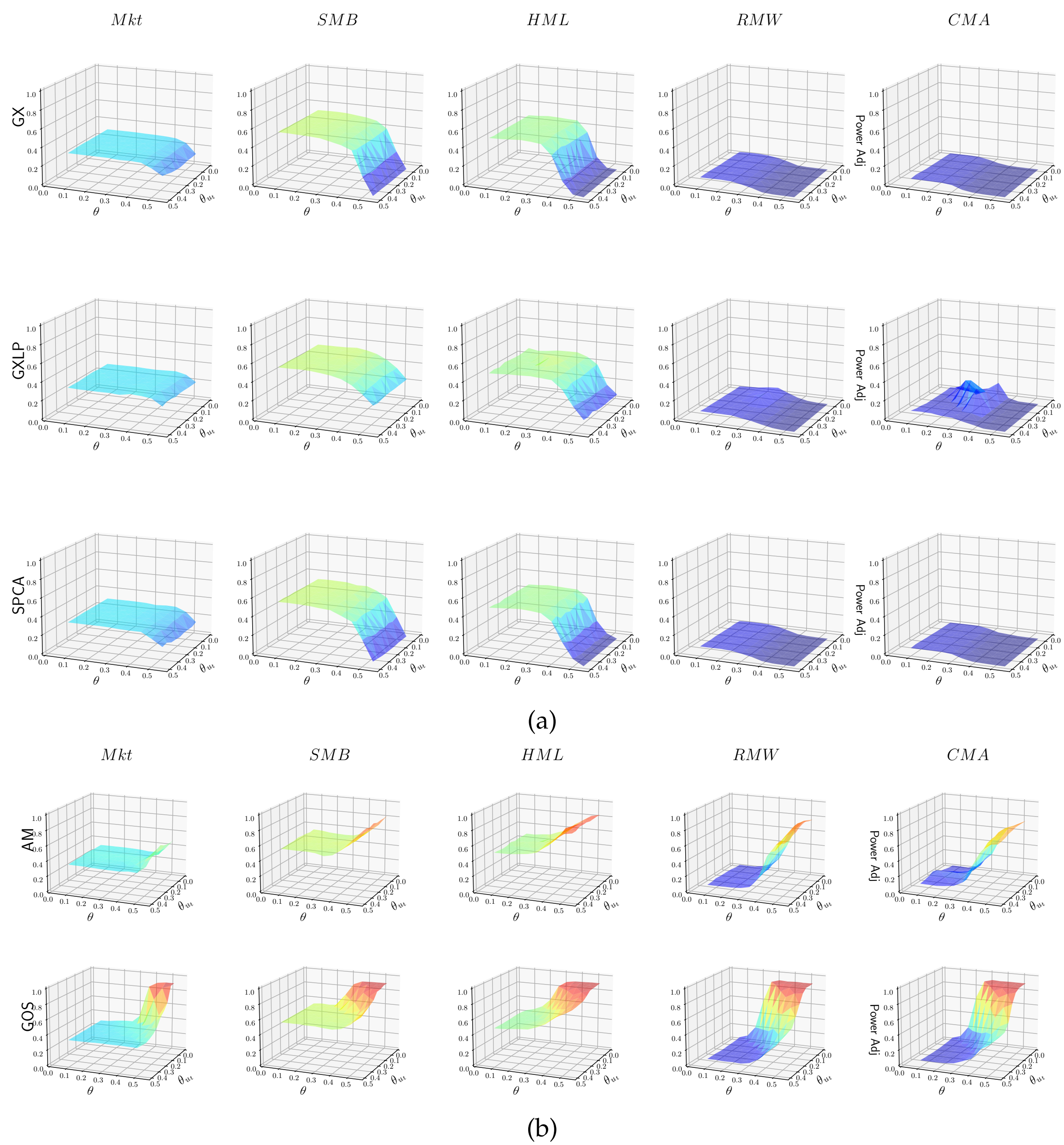
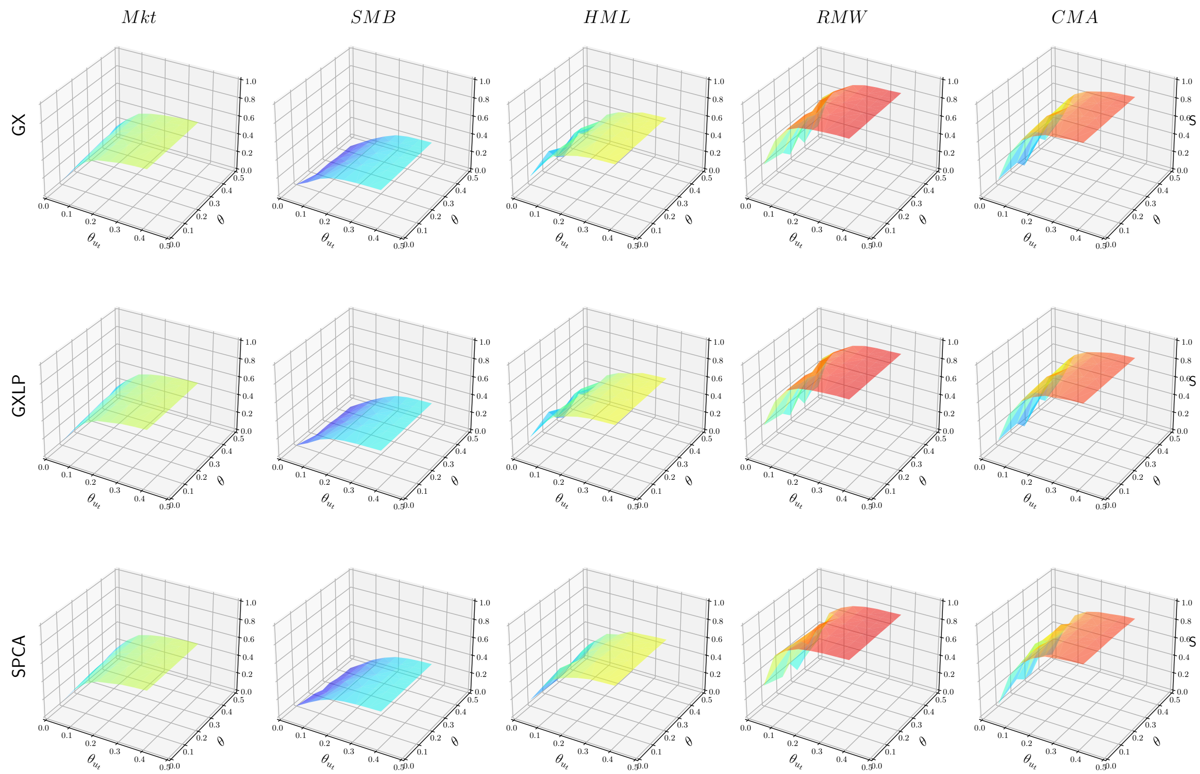
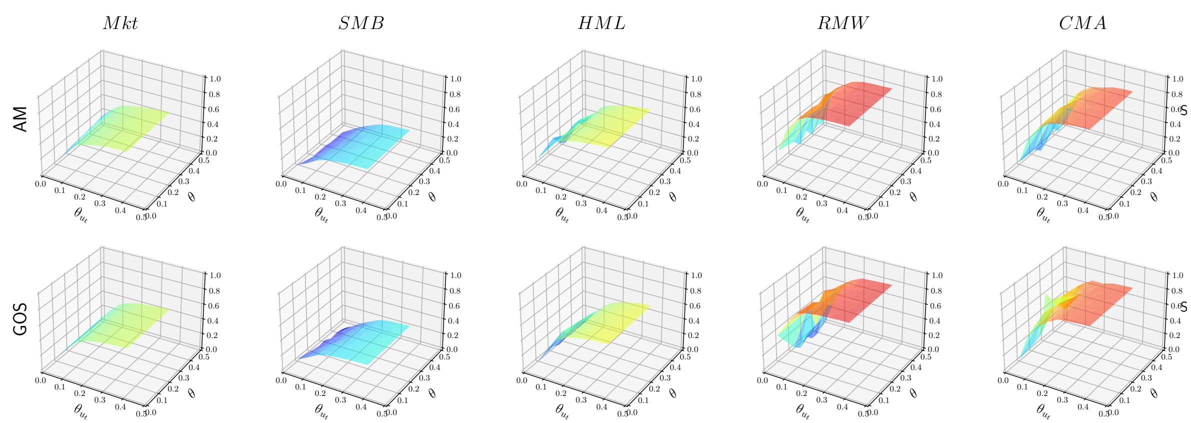


Figure 14: Adjusted Power of the test when loadings are weak/semi-strong and pervasive. $N = 400, T = 400$. PCA-based estimators (panel a) and estimators related to observable factors (panel b).



(a)



(b)

Figure 15: Size of the test when loadings are sparse. $N = 400, T = 400$. PCA-based estimators (panel a) and estimators related to observable factors (panel b).

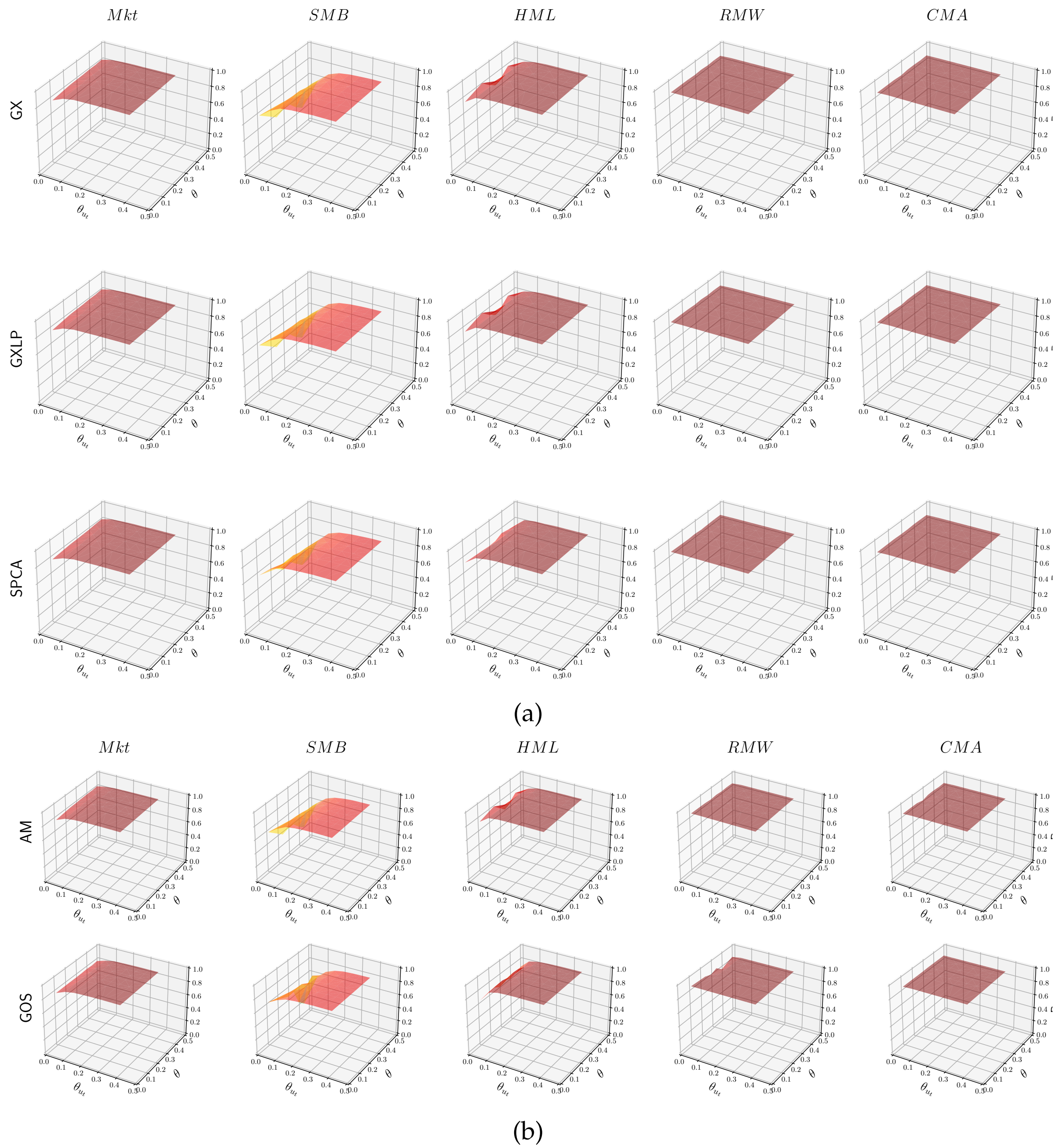


Figure 16: Power of the test when loadings are sparse. $N = 400, T = 400$. PCA-based estimators (panel a) and estimators related to observable factors (panel b).

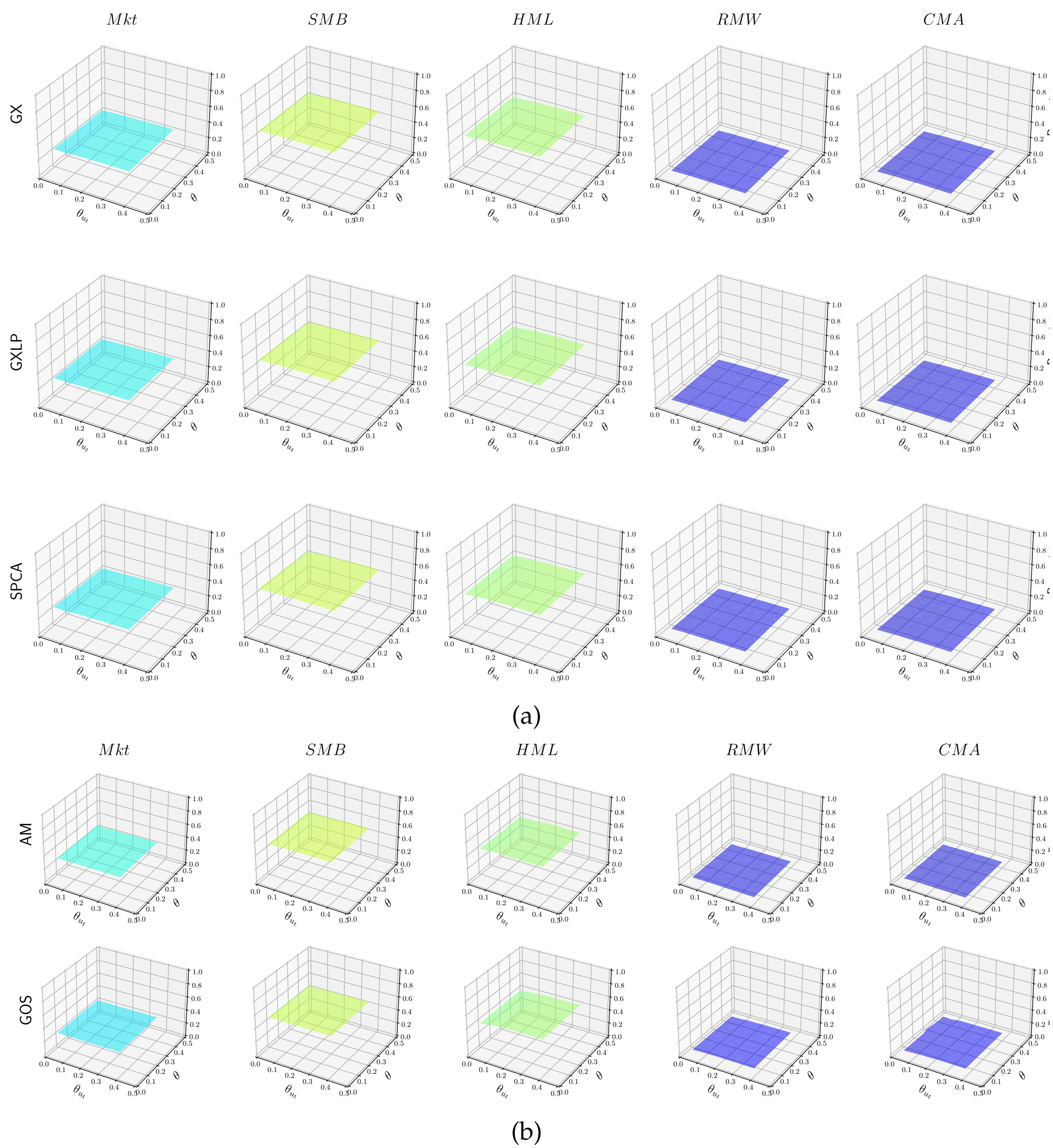


Figure 17: Adjusted Power of the test when loadings are sparse. $N = 400, T = 400$. PCA-based estimators (panel a) and estimators related to observable factors (panel b).

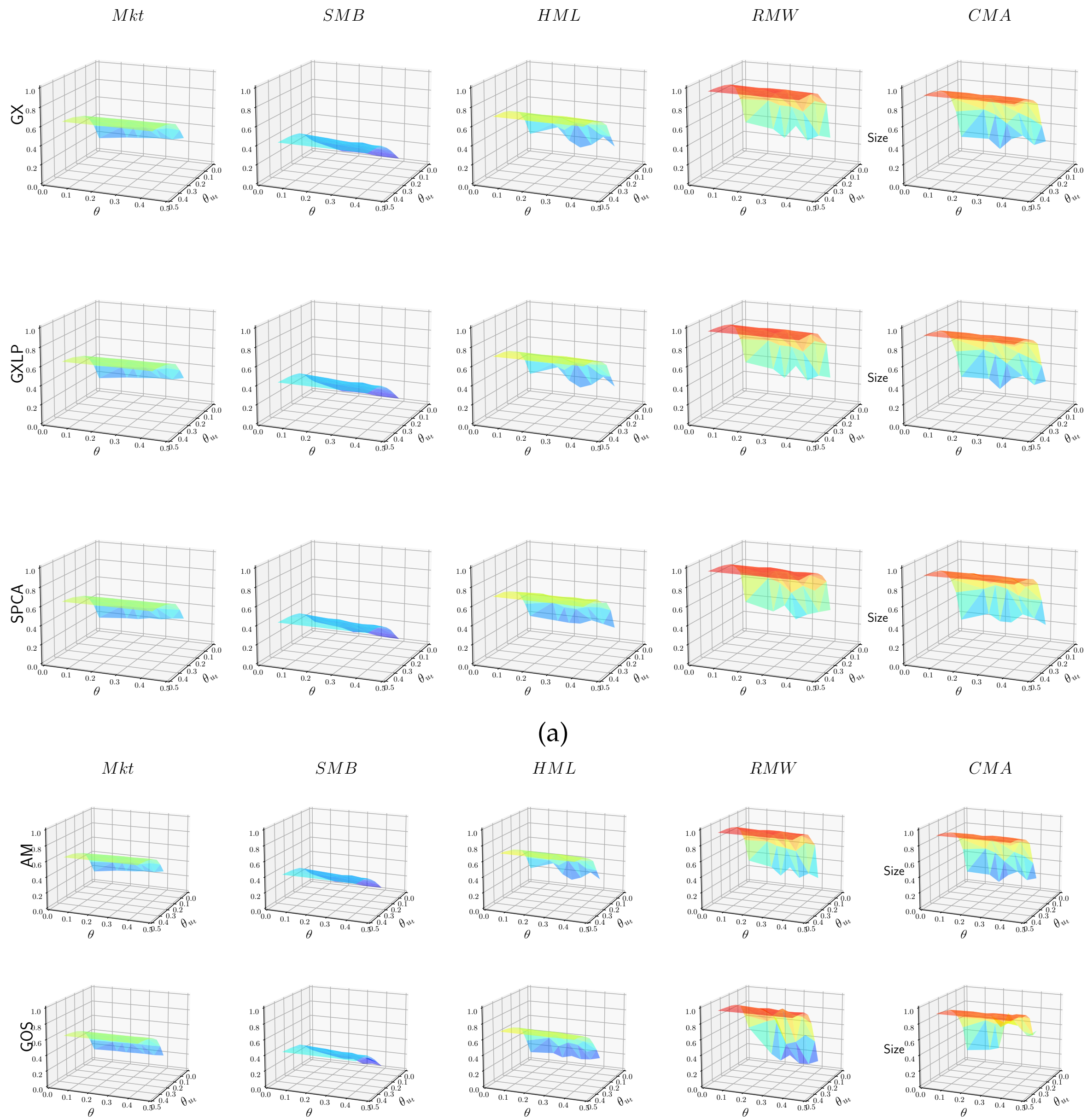


Figure 18: Size of the test when loadings are sparse. $N = 400, T = 400$. PCA-based estimators (panel a) and estimators related to observable factors (panel b).

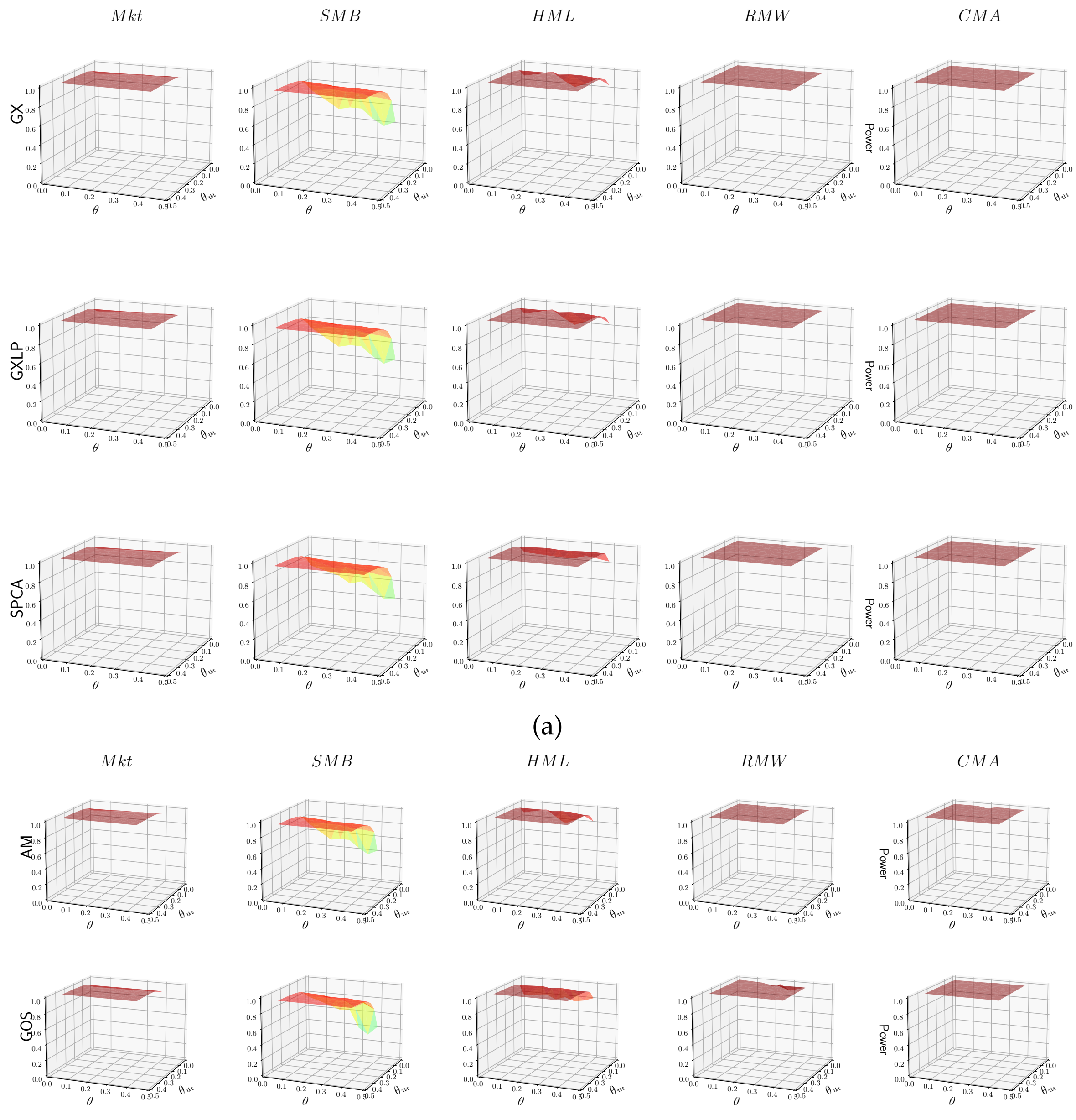


Figure 19: Power of the test when loadings are sparse. $N = 400, T = 400$. PCA-based estimators (panel a) and estimators related to observable factors (panel b).

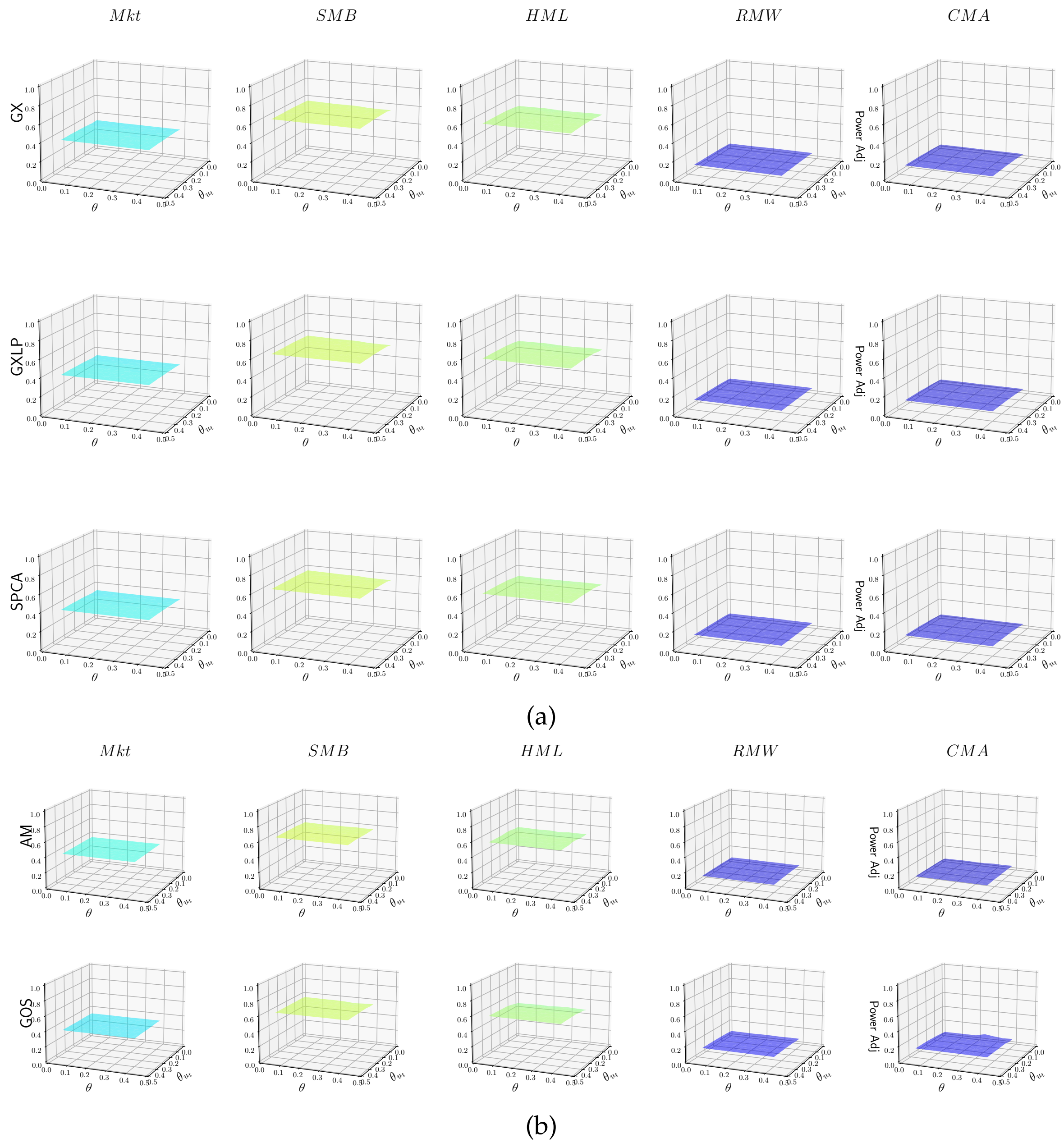
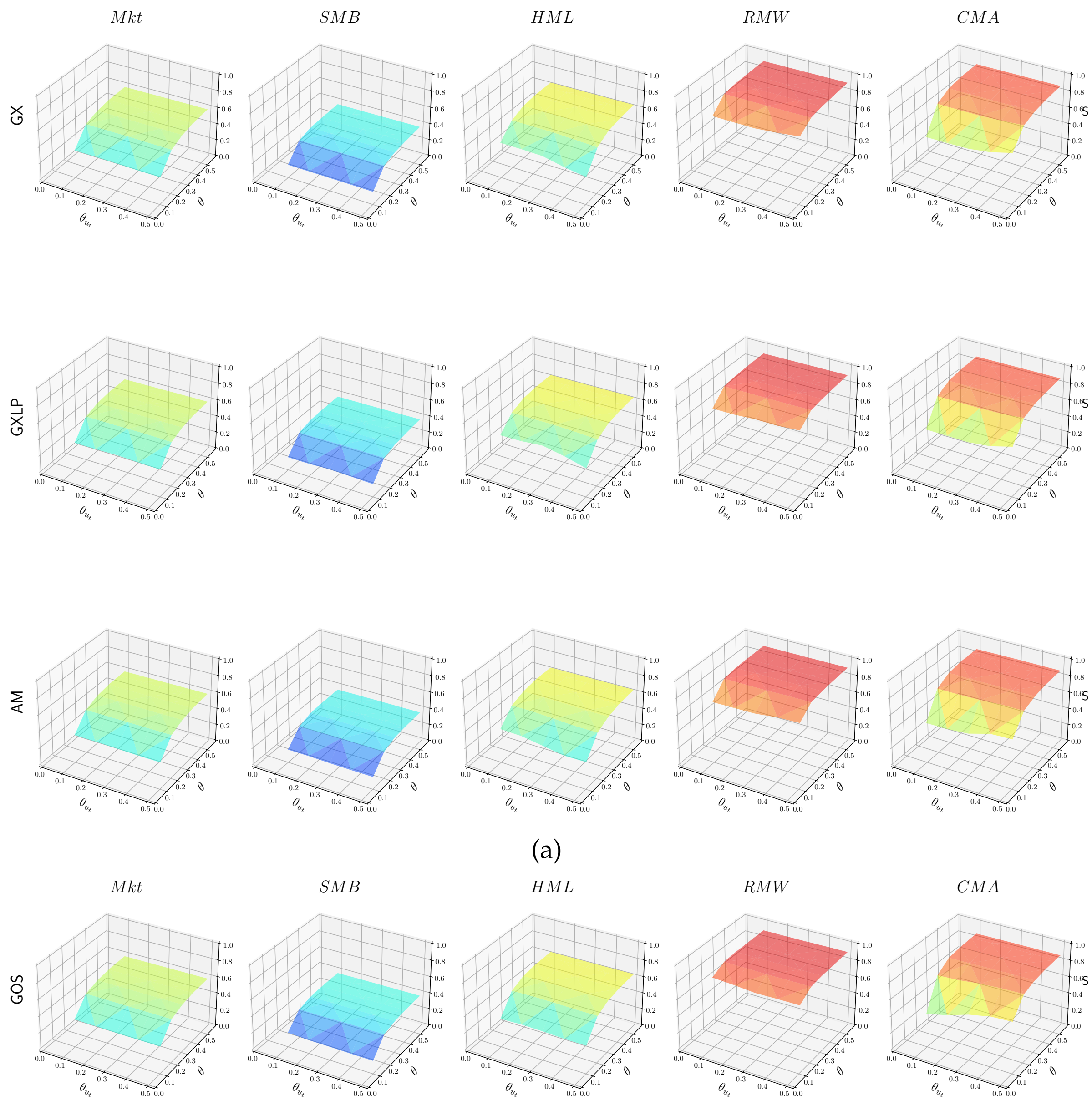


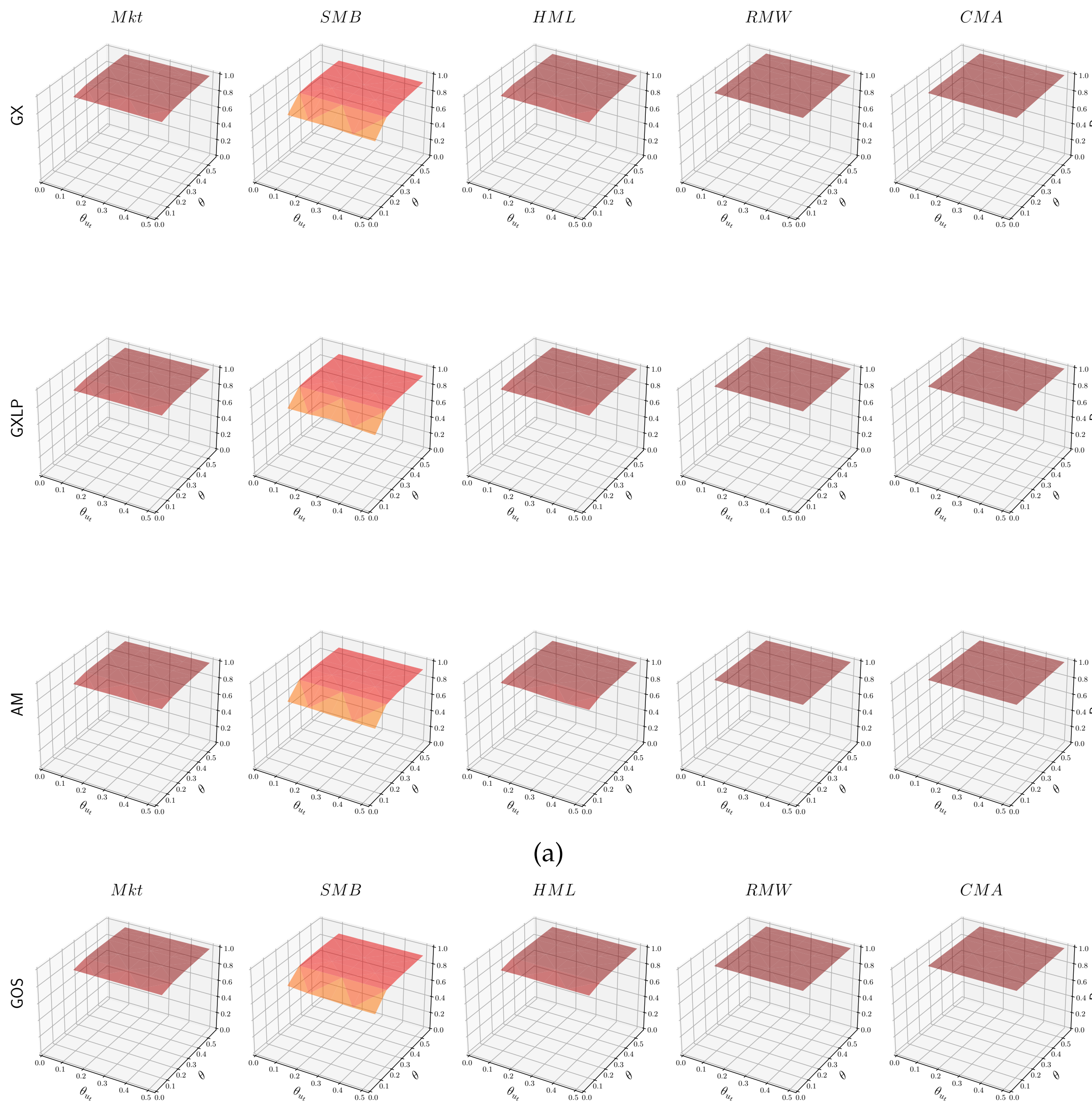
Figure 20: Adjusted Power of the test when loadings are local and strong/semi-strong. $N = 400, T = 400$. PCA-based estimators (panel a) and estimators related to observable factors (panel b).



(a)

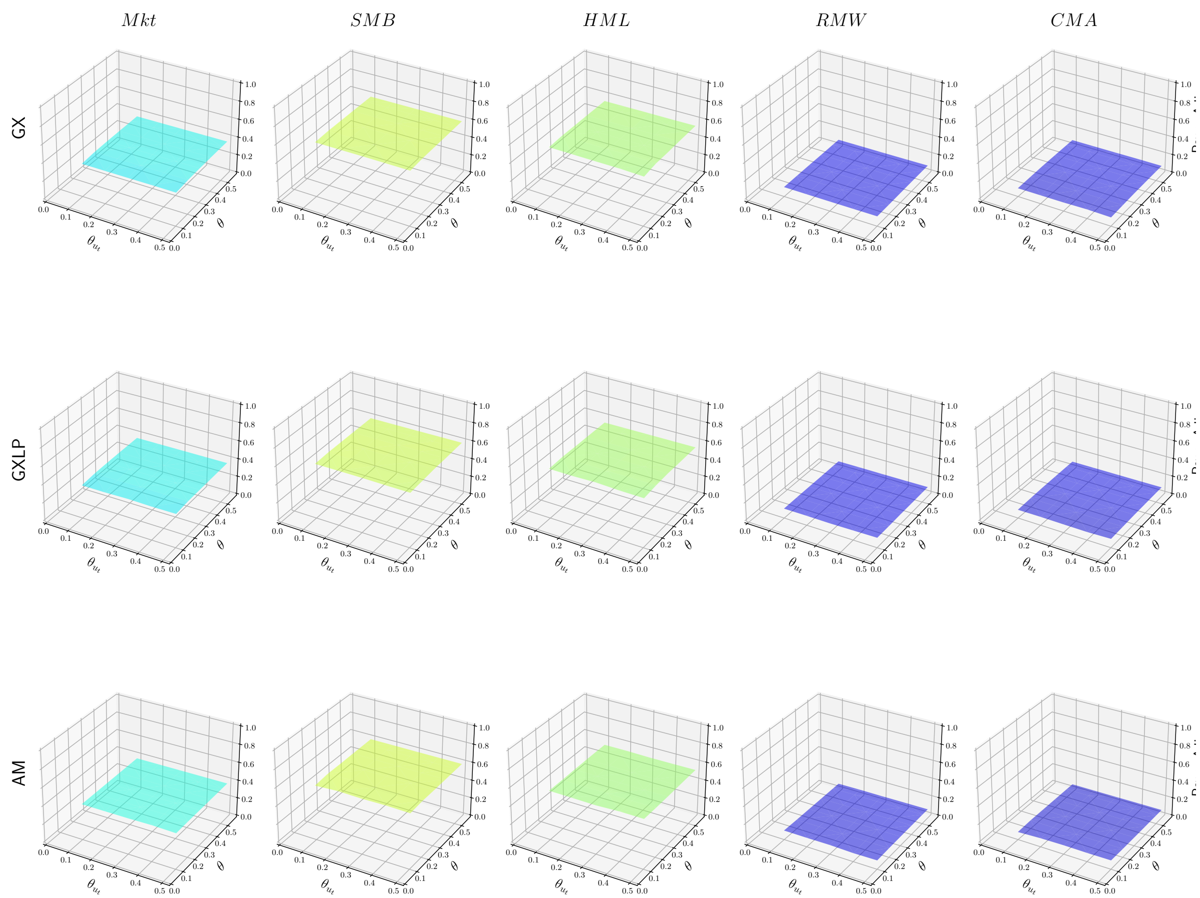
(b)

Figure 21: Size of the test when loadings are local and strong/semi-strong. $N = 400, T = 400$. PCA-based estimators and Anatolyev and Mikusheva (2021)'s (panel a) and Gagliardini et al. (2016) estimator (panel b).

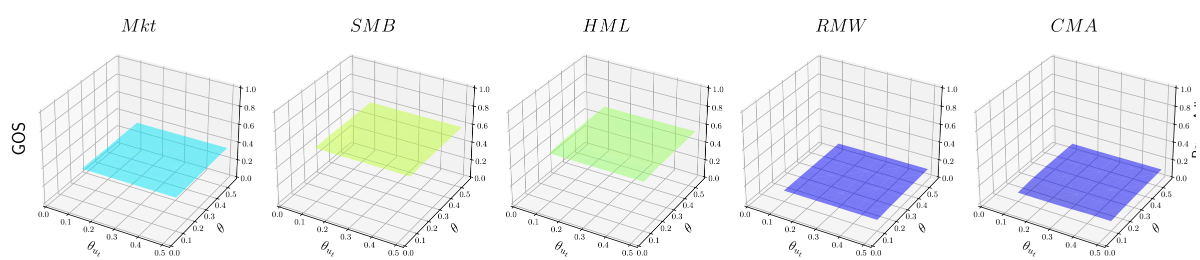


(b)

Figure 22: Power of the test when loadings are local and strong/semi-strong. $N = 400, T = 400$. PCA-based estimators and Anatolyev and Mikusheva (2021)'s (panel a) and Gagliardini et al. (2016) estimator (panel b).

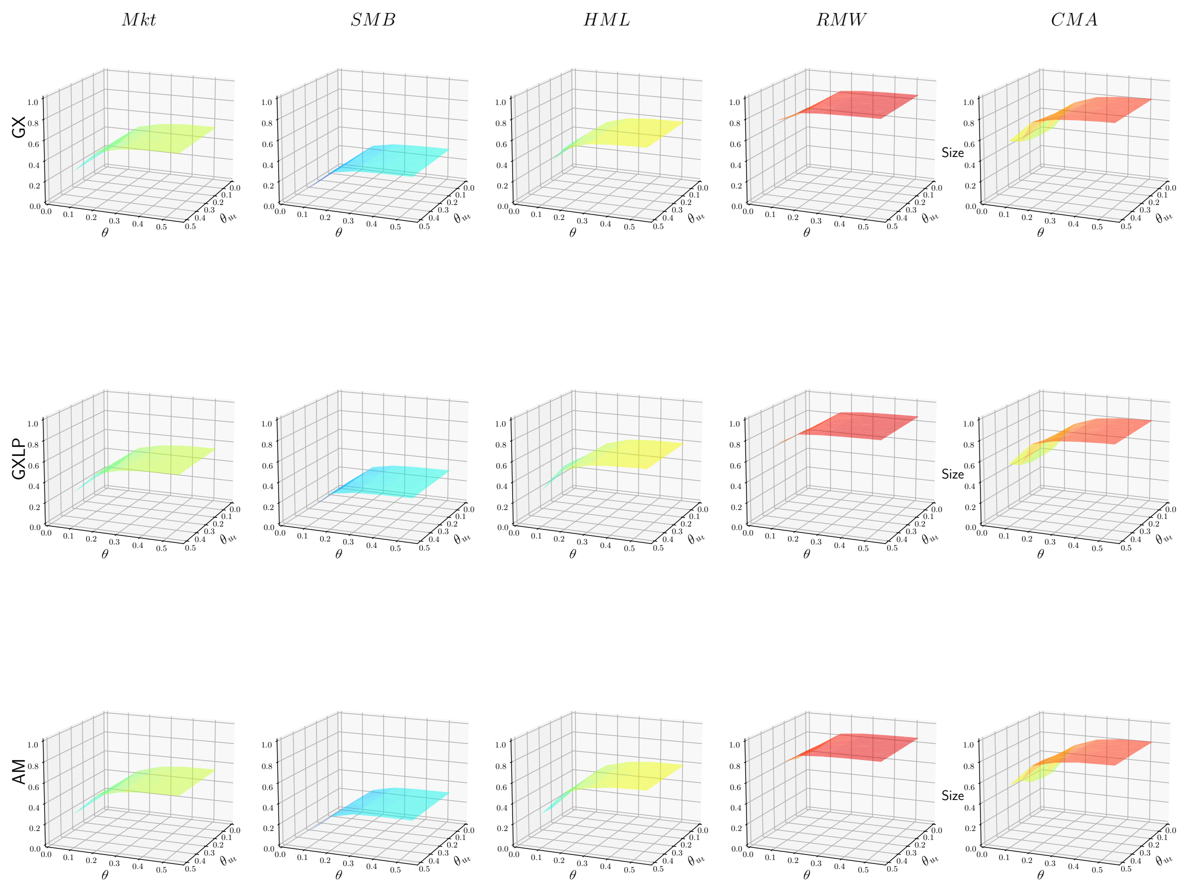


(a)

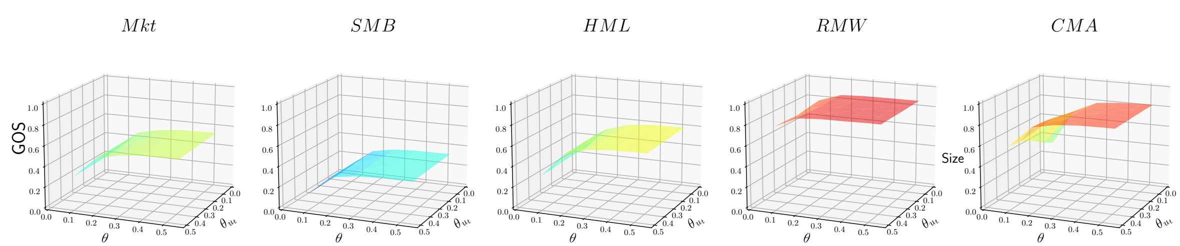


(b)

Figure 23: Adjusted Power of the test when loadings are local and strong/semi-strong. $N = 400, T = 400$. PCA-based estimators and Anatolyev and Mikusheva (2021)'s (panel a) and Gagliardini et al. (2016) estimator (panel b).



(a)



(b)

Figure 24: Size of the test when loadings are local and strong/semi-strong. $N = 400, T = 400$. PCA-based estimators and Anatolyev and Mikusheva (2021)'s (panel a) and Gagliardini et al. (2016) estimator (panel b).

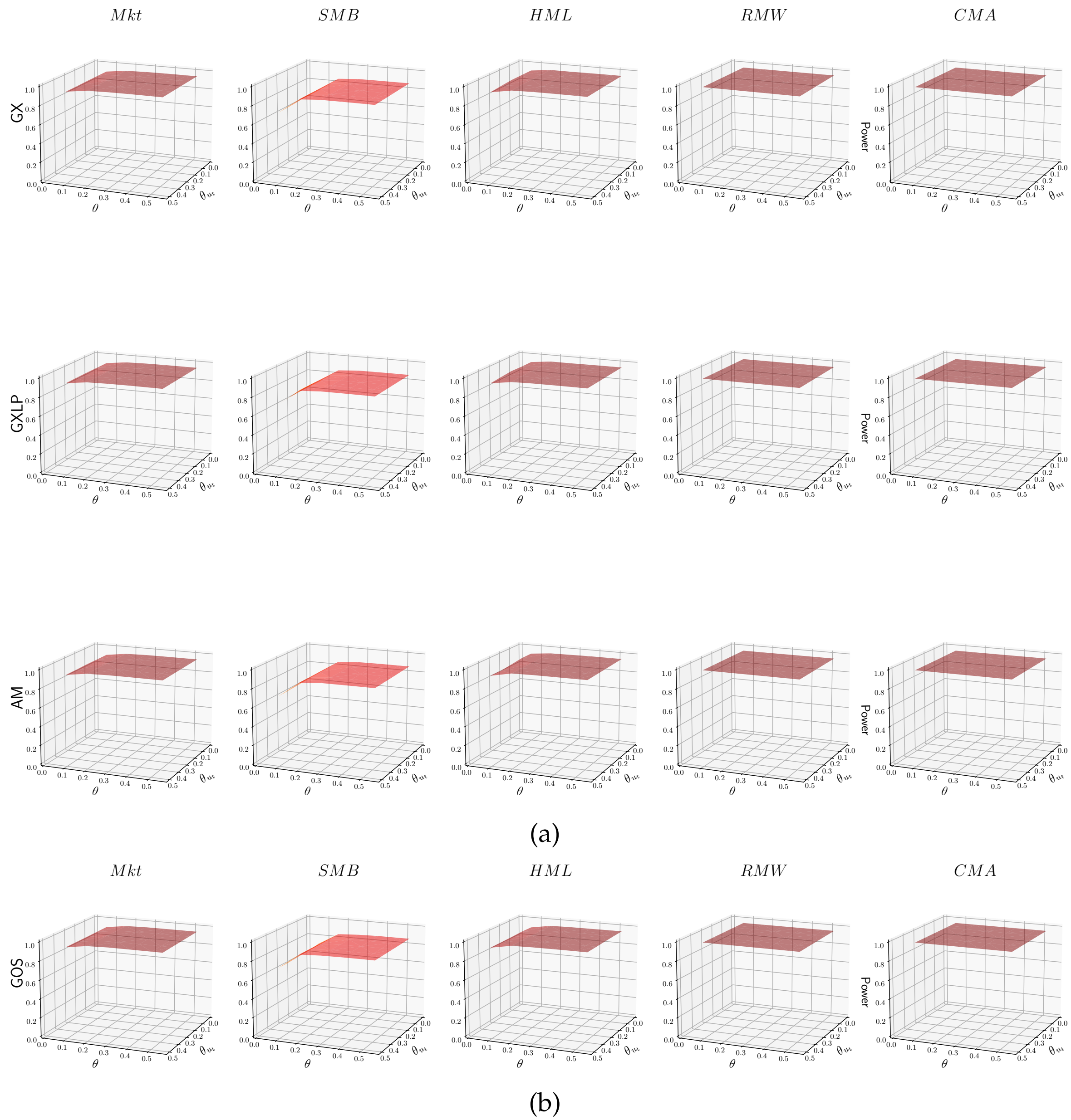
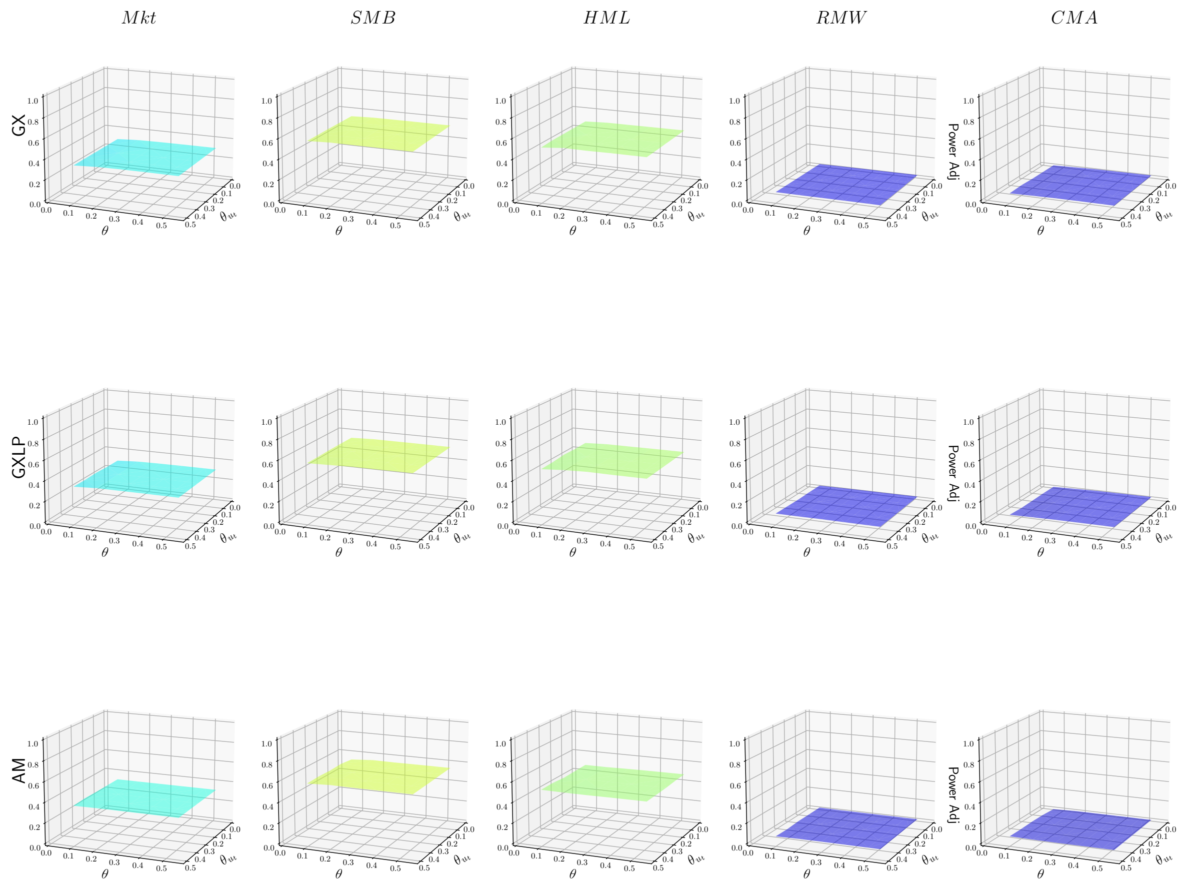
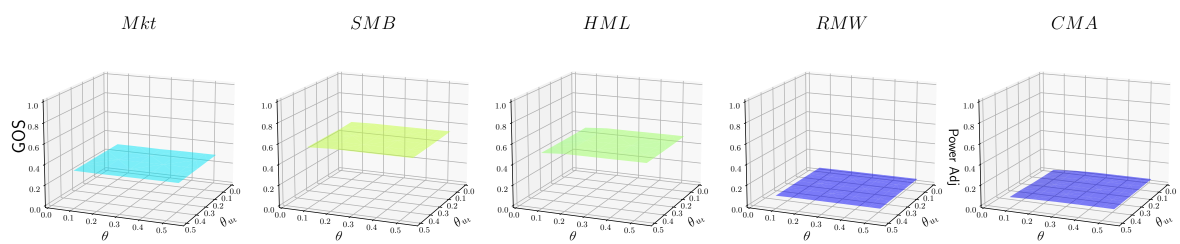


Figure 25: Power of the test when loadings are sparse. $N = 400, T = 400$. PCA-based estimators and Anatolyev and Mikusheva (2021)'s (panel a) and Gagliardini et al. (2016) estimator (panel b).



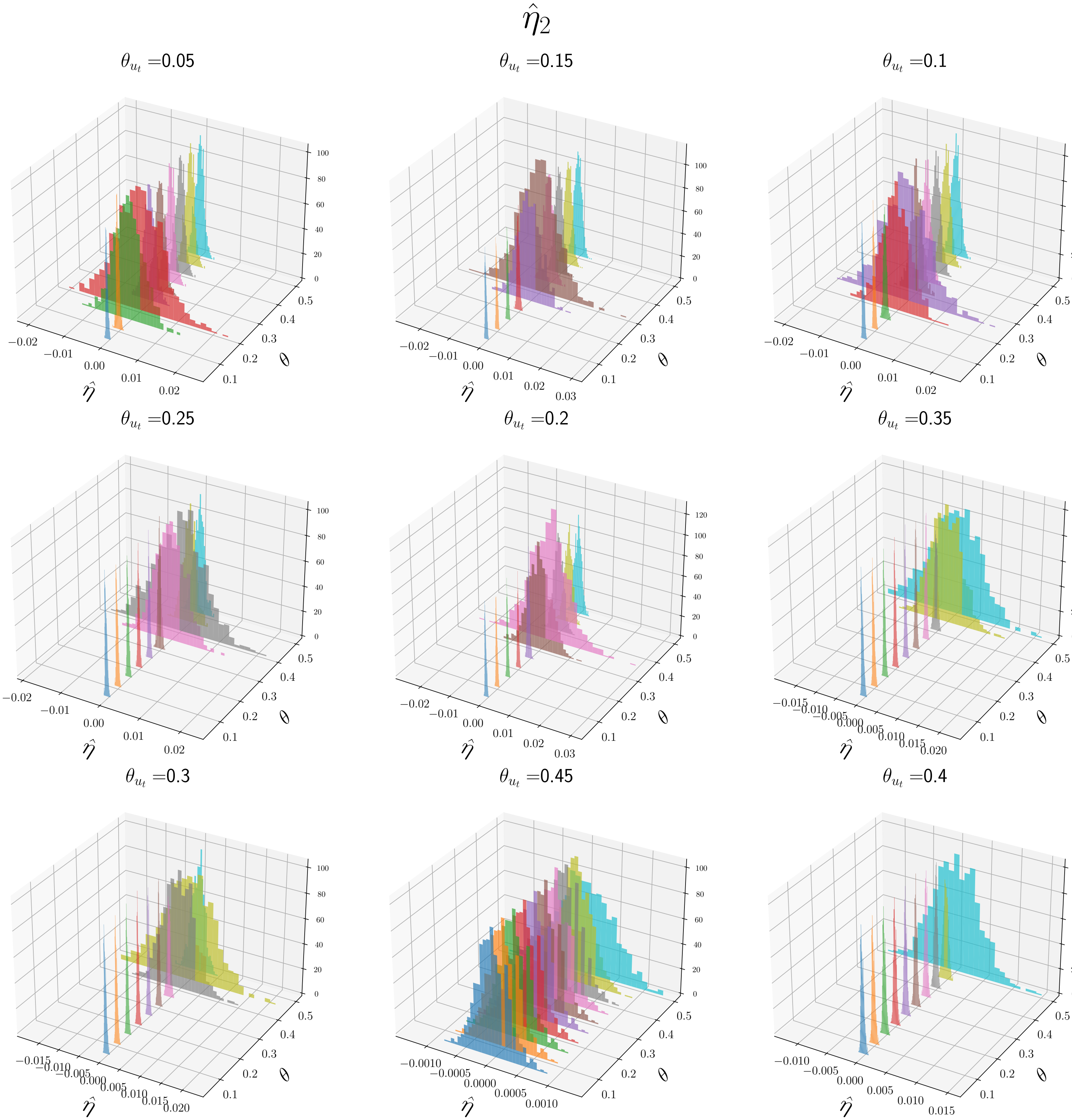
(a)



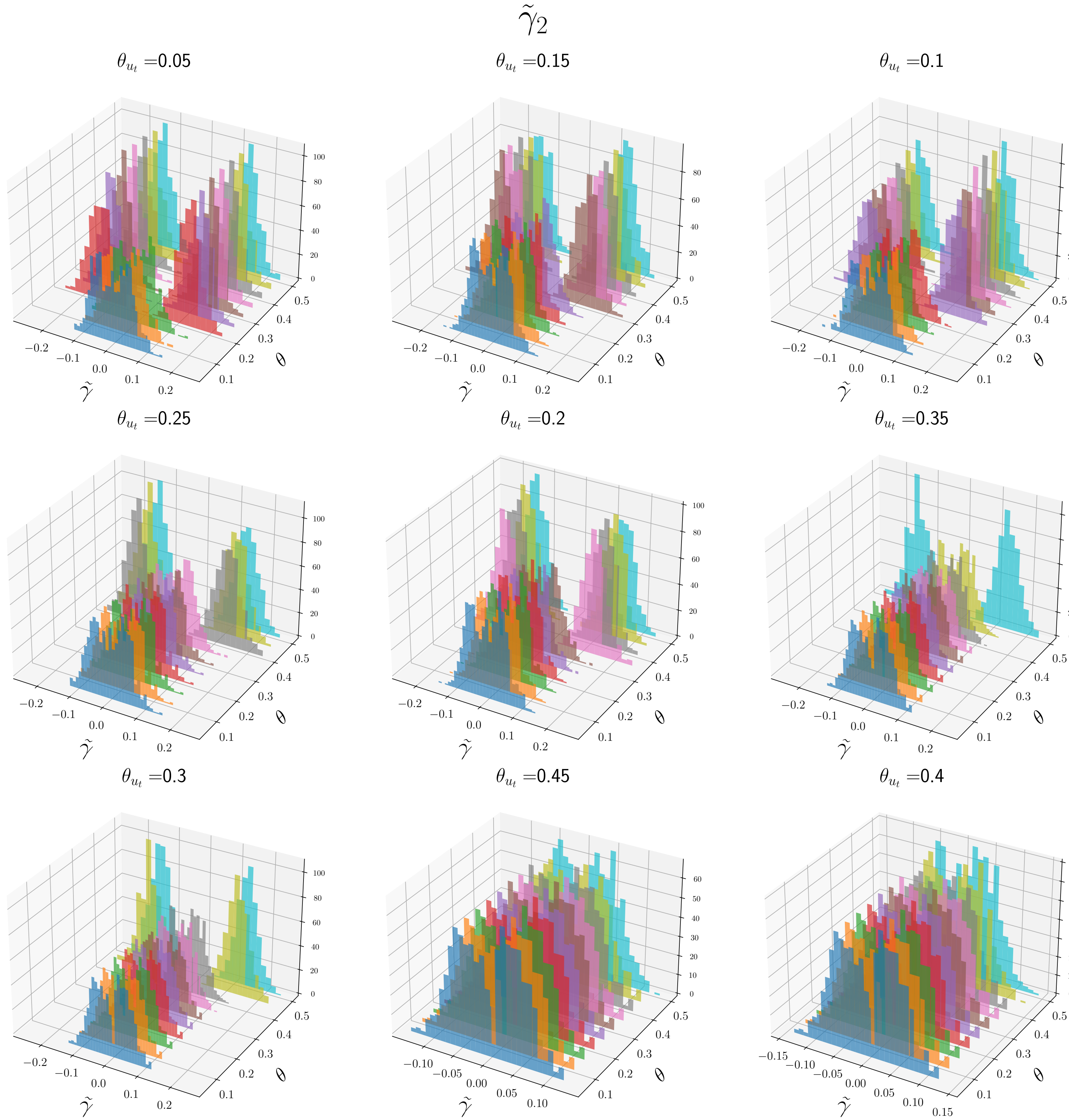
(b)

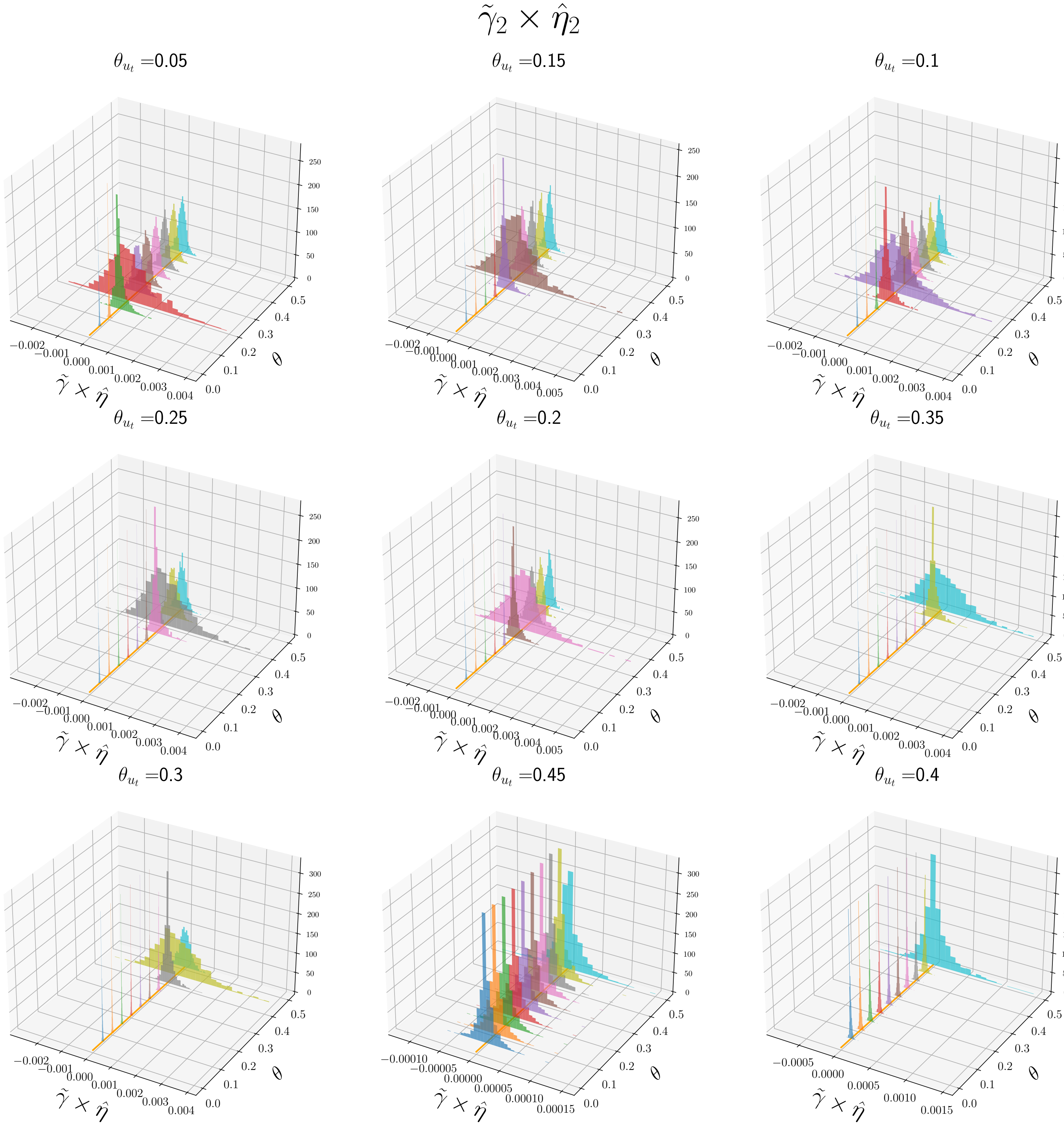
Figure 26: Adjusted Power of the test when loadings are sparse. $N = 400, T = 400$. PCA-based estimators and Anatolyev and Mikusheva (2021)'s (panel a) and Gagliardini et al. (2016) estimator (panel b).

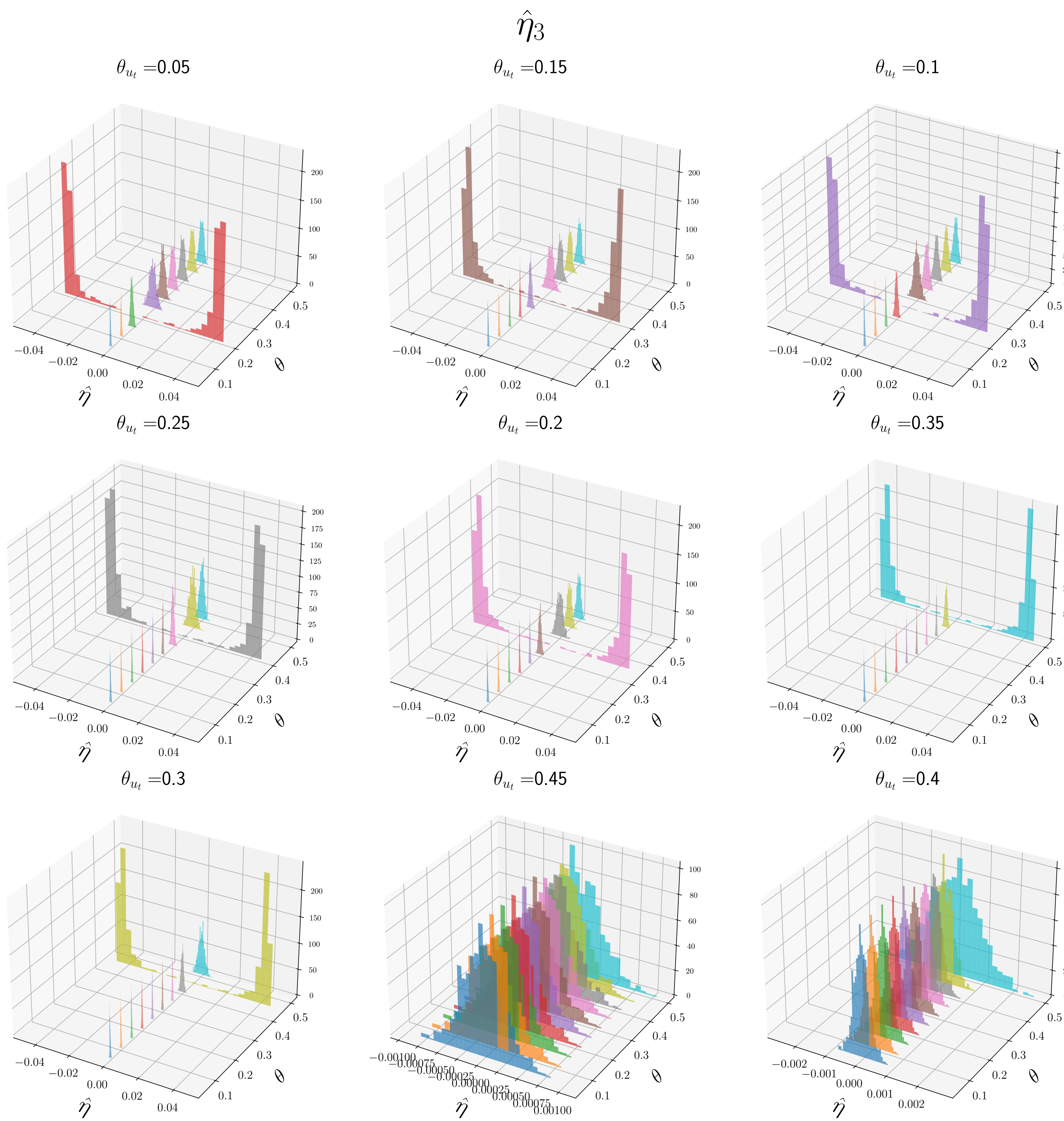
C Appendix: Dissecting Giglio and Xiu (2021)

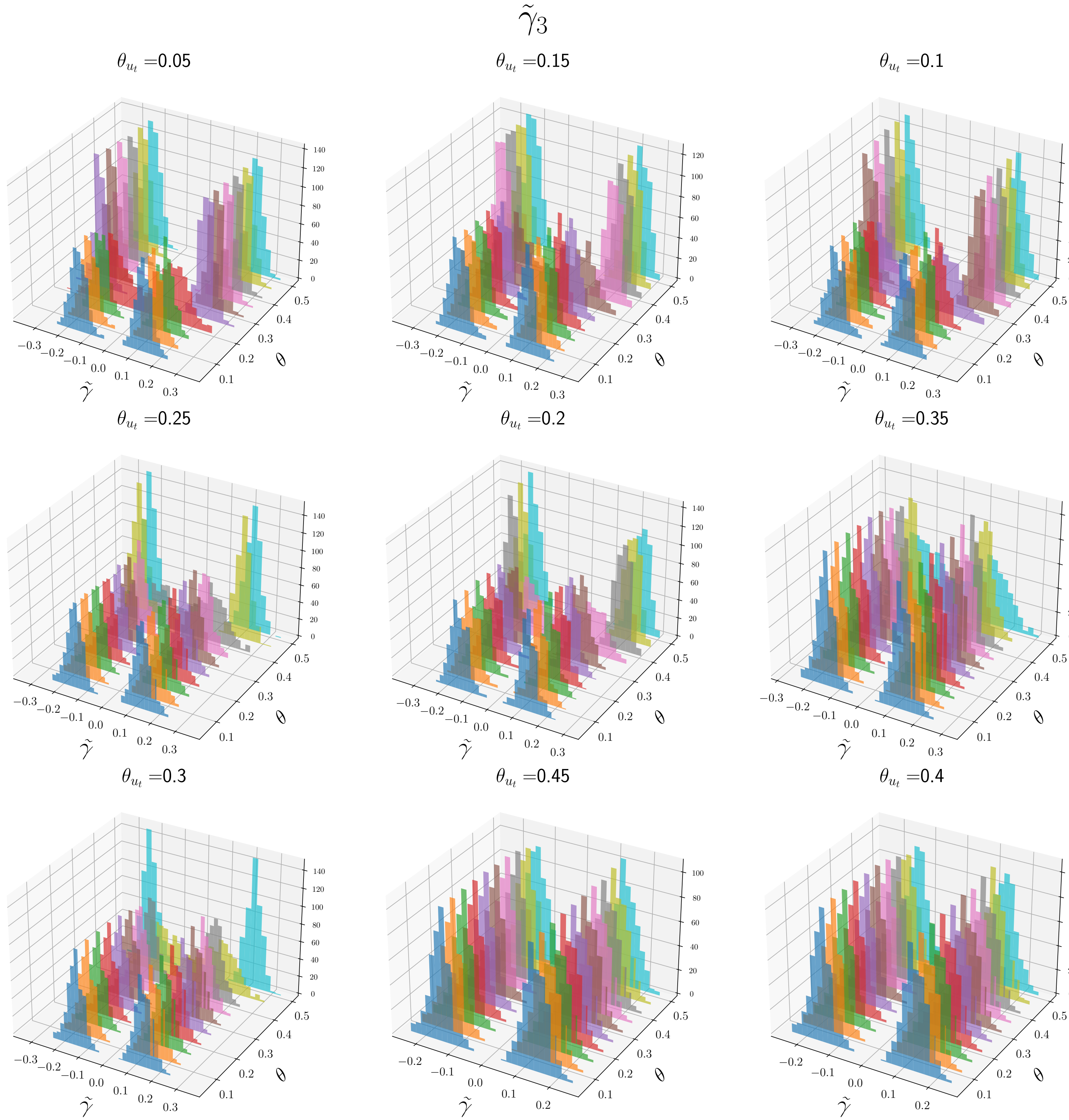


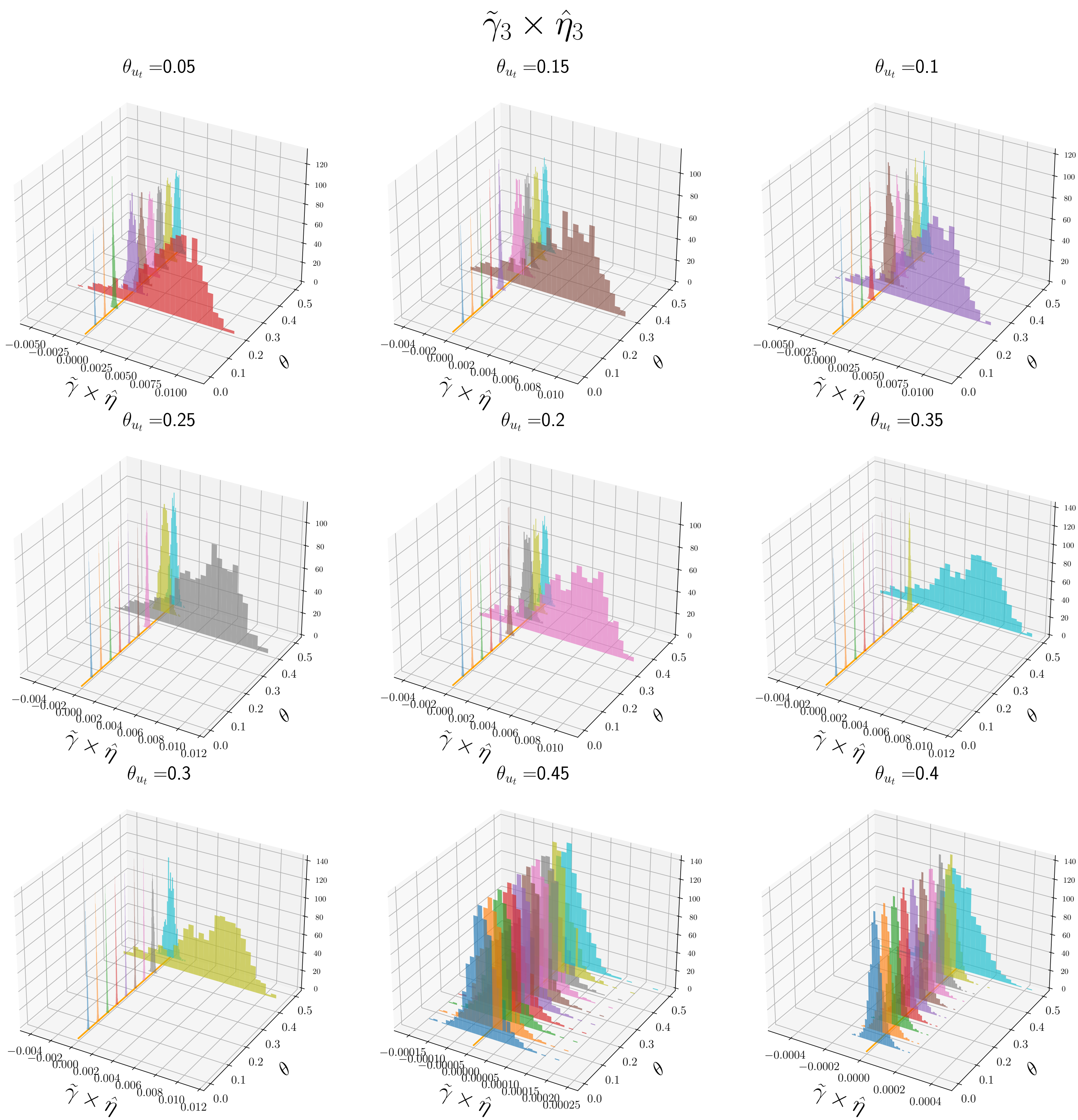
$\hat{\eta}_2$
θ_{u_t} =0.05
θ_{u_t} =0.15
θ_{u_t} =0.1
θ_{u_t} =0.25
θ_{u_t} =0.2
θ_{u_t} =0.35
θ_{u_t} =0.3
θ_{u_t} =0.45
θ_{u_t} =0.4
$\hat{\eta}$
θ


$\tilde{\gamma}_2$
θ_{u_t} =0.05
θ_{u_t} =0.15
θ_{u_t} =0.1
θ_{u_t} =0.25
θ_{u_t} =0.2
θ_{u_t} =0.35
θ_{u_t} =0.3
θ_{u_t} =0.45
θ_{u_t} =0.4
$\tilde{\gamma}$
θ


$\tilde{\gamma}_2 \times \hat{\eta}_2$
θ_{u_t} =0.05
θ_{u_t} =0.15
θ_{u_t} =0.1
θ_{u_t} =0.25
θ_{u_t} =0.2
θ_{u_t} =0.35
θ_{u_t} =0.3
θ_{u_t} =0.45
θ_{u_t} =0.4
$\tilde{\gamma} \times \hat{\eta}$
θ


$\hat{\eta}_3$
θ_{u_t} =0.05
θ_{u_t} =0.15
θ_{u_t} =0.1
θ_{u_t} =0.25
θ_{u_t} =0.2
θ_{u_t} =0.35
θ_{u_t} =0.3
θ_{u_t} =0.45
θ_{u_t} =0.4
$\hat{\eta}$
θ


$\tilde{\gamma}_3$
θ_{u_t} =0.05
θ_{u_t} =0.15
θ_{u_t} =0.1
θ_{u_t} =0.25
θ_{u_t} =0.2
θ_{u_t} =0.35
θ_{u_t} =0.3
θ_{u_t} =0.45
θ_{u_t} =0.4
$\tilde{\gamma}$
θ



$\tilde{\gamma}_3 \times \hat{\eta}_3$
θ_{u_t} =0.05
θ_{u_t} =0.15
θ_{u_t} =0.1
θ_{u_t} =0.25
θ_{u_t} =0.2
θ_{u_t} =0.35
θ_{u_t} =0.3
θ_{u_t} =0.45
θ_{u_t} =0.4
$\tilde{\gamma} \times \hat{\eta}$
θ