Characteristics-based factor modeling via Reduced Rank Regression *

Luca Pezzo[†] Raja Velu[‡] Lei Wang[§] Zhaoque(Chosen) Zhou[¶]

June, 2023

Abstract

We provide a framework for extracting characteristics-based factors via Reduced Rank Regression. This generalizes the Instrumented Principal Component Analysis by Kelly et al. (2019), the Projected Principal Component Analysis in Fan et al. (2016b), can accommodate cross-sectional and time-series dependencies, and recovers the closest lower-dimensional approximation to GLS factors discussed in Kozak and Nagel (2023). The asymptotic theory is derived and a bias in the IPCA inference is corrected. A sparse design is introduced to interpret the factors. Our findings confirm that accounting for cross-sectional dependence results in more efficient estimators leading to a better fit and a higher spanning.

JEL-Classification: C23, G11, G12

Keywords: Cross-sectional returns, Mean-Variance spanning, GLS, Industry-clustering, Factors, Principal components, Sparseness

*We are grateful to Ravi Jagannathan, Bryan T. Kelly, Soohun Kim, Serhiy Kozak, Yuan Liao, Tongshu Ma, Seth Pruitt, Ruey Tsay, Junbo Wang, Lai Xu, Guofu Zhou, and seminar participants at Tulane University, Louisiana State University for their helpful comments.

†University of New Orleans, department of Economics and Finance, +1 314 422 6083, lpezzo@uno.edu.
‡Syracuse University, Whitman School of Management, rpvelu@syr.edu.
§University of New Orleans, department of Economics and Finance, +1 716 589 4183, lwang9@uno.edu.
¶Washington University in St. Louis, John M. Olin Business School, zhaoque@wustl.edu.

1 Introduction

Since the publication of the seminal work of Fama and French (1993) there has been a proliferation of studies about characteristics-based factors. Harvey and Liu (2022) document that over 500 characteristics have been identified as possible factors in the literature that can explain the cross-sectional variation in returns. Historically, testing of multi-factor asset pricing models are done on portfolios of select stocks. To minimize the discretionary bias involved in the selection of the stocks and the predictors, empirical asset pricing models now directly use the entire cross-section of stocks and their associated characteristics.

In a recent study, Daniel et al. (2020) have shown that the mean-variance spanning of characteristics-based factors is biased downward due to the presence of unpriced risk and demonstrated how to construct hedge portfolios to attenuate the bias. Kozak and Nagel (2023) recognize that a source of mean-variance inefficiency stems from the fact that characteristics-based factors do not utilize the information from the covariances of the stock returns. It is pointed out that GLS factors, the slopes of monthly cross-sectional GLS regression of returns on characteristics, are indeed efficient. Because estimating a large covariance matrix for GLS estimation is notoriously a challenging task, Kozak and Nagel (2023) derive the necessary conditions for characteristics-based factors to be efficient, that is, to reach the same spanning as the GLS factors without using the information from the stock covariances.[1] However, such conditions hold in rather special circumstances. In the context of this study, it can happen only when the characteristics included in the model can pick up all the priced risk factors, a fact that cannot be guaranteed.

The insights from these recent studies are developed at the population level, assuming all parameters are known. Here we address the problem from an econometric perspective. We develop a flexible model in the Reduce-Rank-Regression (RRR) framework to study the sample properties of the estimators of characteristics-based factors. We further highlight their connections to leading models such as the PPCA model by Fan et al. (2016b) and the

[1]For early work on the equivalence between OLS and GLS estimators see Rao (1967).

IPCA model by Kelly et al. (2019). Specifically, we show that working directly with the covariance of stock returns is critical. From the commonality studies, it is well-known that the empirically observed covariances of stock returns play an important role in capturing stock co-movements (see Hasbrouck and Seppi (2001)). From an economic perspective, our setup shows how accounting for stock covariances can incorporate information about any unpriced risk in the estimators for the characteristics-based factors, and thus guaranteeing their mean-variance efficiency. Moreover, our model can strike a balance between maximal mean-variance spanning and parsimonious use of characteristics-based factors. It extracts the closest lower-dimensional approximation to the efficient estimators of GLS factors.

The main focus of cross-sectional models in studying the first two moments of the return distribution can be rationalized by the existence of a stochastic discount factor (SDF), linear in a set of characteristics-based factors. The existence of such SDF stems from the assumption that the law of one price holds and market frictions (like transaction costs and price impact) are negligible. When the return distribution is assumed to be stationary, the SDF has a time-invariant structure and the equivalent factor model shares the same data generating process implied by the Arbitrage Pricing Theory (APT) of Ross (1976). The classical RRR setup (see Reinsel et al. (2022), chapter 2) applies in such a context to models where factors are based on a (large) number of common macro variables. If characteristics are allowed to be stock-specific but static, Fan et al. (2016b) show that extracting factors through principal components from projected excess returns delivers more accurate factors. The Projected PCA procedure (PPCA) is a non-parametric regression of returns on characteristics and can be taken as a non-parametric version of a RRR model.

When the SDF structure is set to vary over time, the resulting conditional factor models can accommodate the time-varying nature of the investment opportunity set. This can be achieved by having stock-specific characteristics that vary over time and ascribing such variation to the model intercept and factor loading matrix. Kelly et al. (2019) successfully treat such characteristics as instruments and extract PCA-style latent factors by condensing down the relevant pricing information from L characteristics to $K < L$ linear combinations of them. As shown here in the RRR setup, such dimension reduction is possible due to binding constraints implied via the rank of the regression coefficient matrix from a system of T cross-sectional regressions that are appropriately stacked. Thus the IPCA model also has an implied RRR structure and the determination of K, the number of factors, can be made via a rank test. In the absence of the rank constraint, all L characteristics-based factors can be extracted. The estimated regression coefficient in this case are the BARRA OLS factors used in Fama and French (2020).

To summarize, with time-invariant asset characteristics and with homoscedastic errors, our model can be taken as a linear parametric version of the PPCA model. If the characteristics vary over time, we are in the realm of the IPCA model. By specifying the cross-sectional dependencies in the error term with a positive-definite covariance matrix, our model generates for each time 't' the closest $K < L$-dimensional approximation to the L-dimensional vector of mean-variance efficient GLS factors. Specifically, we assume the same structure for the GLS factor matrix as in the IPCA. This has two components: a time-invariant mispricing factor representing the lower dimensional approximation of the vector of unconditional expected values of the GLS factors and the product of a $L \times K$ factor loading matrix and a K-dimensional vector of factors, which we refer to as the *RRR* factors. The loading matrix maps the L stock characteristics into the K *RRR* factors (and can be equivalently thought of as the loading matrix in a static factor model for L-dimensional factors) and the resulting *RRR* factors capture the K most important innovations present in the GLS factors. Our setup naturally relates the mean-variance efficiency of the extracted *RRR* factors to the absence of mispricing. For the case of $K = L$ the *RRR* factors are proportional to the mean-variance efficient GLS factors, thus achieving the same mean-variance spanning, and at the same time the mispricing factor is zero. When $K < L$ the *RRR* factors are mean-variance efficient lower-dimensional approximations of the GLS factors only when the mispricing factor is zero. Because the IPCA model assumes that the error covariance matrix is proportional to an identity matrix, it follows that the IPCA factors can be taken as a lower-dimensional approximation to the OLS factors, which are not mean-variance efficient. Finally, our setup can also handle time-series variation coming from the error term by imposing a VAR(1) or other simplified structures on the error term as in DeMiguel et al. (2014), while still extracting the lower-dimensional factors as in the IPCA, thus providing a robust approach to utilize the empirical features of the data.

We derive the limiting distribution for the estimates of the parameters of the RRR model. This allows us to employ directly closed form expressions for hypothesis tests that are computationally less intensive. In Kelly et al. (2019), a Wald-type test for mispricing is suggested, however the test appears to have a bias due to an incorrect assumption made on the data generating process. The result from the limiting distribution enables us to correct for the nontrivial bias present in the inference based on bootstrapping samples. The suggested correction realigns the IPCA model results with the results using the correct data generating process. In the spirit of Cochrane (2011)'s presidential address, to better understand which characteristics matter the most, we augment our RRR model, which provides the dimension-reduction, with sparseness methods that address the variable selection problem. Specifically we place a LASSO penalty on the factor loading matrix which maps the L characteristics to the K-RRR factors. In this context, the sparseness structure of the columns of the loading matrix isolates the driving characteristics behind each extracted factor and thus provides a way to interpret the extracted factors themselves.

To assess the performance of the RRR model, we use the same data considered in Kelly et al. (2019). This includes monthly returns and 36 asset characteristics for approximately 10,000 U.S. common stocks between July 1962 and May 2014. The characteristics do generally exhibit significant time variation (making an IPCA/RRR model structure more suitable than a PPCA one). We account for the cross-sectional covariances through an industry-wide block-diagonal structure for the error covariance matrix. Industry-wide clustering is motivated among other studies, from the observation made in Daniel et al. (2020) that industry exposure represents a source of unpriced risk and it also reflects the natural structure of the stock market. We select an industry structure that provides the best in-sample fit (GLS R^2, BIC and AIC) and results in the best mean-variance spanning (maximal Sharpe ratios).

Under this structure, we show empirically that feasible versions of the GLS factors result in a higher mean-variance span and fit than the OLS factors and than the OLS-hedge factors proposed in Kozak and Nagel (2023) and Daniel et al. (2020). More generally, if at least five factors are extracted, the RRR model without mispricing always displays the best fit and spanning. What mainly drives these results is the higher efficiency of the GLS estimators; The higher spanning of the factors that use GLS estimators come from lower volatilities in the tangency portfolios (the denominator of the Sharpe ratio). As a matter of fact, the tangency portfolio averages are found to be generally lower for GLS factors than OLS factors. According to the mean-variance theory these results imply that investors who have access to a better proxy for risk (through the residual covariance matrix specification of our setup) should become more cautious in taking on risk.

The model that strikes the best trade off between parsimonious number of factors and mean-variance spanning is the RRR model with five factors (and no mispricing factor). Such a model is shown to outperform the IPCA analog which in turn Kelly et al. (2019) show to outperform the CAPM and the Fama and French (3 through 6) factor models. The sparseness as well as the validation analysis are then carried out on the best model directly. The sparseness analysis of the RRR model reveals a much simpler factor structure than the sparseness structure that is detected for the IPCA model, that lends itself to an elegant interpretation: we are able to identify the first factor as the market, the second as size, the third and fourth as momentum and the fifth as illiquidity. The superior performance of the best RRR model over the IPCA analog is confirmed both out of sample and via simulations. We provide two types of out of sample validations: a cross-sectional design where the model is estimated using a random subsample containing 80% of the available stocks and then the model is tested on the excluded 20% of the stocks; and a time series design where the model is recursively estimated over a 20-year rolling window and tested over one-month-ahead forecasts. Simulations also confirm that the RRR model is stable and gets closer to the true data generating process. In general, the superior spanning comes from the lower variance of the RRR estimators but the tangency portfolio averages are generally found smaller confirming the in-sample pattern supporting the fact that investors having access to a better proxy for risk appear to be more risk averse.

There are some recent studies that have focused on conditional asset pricing modeling with a large number of asset characteristics using both parametric and non-parametric methods. The non-parametric methods such as splines, adaptive group LASSO, etc are shown to result in better out-of-sample predictions of the conditional returns. Freyberger et al. (2020) use 62 asset characteristics and isolate the power of individual characteristics to forecast future returns. Clarke and Momeni (2021) develop tests for asset pricing models using individual assets rather than sorted portfolios, but conclude that the IPCA model which uses information on characteristics across assets fares better. Clarke and Linn (2023) use pairwise return covariances and relate them to a large number of asset characteristics, with an assumption that if a characteristic can predict future returns, it can also help explain the covariances. These studies supposedly complement the dimension reduction methods such as IPCA, RRR, etc. The emerging interest in evaluating a large number of asset characteristics for their predictive power for future returns, thus involve both theoretical reasoning, such as Daniel et al. (2020) and Kozak and Nagel (2023), and empirical reasoning, followed in this paper. On the methodology side, Zhang (2023) has extended the IPCA to include time-varying alphas to capture the pricing errors that can come from outside the space spanned by the asset characteristics. These alphas are taken to be orthogonal to factor loadings of the asset characteristics and do not have exposure to the resulting common factors. This extension is shown to result in the null pricing error hypothesis being rejected for several, commonly used, characteristics-based models.

There are multiple contributions to the literature from this work. By casting the dimension reduction (from characteristics to factor) models in the regression framework, we provide a richer modeling framework through a pooled Reduced-Rank Regression as well as an asymptotic theory for the estimators. In this framework, we can accommodate the cross-sectional covariances as dictated by theory as well by empiric but also chronological dependencies. The dimension-reduction feature is further augmented by the sparseness methods that lead to variable selection for elegant interpretations. We empirically demonstrate how the methodology has good predictive power. The in-sample and out-of-sample performance of the estimators convey that the RRR estimators are stable and the inference is robust.

The rest of the paper is structured as follows. Section 2 formally introduces the model. In Section 3, details of the estimation of the model parameters, highlighting their properties and their interpretation are provided. In Section 4, we present the asymptotic results, point out the bias in the IPCA inference and offer some mispricing tests. In Section 5, a sparse version of our model is introduced. Section 6 describes the empirical design, including the data, the performance metrics that are used to evaluate the models, the specification of the residual covariance matrix of stock returns, and the validation assumption for the sparseness setup. Section 7 presents the empirical results of the study, starting with the in-sample analysis and then moving on to the out-of-sample analysis. Finally, simulation results are presented for the best selected model from the study. Section 8 concludes. Some technical details can be found in the Appendix.

2 Reduced Rank Regression Model Formulation

One of the main goals of empirical assert pricing research is to understand the risk-return relationship by looking at the first two moments of the stock return distribution. The classical theory assumes that the law of one price holds and the market is frictionless.[2] In this context, it can be shown that there exists a stochastic discount factor, linear in a (small) number of factors that perfectly price the cross-section of returns. Moreover, if the covariance between such factors and the returns is proportional to the stock risk premia, then risk premia can be consistently modeled as a linear function of the factors. Because the theory does not provide any guidance on the nature of such factors, a widely used approach is

[2]This means, among other things, investors are treated as price takers and the effect of transaction costs is considered to be of second order.

to retrieve them indirectly through dominant principal components of the covariance matrix of returns. This cross-sectional analysis results in unconditional factor model with static factor loadings.

However, to capture the time-varying nature of the investment opportunity set, a conditional factor model is more suitable. In order to estimate the resulting factor model, two approaches are taken in the literature; either the factors are to be assumed known or factor loadings are taken to be time-varying that can be related to asset specific characteristics. In the absence of economic theory, specifying correctly the exact (latent) factors is an arduous task. Therefore, several asset characteristics(instruments) are considered that can possibly represent these factors. In this section, we formulate these models in the multivariate regression framework that unifies various applications found in the literature.

The basic data consists of $(r_{i,t+1}, Z'_{i,t})$, where $r_{i,t+1}$ is the excess return of the i^{th} asset $(i = 1, ..., N)$ at time t $(t = 1, ..., T)$ with $Z_{i,t}$ being the $L \times 1$ vector of associated asset characteristics. Let Z_t be the $N \times L$ matrix of characteristics associated with 'N' assets. The first two moments of the returns, conditioned on the characteristics, are given as:

$$\mu_r(t) \equiv E[r_{t+1}|Z_t] \ , \ \Sigma_{rr}(t) \equiv var(r_{t+1}|Z_t) \tag{1}$$

where $\Sigma_{rr}(t)$ the conditional covariance matrix is assumed to be a positive definite matrix. If the law of one price holds (and market frictions are of second order), then a Stochastic Discount Factor (SDF), m_{t+1}, exists and is unique in the span of the excess returns r_{t+1} such that $E_t[m_{t+1}r_{t+1}|Z_t] = 0$, and is linear in a $K \leq L$ dimensional factors f_{t+1}. Therefore, under the assumption that stock risk premia are proportional to the covariances between the stock returns and the factors, the following model representation for the excess returns r_{t+1} is equivalent to the existence of such SDF (see Section 6.3 in Cochrane (2005)),

$$r_{t+1} = \alpha_t + A_t f_{t+1} + \varepsilon_{t+1} \ , \ \varepsilon_{t+1} \sim N(0, \Sigma_{\varepsilon\varepsilon}(t)). \tag{2}$$

Here α_t is a $N \times 1$ vector capturing the portion of excess returns that are not coming from the exposure to the factors and A_t is a $N \times K$ conditional factor loading matrix. When $\alpha_t = \alpha, A_t = A$ and $\Sigma_{\varepsilon\varepsilon}(t) = \Sigma_{\varepsilon\varepsilon} = \Sigma$, model (2) reduces to

$$r_{t+1} = \alpha + A f_{t+1} + \varepsilon_{t+1} \ , \ \varepsilon_{t+1} \sim N(0, \Sigma), \tag{3}$$

the APT-implied data generating process for the excess returns in the unconditional setup. This is equivalent to a SDF with time-invariant intercept and a L-dimensional vector of slopes. Observe that in (2) and (3), only r_{t+1} is observable. The factors f_{t+1} in model (3) are estimated as the principal components of the covariance matrix of the returns. Alternatively if the factors are taken to be related to exogenous macro characteristics, we show how it naturally leads to a reduced rank multivariate regression model.

Suppose the factors f_{t+1} are assumed to be approximately linear in Z_t via

$$f_{t+1} = B \cdot Z_t + a_{t+1}, \tag{4}$$

then model (3) along with (4) becomes

$$r_{t+1} = \alpha + AB \cdot Z_t + \varepsilon^*_{t+1} = \alpha + C \cdot Z_t + \varepsilon^*_{t+1}. \tag{5}$$

The identifying restriction in model (5) is represented by the rank of the regression coefficient matrix C, that is $rank(C) = K < L$. In practice, the value of K can be specified based on an apriori theory or it can be determined empirically based on the canonical correlations between r_t and Z_t. This rank condition has two practical implications with elegant interpretations:

$$\underset{N\times L}{C} = \underset{N\times K}{A}\underset{K\times L}{B}; \qquad l_i' C = 0, \quad i = 1, 2, ..., (L-K). \tag{6}$$

Observe that the term $BZ_t = f_{t+1}$ can be interpreted as the vector of known factors recovering most of the useful information from Z_t, and the linear combinations, $l' r_{t+1} = l'\alpha + l'\varepsilon^*_{t+1}$, are independent of the Z_t characteristics, thus indicating that certain stock returns can be modeled without any reference to asset characteristics. The rows of A can be interpreted as the influence of factors f_{t+1} on each asset. As mentioned, if we let $\underset{N\times T}{R} = [r_2, ..., r_{T+1}]$ and $\underset{L\times T}{Z} = [Z_1, ..., Z_T]$ as data matrices, the rank of the C matrix can be simply tested by the significant canonical correlations between r_{t+1} and Z_t, or equivalently, through the singular values of $W_1^{\frac{1}{2}} \cdot C \cdot W_2^{\frac{1}{2}}$, where W_1 and W_2 are appropriately chosen weight matrices. These choices are usually based on the asymptotic distribution of the LS estimator of the C coefficient matrix. For model (5), the choices are, $W_1 = \Sigma^{-1}$, the inverse of the covariance matrix of the excess returns R (or the errors) and $W_2 = \Sigma_{ZZ}$, the covariance matrix of the stock-characteristics Z_t. Details on the estimation and inference of model (5) with constraint (6) can be found in (Reinsel et al., 2022, Chapter 2). Note that if $Z_t = r_t$, model (5) is a vector autoregressive model (VAR(1)) with additional constraint on the stationarity of r_t. It is well-known that there is a close relationship between the non-stationarity of the process and the canonical correlations between r_{t+1} and r_t, and hence the singular values of $W_1^{\frac{1}{2}} \cdot C \cdot W_2^{\frac{1}{2}}$.

The RRR model in (5) has found some applications in finance. For an elegant application of model (5) in the bond market, where $rank(C) = 1$ arises naturally, see Cochrane and Piazzesi (2005). Another example, the "sieve reduced rank regression" model of Adrian et al. (2019) is a non-linear version of (5) with Z_t representing the VIX, the implied volatility index at time t. There are other recent models extending principal component in the literature that can be formulated in the reduced-rank regression framework. A prominent example is given by the projected PCA (PPCA) model of Fan et al. (2016b). The model assumes the L characteristics to be time-invariant but stock specific, that is $Z_{i,t} = Z_i, i = 1, ..., N$ stacked in $Z = [Z_1, ..., Z_N]'$. The factors are related to $g(Z)$, not necessarily linear. This amounts to re-write model (4) as,

$$f_{t+1} = g(Z) + a^*_{t+1}. \tag{7}$$

Thus model (3) can now be written in matrix form as

$$\underset{N\times T}{R} = \underset{N\times K}{A}(\underset{K\times T}{G(Z)} + \Gamma)' + \underset{N\times T}{U} \tag{8}$$

where Γ is a $T \times K$ matrix that contains the factors that cannot be explained by the variables Z and U represents the idiosyncratic errors. Approximating $G(Z) \approx \Phi(Z)B'$ where $\Phi(Z)$ is a matrix of basis functions, observe that

$$R = A \cdot B \cdot \Phi'(Z) + U \tag{9}$$

which is in the RRR form of model (5). If $P = \Phi(Z)(\Phi'(Z)\Phi(Z))^{-1}\Phi'(Z)$ is the projection matrix, then

$$R \cdot P = A \cdot B \cdot \Phi'(Z) + E. \tag{10}$$

The projected data, $R \cdot P$ is used to estimate A and B; the resulting estimates are called projected principal components. Observe that the projected data $R \cdot P$ is smoother than R, resulting in estimators having certain nice properties. The PPCA model of Fan et al. (2016b) can be taken essentially as a non-parametric version of a RRR model when stock characteristics are time-invariant. This model is used for forming arbitrage portfolios, using the static asset characteristics, by Kim et al. (2021).

2.1 RRR Model for Conditional Risk and Return

The basic factor model stated in equation (2) is quite general. All quantities that are related to returns vary over time. The reduced-rank model in equation (5), results from relating time-varying factors to asset characteristics. An alternate formulation is to relate the factor loadings to asset characteristics, thus providing a rich linkage between asset-specific loadings and the characteristics. This is the essence of the IPCA model of Kelly et al. (2019). The parameters of model (2) are specifically structured as:

$$\alpha_t = Z_t \Gamma_\alpha + v_{\alpha_t}, \qquad A_t = Z_t \Gamma_\beta + v_{A_t}. \tag{11}$$

This is different from the traditional factor models where the factors are assumed to be unknown as in (3) or known to be related to time-varying characteristics as in (4). The cross-sectional regression model then can be written as,

$$\underset{N\times 1}{r_{t+1}} = \underset{N\times L}{Z_t} \cdot \underset{L\times 1}{\beta_{t+1}} + \underset{N\times 1}{\epsilon_{t+1}}, \qquad \epsilon_{t+1} \sim N(0, \Sigma_t) \tag{12}$$

with

$$\underset{L\times 1}{\beta_{t+1}} = \underset{L\times 1}{\Gamma_\alpha} + \underset{L\times K}{\Gamma_\beta} \cdot \underset{K\times 1}{f_{t+1}} \tag{13}$$

and $K < L$. In the IPCA model of Kelly et al. (2019) it is assumed that $\Sigma_t = \sigma^2 I_N$, ignoring the return covariances.

In the absence of restriction (13) model (12) defines a system of panel regressions which can be run independently. Thus the L-dimensional GLS estimator, for a known Σ_t, is

$$\tilde{\beta}_{t+1}^{GLS} = (Z_t' \Gamma_t Z_t)^{-1} Z_t' \Gamma_t r_{t+1} \sim N\left(\beta_{t+1}, (Z_t' \Gamma_t Z_t)^{-1}\right) \tag{14}$$

where $\Gamma_t = \Sigma_t^{-1}$. Equation (14) defines the distribution of the unconstrained GLS factors. Restriction (13) defines β_{t+1} as a linear function of the K-dimensional vector of RRR factors f_{t+1}. When $\Gamma_\alpha = 0$, observe that equation (13) can be treated as a static factor model of the regression coefficients β_{t+1} with Γ_β representing the matrix of loadings mapping the GLS factors into the RRR factors. If $\Gamma_\alpha \neq 0$, equation (13) can be considered as a static factor model of the mean adjusted time-varying regression coefficients, $(\beta_{t+1} - \bar{\beta})$, where $\bar{\beta} = \frac{1}{T}\sum_t \beta_{t+1}$. Indeed as we show later, we can think of Γ_α as a close approximation of $\bar{\beta}$, proxying for mispricing. In Section 3, we show how $\hat{\Gamma}_\alpha + \hat{\Gamma}_\beta \hat{f}_{t+1}$ represents the closest K-dimensional approximation of the L-dimensional vector β_{t+1}. Specifically, f_{t+1} captures the K most important innovations in β_{t+1}. Moreover, such innovations themselves represent mean-variance efficient factors for a given K when $\Gamma_\alpha = 0$ and $\beta_{t+1} = \tilde{\beta}^{GLS}_{t+1}$.

The RRR nature of model (12), and thus of the IPCA model, becomes apparent once we reformulate the restriction (13) as a rank condition on the coefficients matrix stacked as,

$$\underset{T\times L}{\beta} \equiv \begin{bmatrix} \beta_2' \\ \vdots \\ \beta_{T+1}' \end{bmatrix} \equiv 1_T\Gamma_\alpha' + \begin{bmatrix} f_2' \\ \vdots \\ f_{T+1}' \end{bmatrix} \cdot \Gamma_\beta' \equiv \underset{T\times 1}{1_T} \cdot \underset{1\times L}{\Gamma_\alpha'} + \underset{T\times K}{F'} \cdot \underset{K\times L}{\Gamma_\beta'} \tag{15}$$

with the following rank constraint

$$rank(\beta - 1_T\Gamma_\alpha') \equiv rank(\beta_\alpha) = K < L. \tag{16}$$

The RRR model version of IPCA is not in the standard form of the classical RRR setup as defined in equation (5), because the time-varying predictors are related to loadings rather than factors. Although the slope vectors, β_{t+1} vary, their similarity empirically observed over time would suggest that a lower rank for the β matrix is possible. This is an empirical observation, not based on any apriori theory. As we show later, we can estimate an approximate value for K through a rank test based on the matrix formed by stacking as rows the GLS estimates $\tilde{\beta}^{GLS}_{t+1}$ for $t = 2, ..., T+1$. As discussed in Fan et al. (2021), knowing the true value for K would guarantee the consistency of the factor estimators.

We want to also comment on the usefulness of the RRR model formulation. The conditional RRR factor model presented in this section can also accommodate time-series dependence in the error terms as follows: If we denote the error term in (12) as, u_{t+1}, we can specify a model for the errors as,

$$u_{t+1} = \Phi_t u_t + \epsilon_{t+1} \ , \ \epsilon_{t+1} \sim N(0, \Gamma_t^{-1}). \tag{17}$$

Then model (12) along with (14) can be rewritten as

$$r^*_{t+1} \equiv r_{t+1} - \Phi_t r_t = Z_t \beta_{t+1} - \Phi_t Z_{t-1} \beta_t + \epsilon_{t+1}. \tag{18}$$

The estimator of this model along with the constraints (13) can be computed with the RRR methodology.

2.2 Why the covariance structure is important?

We want to emphasize the need to account for the covariance structure in the modeling of conditional returns. This represents the main difference between the RRR model studied here and the IPCA model. As observed in Hasbrouck and Seppi (2001) stock returns do exhibit strong commonalities measured in terms of the principal components of the covariance matrix. Ignoring this information can lead to a loss in statistical and economic efficiency of the portfolio-based estimators. From a statistical point of view, exploiting a more general positive definite covariance structure if it is present should result in estimators that are more efficient, which we show explicitly later. From the economic perspective, as pointed out in Daniel et al. (2020) and Kozak and Nagel (2023), neglecting stock covariances in the construction of factors can lead to the loss of mean-variance efficiency.

It is instructive to relate our model to the models studied in Daniel et al. (2020) and in Kozak and Nagel (2023). These authors advocate incorporating the covariances from economic perspective. The conditional version of the model in Daniel et al. (2020) can be stated as

$$r_{t+1} = A_t f_{t+1} + B_t g_{t+1} + u_{t+1} = Z_t \beta_{t+1} + \varepsilon_{t+1} \tag{19}$$

where g_{t+1} represents the unpriced factors, independent of the priced factors f_{t+1}, with $E[g_{t+1}] = 0$ and $Var[g_{t+1}] = V$, a positive definite matrix. It is assumed that $u_{t+1} \sim N(0, \sigma^2 I_N)$; with that $\Sigma_t = B_t V B_t' + \sigma^2 I_N$, with non-null covariances. It is well-known (see Rao (1967)) that the OLS estimate of f_{t+1} in (19) is efficient if $A_t' B_t = 0$, which implies that all the priced factors are captured through the covariates A_t. If A_t represents certain covariates that may contain unpriced factors, then the estimated factors $\hat{f}_{t+1}$ will not generate a mean-variance efficient tangency portfolio. To that end, Daniel et al. (2020) propose a hedging method for removing the unpriced risk, which results in higher Sharpe ratios. With no theoretical guidance on characteristic that represent the priced factors only, it is safer to consider all the covariates in the construction of factors that guarantee the mean-variance efficiency. This is equivalent to constructing GLS factors rather than OLS factors. This is the approach taken by Kozak and Nagel (2023), which we describe below.

Observe that the implied variance of r_{t+1} in model (19) can be written as

$$\Sigma_t = Z_t \Phi_t Z_t' + U_t \Omega_t U_t' \tag{20}$$

where $Z_t \Phi_t Z_t'$ captures the systematic component coming from $A_t f_{t+1} = Z_t f_{t+1}$ under the additional assumption that f_{t+1}' are random with covariance matrix Φ_t, and $U_t \Omega_t U_t'$ the idiosyncratic component coming from $B_t g_{t+1} + u_{t+1}$, that is, $B_t \Sigma_{gt} B_t + \sigma^2 I_N$. Kozak and Nagel (2023) also state, as observed earlier, that in the special case of $A_t' U_t = 0$ the OLS factors from model (19) are mean-variance efficient. The equivalence of OLS and GLS estimators of f_{t+1} in model (19) has been studied by Rao (1967). To prove the equivalence, we need to determine the structure of Σ_t, that satisfies (leaving the subscript 't' out)

$$(A'\Sigma^{-1}A)^{-1}A'\Sigma_{rr}^{-1} = (A'A)^{-1}A' \tag{21}$$

Lemma 5a in Rao (1967) states that if B is of rank $r = (N - rank(Z))$ such that $B'A = 0$, then the set of Σ_{rr} matrices are of the form,

$$\Sigma_{rr} = A\Phi A' + BVB' + \sigma^2 I_N \tag{22}$$

as stated in Kozak and Nagel (2023).

Kozak and Nagel (2023) point out that such factors constructed taking into account the covariances are (conditionally) mean-variance efficient as long as stock risk premia are linear functions of stock characteristics Z_t. Under the structure of model (12), without the structure implied in (13), Lemma 1 in Kozak and Nagel (2023) is satisfied by simply defining the L-dimensional price of risk vector as $b_t = (Z_t'\Gamma_t Z_t)\beta_{t+1}$. Thus, as long as the covariance of stock returns r_{t+1} and the factors β_{t+1} is proportional to the stock risk premia, the usual minimal requirement for any factor model to hold, the tangency portfolio formed by the factors $\tilde{\beta}_{t+1}^{GLS}$ achieves the maximal squared Sharpe ratio $\mu_t'\Sigma_t^{-1}\mu_t$ spanned by the SDF m_{t+1} in the span of r_{t+1}. If we further assume, as in the IPCA setup, that $\Gamma_t = (1/\sigma^2)I_N$ equation (14) defines the mean-variance inefficient distribution of the OLS factors, also know as the BARRA factors (after Bar Rosenberg, founder of Barra Inc. that first used them) and recently used in Fama and French (2020).

In our application we follow these insights and estimate Σ_t using industry clustering as Daniel et al. (2020) argue that industry exposure is a source of unpriced risk. Our results show how directly modeling the impact of unpriced factors through the residual covariance matrix is more effective than indirectly adjusting OLS factors via the hedged portfolios designed by Daniel et al. (2020) and refined by Kozak and Nagel (2023).

3 Estimation

For the ease of exposition, let us assume that the error term, ϵ_{t+1} has a time-invariant covariance matrix ($\Sigma_t = \Sigma = \Gamma^{-1}$). Now let $\underset{NT\times 1}{r} = [r_2, \cdots, r_{T+1}]'$ and $\underset{NT\times TL}{\bar{Z}} = Diag(Z_1, \cdots, Z_T)$ be the data matrices. Stacking the regression related parameters with the constraint (13) in the vector form, denote $\underset{TL\times 1}{\beta_{vec}}(\theta) = (1_T \otimes I_L)\Gamma_\alpha + (F' \otimes I_T)vec(\Gamma_\beta)'$, and the errors stacked as $\underset{NT\times 1}{e} = [\varepsilon_2, \cdots, \varepsilon_{T+1}]'$. Then the RRR model in (12) can be more compactly written as

$$r = \bar{Z} \cdot \beta_{vec}(\theta) + e, \tag{23}$$

and the cross-sectional contemporaneous correlation assumption can be expressed as

$$cov(e) = I_T \otimes \Sigma. \tag{24}$$

Then RRR parameters can be arranged in the following, $s = L + K(T + L)$ dimensional vector as, $\theta' = [\Gamma'_\alpha, vec(F)', vec(\Gamma'_\beta)']$ and can be found by minimizing the GLS criterion

$$\begin{aligned} S(\theta) &= \frac{1}{2T} \cdot (r - \bar{Z} \cdot \beta_{vec}(\theta))' \cdot (I_T \otimes \Gamma) \cdot (r - \bar{Z} \cdot \beta_{vec}(\theta)) \\ &\propto \frac{1}{2T} \sum_{t=2}^{T+1} (\tilde{\beta}^{GLS}_{t+1} - \Gamma_\alpha - \Gamma_\beta f_{t+1})' \cdot (Z'_t \Gamma_t Z_t) \cdot (\tilde{\beta}^{GLS}_{t+1} - \Gamma_\alpha - \Gamma_\beta f_{t+1}). \end{aligned} \tag{25}$$

where $\tilde{\beta}^{GLS}_{t+1}$ is as given in (14). Observe that in the estimation of model parameters, Γ_β, f_{t+1} cannot be uniquely determined as $(\Gamma_\beta V)(V^{-1} f_{t+1})$ for any non-singular L-dimensional V matrix holds, but the product $\Gamma_\beta f_{t+1}$ is uniquely determined. There are several methods to identify uniquely the parameters, Γ_α, Γ_β and $F = [f_2, ..., f_{t+1}]$, we follow Kelly et al. (2019), and impose the following constraints:

$$\Gamma'_\alpha \Gamma_\beta = 0, \qquad \Gamma'_\beta \Gamma_\beta = I_K, \qquad FF' = diag(\lambda_1, ..., \lambda_K) > 0. \tag{26}$$

These constraints restrict the information content coming from the vector Γ_α to be orthogonal to the information coming from the factors. The norm of the loading matrix Γ_β is set to be unity for better interpretation, and the time series of extracted factors are taken to be independent of each other and independent over time.

The values of θ that minimize the GLS criterion satisfy the following first-order conditions[3], for all t,

$$\hat{f}_{t+1} = [\hat{\Gamma}_\beta{}'(Z_t'\Gamma_t Z_t)\hat{\Gamma}_\beta]^{-1}\hat{\Gamma}_\beta{}'(Z_t'\Gamma_t Z_t)\cdot(\tilde{\beta}_{t+1}^{GLS} - \hat{\Gamma}_\alpha)$$
$$vec(\hat{\Gamma}_\beta') = \left[\sum_{t=1}^{T}(Z_t'\Gamma_t Z_t \otimes \hat{f}_{t+1}\hat{f}_{t+1}')\right]^{-1}\cdot\left[\sum_{t=1}^{T}\left(Z_t'\Gamma_t Z_t \otimes \hat{f}_{t+1}\right)(\tilde{\beta}_{t+1}^{GLS} - \hat{\Gamma}_\alpha)\right] \quad (27)$$
$$\hat{\Gamma}_\alpha = \left[\sum_{t=1}^{T}(Z_t'\Gamma_t Z_t)\right]^{-1}\cdot\left[\sum_{t=1}^{T}(Z_t'\Gamma_t Z_t)(\tilde{\beta}_{t+1}^{GLS} - \hat{\Gamma}_\beta\hat{f}_{t+1})\right].$$

These estimates are then normalized to satisfy the identification constraints stated in (26). This is achieved by re-scaling $\hat{\Gamma}_\beta$ and $\hat{F}$ via terms extracted from a Cholesky decomposition on $\hat{\Gamma}_\beta'\hat{\Gamma}_\beta$ and a singular value decomposition on $\hat{F}\hat{F}'$, and re-defining the L vector of intercepts as the residuals from regressing $\hat{\Gamma}_\alpha$ on the adjusted version of $\hat{\Gamma}_\beta$.[4] Because the estimates in (27) are functions of each other, the partial least squares (PLS) method is used to iteratively arrive at a solution. For the PLS method to converge, it is essential to have good starting values for the estimates. Because of the form of structure imposed on $\beta_{t+1} = \Gamma_\alpha + \Gamma_\beta f_{t+1}$ we suggest selecting the following initial values; Let $\bar{\beta}^{GLS} = \frac{1}{T}\sum_t \tilde{\beta}_{t+1}^{GLS}$; decompose $\tilde{\beta}_{t+1}^{GLS} = \bar{\beta}^{GLS} + (\tilde{\beta}_{t+1}^{GLS} - \bar{\beta}^{GLS}) = \bar{\beta}^{GLS} + \hat{\gamma}_{t+1}$. Start with $\hat{\Gamma}_\alpha = \bar{\beta}^{GLS}$; observe that $\hat{\gamma}_{t+1} = \hat{\Gamma}_\beta\hat{f}_{t+1}$ and thus the matrix stacking up the estimates $\hat{\gamma}_{t+1}$, $\underset{L\times T}{\hat{\beta}_\alpha} = [\hat{\gamma}_2, ..., \hat{\gamma}_{T+1}]'$, is of reduced-rank K. If $\underset{L\times K}{\hat{V}} = [\hat{V}_1, ..., \hat{V}_K]$ are the eigenvectors corresponding to the first K eigenvalues of $\hat{\beta}_\alpha'\hat{\beta}_\alpha$ (or the singular values of $\hat{\beta}_\alpha'$), set $\hat{\Gamma}_\beta = \hat{V}$ and $\hat{F} = \hat{V}'\hat{\beta}_\alpha'$. This procedure works reasonably well in practice.

To show for any t, that $\tilde{\beta}_{t+1}^{GLS}$ is more efficient than its OLS analog, we follow the standard approach given in the literature. Observe that $\tilde{\beta}_{vec}^{GLS} = (\bar{Z}'(I_T\otimes\Gamma)\bar{Z})^{-1}\bar{Z}'(I_T\otimes\Gamma)r \sim N(\beta_{vec}, (\bar{Z}'(I_T\otimes\Gamma)\bar{Z})^{-1})$ and $\tilde{\beta}_{vec}^{OLS} = (\bar{Z}'\bar{Z})^{-1}\bar{Z}'r \sim N(\beta_{vec}, (\bar{Z}'\bar{Z})^{-1}\bar{Z}'(I_T\otimes\Gamma)\bar{Z}(\bar{Z}'\bar{Z})^{-1})$. Define $X = \left(\bar{Z}'\bar{Z}\right)^{-1}\bar{Z}' - \left[\bar{Z}'(I_T\otimes\Gamma)\bar{Z}\right]^{-1}\bar{Z}'(I_T\otimes\Gamma)$. Note $X\bar{Z} = 0$ and $X(I_T\otimes\Gamma^{-1})X' = (\bar{Z}'\bar{Z})^{-1}\bar{Z}'(I_T\otimes\Gamma)\bar{Z}(\bar{Z}'\bar{Z})^{-1} - \left[\bar{Z}'(I_T\otimes\Gamma)\bar{Z}\right]^{-1} \equiv Var(\tilde{\beta}_{vec}^{OLS}) - Var(\tilde{\beta}_{vec}^{GLS})$. Because $(I_T \otimes$

[3]By setting $\Gamma_t = \frac{1}{\sigma^2}I_{N_{t+1}}$, these are exactly the conditions for OLS as stated in Kelly et al. (2019)(section 2); we additionally provide an explicit formula for Γ_α.

[4]The sign of the columns of Γ_β and F are altered so that the means of the rows of F, representing the average excess factor returns, are non-negative.

Γ^{-1}) is positive definite, $X(I_T \otimes \Gamma^{-1})X' > 0$. Thus $Var(\tilde{\beta}_{vec}^{OLS}) > Var(\tilde{\beta}_{vec}^{GLS})$. Note that both $\tilde{\beta}_{vec}^{GLS}$ and $\tilde{\beta}_{vec}^{OLS}$ are unbiased with the same expected value β_{vec}, we can then see how the smaller covariance for $\tilde{\beta}_{vec}^{GLS}$ leads to a higher Sharpe ratio.

By the same logic, in the context of model (12) and (13) when Γ_α and Γ_β are known, we can regress $r_{t+1}^{**} \equiv r_{t+1} - Z_t\Gamma_\alpha$ on $Z_t\Gamma_\beta$ and conclude that $Var(\hat{f}_{t+1}|\Gamma_\alpha, \Gamma_\beta) = \left[\Gamma_\beta'(Z_t'\Gamma_t Z_t)\Gamma_\beta\right]^{-1}$ is lower than $Var(\hat{f}_{t+1}^{IPCA}|\Gamma_\alpha, \Gamma_\beta) = \sigma^2\left[\Gamma_\beta' Z_t' Z_t \Gamma_\beta\right]^{-1}$. Therefore, the RRR factors are statistically and economically more efficient than IPCA factors in theory. Our findings builds on those of Kozak and Nagel (2023) as we demonstrated that the GLS and RRR factors, $\tilde{\beta}_{vec}^{GLS}$ and $\hat{f}_{t+1}$ respectively, not only achieve a higher mean-variance spanning but are also more efficient estimators.

3.1 Economic interpretation of the estimates

Observe that the RRR restriction $\beta_{t+1} = \Gamma_\alpha + \Gamma_\beta f_{t+1}$ decomposes the L-dimensional coefficients from the regression model (12) into two parts, the time-invariant latent factor Γ_α and the dynamic component $\Gamma_\beta f_{t+1}$. Indeed the GLS criterion (25) defines $\hat{\Gamma}_\beta \hat{f}_{t+1}$ as the closest K-dimensional approximation to the L-dimensional adjusted GLS coefficients, $\tilde{\beta}_{t+1}^{GLS} - \hat{\Gamma}_\alpha$. The discussion below, we believe can help better interpret the components when estimated from empirical data.

Intercept $\hat{\Gamma}_\alpha$: Observe that $(Z_t'\Gamma_t Z_t)(\tilde{\beta}_{t+1}^{GLS} - \hat{\Gamma}_\beta \hat{f}_{t+1}) = Z_t'\Gamma_t\left(r_{t+1} - Z_t\hat{\Gamma}_\beta \hat{f}_{t+1}\right)$; hence from the first order conditions (27) we see how $\hat{\Gamma}_\alpha$ is a weighted average of residuals from model (12) without the term Γ_α. A matrix algebra result from Fujikoshi (1974) explicitly relates the unconstrained matrix of GLS factors $\tilde{\beta}^{GLS} = [\tilde{\beta}_2^{GLS}, ..., \tilde{\beta}_{T+1}^{GLS}]'$ to the constrained matrix β defined in equation (15) as

$$\hat{\beta} = \frac{1_T 1_T'}{T}\tilde{\beta}^{GLS} + \left(I_T - \frac{1_T 1_T'}{T}\right)\tilde{\beta}^{GLS} V_K V_K' \tag{28}$$

where V_K consists of K eigenvectors that correspond to the K largest eigenvalues of $D =$

$(\tilde{\beta}^{GLS} - \hat{\beta})(\tilde{\beta}^{GLS} - \hat{\beta})'$. [5] Inspecting the structure of $\hat{\beta}$ from (28) allows to identify $\frac{1_T' \tilde{\beta}^{GLS}}{T}$ as a proxy for Γ_α'. That is, Γ_α can be thought of as the temporal average of the GLS factors, which was defined as $\bar{\beta}^{GLS}$ in the previous section. Thus equation (28) formally justifies the suggestion to assign $\bar{\beta}^{GLS}$ as the initial value for the iterative procedure, and provides a simple interpretation for Γ_α.

Factors $\hat{\mathbf{f}}_{\mathbf{t+1}}$: From the first order conditions (27) we also see that the estimated RRR factors $\hat{f}_{t+1}$ are weighted exposures of the GLS factors, $\tilde{\beta}^{GLS}_{t+1}$ in excess of their approximate mean $\hat{\Gamma}_\alpha$. As a matter of fact, we can interpret $\hat{f}_{t+1}$ as capturing the K most important innovations in the L-dimensional GLS coefficients vector $\tilde{\beta}^{GLS}_{t+1}$. More importantly, under the usual assumption that stock risk prima are proportional to the covariance between the stock excess returns r_{t+1} and the K-dimensional factors f_{t+1}, the Sharpe ratio of the mean-variance portfolio that uses $\hat{f}_{t+1}$ when $\Gamma_\alpha = 0$ equals the maximum squared conditional Sharpe ratio $\mu_t' \Sigma_t^{-1} \mu_t$.[6] Therefore, when $\Gamma_\alpha = 0$ the RRR factors are conditionally mean-variance efficient. Moreover, it follows that when $\Gamma_\alpha \neq 0$ the RRR factors are not efficient so that a nonzero estimated mean $\hat{\Gamma}_\alpha$ for the GLS factors signals mispricing. Also notice that when $K = L$ the first order conditions in (27) show that the RRR factors $\hat{f}_{t+1}$ are proportional to the GLS factors and at the same time Γ_α must be zero. This implies that also for $K = L$ the RRR factors are mean-variance efficient (achieving the exact same spanning as the GLS factors) and there is no mispricing.[7] Therefore our setup nicely relates the mean-variance efficiency of the $K \leq L$ dimensional vector of extracted RRR factors to the absence of mispricing.

[5] For the proof, see Fujikoshi (1974) (Lemma 1 and Lemma 2) with $L \times L$ positive matrix Q set to identity.

[6] This can be easily seen by defining $W_t = \Gamma_t Z_t \Gamma_\beta (\Gamma_\beta' Z_t' \Gamma_t Z_t \Gamma_\beta)^{-1}$ and $\Sigma_t^{-1} = \Gamma_t$ in Lemma 1 of Kozak and Nagel (2023) and defining the K-dimensional price of risk vector as $b_t = (\Gamma_\beta' Z_t' \Gamma_t Z_t \Gamma_\beta)\beta_{t+1}$.

[7] To see why $\Gamma_\alpha = 0$ when $K = L$, we notice that the only viable singular value decomposition of β is of the form UDV' where U and V are full-rank orthogonal matrices and D is a rectangular diagonal matrix. Then defining $F = DV'$ and $\Gamma_\beta = U$, we can observe that $(\beta - \Gamma_\beta F) = 0$. Such decomposition leads to $\beta^{GLS}_{t+1} = \Gamma_\beta f_{t+1}$ for all t, where f_{t+1} is the t-th column of the F matrix. To see why the RRR factors f_{t+1} and the GLS factors $\tilde{\beta}^{GLS}_{t+1}$ share the same mean-variance spanning when $K = L$, define w_t and q_t to be the L-dimensional row weight vectors that yield the highest Sharpe ratio for $\tilde{\beta}^{GLS}_{t+1}$ and f_{t+1} respectively. Since $\tilde{\beta}^{GLS}_{t+1} = \Gamma_\beta f_{t+1}$ it follows that $SR(w_t \tilde{\beta}^{GLS}_{t+1}) = SR(w_t \Gamma_\beta f_{t+1})$. Because q_t yields the highest Sharpe ratio for f_{t+1} then $SR(w_t \Gamma_\beta f_{t+1}) \leq SR(q_t f_{t+1})$ and therefore $SR(w_t \tilde{\beta}^{GLS}_{t+1}) \leq SR(q_t f_{t+1})$. On the other hand, since $\Gamma_\beta' \Gamma_\beta = I$ then $\Gamma_\beta' \tilde{\beta}^{GLS}_{t+1} = f_{t+1}$. It then follows that $SR(w_t \tilde{\beta}^{GLS}_{t+1}) \geq SR(q_t \Gamma_\beta' \tilde{\beta}^{GLS}_{t+1}) = SR(q_t f_{t+1})$. So $SR(w_t \tilde{\beta}^{GLS}_{t+1}) = SR(q_t f_{t+1})$.

Binding Constraint: $\hat{\mathbf{\Gamma}}_\beta$: The matrix $\hat{\Gamma}_\beta$ serves to map a large number of L stock characteristics to a small number of K risk factors. Constraint (13) suggests that Γ_β can also be interpreted as the loading matrix of a static factor model of the factors: a model that relates the data generating process for the L-dimensional vector of GLS factors $\tilde{\beta}_{t+1}$ to that of the K-dimensional vector of RRR factors f_{t+1}. The estimation of $\hat{\Gamma}_\beta$ involves identifying the time-invariant structure of a few linear combinations of candidate characteristics that are most effective at describing the latent factor loading structure in f_{t+1}.

4 Asymptotic Distributions of Estimators

In the financial panel data in general and in the data that we analyze, $N > T$. We use the asymptotic results presented in Silvey (1959) and Reinsel et al. (2022). With the same available data, (r_{it}, Z'_{it}) as here, Reinsel et al. (2022) in chapter 7 provide the asymptotics for a time-series version of the RRR model where the cross-sectional regression (12) is implemented for a given unit i over time (instead of for all units at a given point in time t). With some regularly conditions added, the results given there can be adapted here as well. We assume that N is fixed but $T \to \infty$. The results follow from the distribution of $\frac{\delta S(\theta)}{\delta \theta}$, the vector of first partial derivatives (27) of the criterion function defined in equation (25), which can be more compactly written as

$$\frac{\delta S(\theta)}{\delta \theta} = -\frac{1}{T} \cdot M' \bar{Z}' (I_T \otimes \Gamma) e \tag{29}$$

where $M = [(1_T \otimes I_L), (I_T \otimes \Gamma_\beta), (F' \otimes I_L)]$. Thus the asymptotic distribution depends on the distribution of $\frac{1}{\sqrt{T}} \bar{Z}'(I_T \otimes \Gamma)e$ which is normal with mean zero and variance-covariance matrix $V^* = \bar{Z}'(I_T \otimes \Gamma)\bar{Z}$.

Result 1: Given $\Gamma = \Sigma^{-1}$, as $T \to \infty$, $\hat{\theta} \to \theta$ almost surely and $\sqrt{T}(\hat{\theta} - \theta)$ has a limiting multivariate normal distribution with a mean zero vector and a singular covariance matrix

W^* defined as

$$W^* = (B_\theta + H_\theta H_\theta')^{-1} B_\theta (B_\theta + H_\theta H_\theta')^{-1} \tag{30}$$

where $B_\theta = \lim_{T\to\infty} \frac{\delta^2 S_T(\theta)}{\delta\theta\delta\theta'} = M'V^*M$, $h(\theta)$ is a $K(K+1)$ column vector of the normalization conditions in (26) stacked without duplication and $H_\theta = \{\partial h_j(\theta)/\partial\theta_i\}$ is a matrix containing the partial derivatives of $h(\theta) = 0$.

More details on the matrices B_θ and H_θ are given in the Appendix B. The singularity of B_θ arises mainly due to imposing the identification constraints (26) and it is removed by adding the $H_\theta H_\theta'$ matrix. Yet observe that W^* is a singular matrix. The asymptotic theory in Result 1 applies to the IPCA model as well.[8] Kelly et al. (2019) base their inference using bootstrapping samples. Our closed-form limiting distribution allows us to provide diagnostics for the inferences.

4.1 Testing for Γ_α

Kelly et al. (2019) consider a large sample test for the no mispricing hypothesis, H_0 : $\Gamma_\alpha = 0$. Recall testing this hypothesis is equivalent to assessing whether the instruments Z_t capture the variation in average returns that may be not correlated to factor exposures. Specifically, it is a test to infer whether the lower dimensional approximation of the mean of the GLS factors β_{t+1} is exactly zero. This condition is economically important as it rules out mispricing coming from the stock characteristics Z_t and is equivalent to categorizing all the variation present in stock characteristics as risk exposure. Kelly et al. (2019) construct a test for H_0 using the following Wald-type test statistics, $W_\alpha = \hat{\Gamma}_\alpha' \hat{\Gamma}_\alpha$ and implement the test via bootstrapping samples. The bootstrapping samples are not drawn from the basic models (12) and (13) but based on the derived model

$$X_{t+1} = Z_t' r_{t+1} = (Z_t' Z_t)\Gamma_\alpha + (Z_t' Z_t)\Gamma_\beta f_{t+1} + Z_t' \epsilon_{t+1}. \tag{31}$$

[8] Set $\Gamma = \frac{1}{\sigma^2} I_N$ and an estimate of σ^2 can be obtained by pooling the residuals.

Observe that X_{t+1} is the return from a characteristic-based portfolio and the resampling is based on the model for X_{t+1} under H_0 and not under the original model for r_{t+1}. Note that in practice the number of stocks available in a given time period t is time-varying (as some firms get delisted, others get listed and/or merged etc) resulting in an unbalanced panel. The portfolio formulation in (31) avoids dealing with this complication; however it induces a time varying residual covariance matrix function of the stock characteristics Z_t which differs from the structure implied by the IPCA model for the stock returns' residuals. That is, $Var(Z_t'\varepsilon_{t+1}) = \sigma^2 Z_t'Z_t$ may not be an identity matrix with a scalar multiplier. The bootstrap sampling of $Z_t'\varepsilon_{t+1}$ from a t-distribution as in Kelly et al. (2019) accounts for fat tails but not for this type of induced heteroscedasticity ($Z_t'Z_t$). This bias has serious implications as illustrated in the top graph of Figure 1.

The figure compares the bootstrapped distribution of the test statistic proposed in Kelly et al. (2019) (blue line) for the null of $\Gamma_\alpha = 0$ in an IPCA model with $K = 6$ versus the distribution for the same statistic implied by the asymptotic theory (black line) before (top graph) and after (middle graph) the correction for the bias.[9] The vertical red line reports the critical value for the test.[10] Notice that according to the test proposed in Kelly et al. (2019) we would not reject the null (with a p-value of 0.505), however, based on the asymptotic distribution from Result 1, it is rejected at virtually any level of confidence. Moreover, as we later show in the empirical section,[11] the difference in the annualized maximal Sharpe ratios due to the presence of a non-zero Γ_α for the IPCA model with $K = 6$ is approximately 1 and it is statistically significant at any conventional level. In contrast, the bottom graph of Figure 1 plots the same quantities as the top graph but under our bootstrap design where we simulate the error term $Z_t'\varepsilon_{t+1}$ from a normal distribution with mean zero and covariance $\sigma^2 Z_t'Z_t$. Thus the bootstrapping results depend upon the valid assumption made on the

[9]We choose to show the case where $K = 6$ as this is the first instance in Kelly et al. (2019) where they cannot reject the null of $\Gamma_\alpha = 0$ at any conventional level.

[10]More details available in Section 4.2.

[11]See the bottom graph of Figure 5, the star on the differential maximal Sharpe ratio estimate for the IPCA model with $K = 6$ represents statistical significance at least at the 5% level using the tests developed in Barillas et al. (2020).

error term that depends the time-varying stock characteristics Z_t.

4.2 Formal test for mispricing

We can exploit the asymptotic distribution in Result 1 to derive the distribution for W_α, the statistic proposed in Kelly et al. (2022), and use it for testing $H_0 : \Gamma_\alpha = 0$. In particular, observe $\hat{\Gamma}_\alpha \sim N(\Gamma_\alpha, W^*_\alpha)$ where W^*_α represents the first L rows and L columns of W^*, obtained by partitioning the matrix of W^* defined in equation (30). From Theorem 2.1 and 3.1 in Box (1954) it follows that under the null $\Gamma_\alpha = 0$, the test statistic $\hat{W}_\alpha$ can be written as a mixture of Chi-squared distributions, which can be well-approximated by a scaled central Chi-squared distribution. Thus

$$\hat{W}_\alpha \equiv \hat{\Gamma}'_\alpha \hat{\Gamma}_\alpha \sim \sum_{l=1}^{L} \lambda_l \chi^2(1) \approx g\chi^2(h) \tag{32}$$

where λ_l is the l-th eigenvalue of W^*_α, $g = \frac{K_2}{2K_1}, h = \frac{2K_1^2}{K_2}$ and $K_1 = \sum_{l=1}^{L} \lambda_l, K_2 = 2\sum_{l=1}^{L} \lambda_l^2$ and hence $\frac{\hat{W}_\alpha}{g} \approx \chi^2(h)$. Thus this procedure makes use of the overall empirical estimates more directly.

Alternatively we can use Result 1 to directly obtain the Wald statistic and its distribution:

$$WS = \hat{\Gamma}'_\alpha \left(W^*_\alpha\right)^{-1} \hat{\Gamma}_\alpha \sim \chi^2_L \tag{33}$$

This statistic is similar to the large sample test employed in Gibbons et al. (1989) to test for the joint significance of the time-invariant regression intercepts in a multivariate regression setup. In Figure 2, we compare the power function of the two tests. It is clear that the Wald statistic in (33) has better power, but the suggested approximation to W_α also fares well. The improvement resulting from the approximation can be seen from the bottom graph of Figure 1, of the distributions for the previously discussed test for an IPCA model with $K = 6$ (after we correctly account for the bias present in the IPCA bootstrapping setup).

To sum up, even if the inference based on the bootstrapped distribution for W_α with the suggested modification appears to be fairly accurate, it remains much more computationally intensive than that based on the asymptotic distribution given in the Wald Test; the test still exhibits less power, and it may underestimate the significance of Γ_α.[12] Thus we suggest using the asymptotic Wald test (33). An alternative informal procedure to test $H_0 : \Gamma_\alpha = 0$ in model (12) can be suggested. The difference between the model with and without Γ_α is in how the second term, $\Gamma_\beta f_{t+1}$, is extracted. Recall our earlier discussion that when Γ_α is present, it can be estimated by $\hat{\Gamma}_\alpha = \bar{\beta}^{GLS}$, where $\bar{\beta}^{GLS}$ is the temporal average of the regression coefficients $\tilde{\beta}_{t+1}^{GLS}$. Therefore, we could simply compare the singular values of $\tilde{\beta}^{GLS} = [\tilde{\beta}_2^{GLS}, ..., \tilde{\beta}_{T+1}^{GLS}]'$ with the singular values of $\tilde{\beta}_\alpha^{GLS} = [(\tilde{\beta}_2^{GLS} - \bar{\beta}^{GLS}), ..., (\tilde{\beta}_{T+1}^{GLS} - \bar{\beta}^{GLS})]'$. If $H_0 : \Gamma_\alpha = 0$ holds, we should expect the two sets of singular values to be the same.

Two other tests that are central to the IPCA or RRR model involve testing for observable factors beyond the instruments and testing for the significance of the instruments themselves. The former can be tested via the RRR model setup in (12) and (13); adding or deleting certain factors involves merely adjusting the returns and the instruments for the partial effect of those factors. Such an extended model in the classic RRR setup is covered in (Reinsel et al., 2022, Ch3). For testing the significance of the instruments, Kelly et al. (2019) follow the cumbersome procedure that compares the models with and without a particular instrument again using the computationally intensive bootstrap procedure. With the asymptotic results given in this paper, it is easy to carry out this test for partial impact of an instrument or a select set of instruments. We explore the sparseness of the Γ_β matrix more directly in the

[12] An additional comparison between the two tests suggests the Wald-type test based on $\hat{W}_\alpha$ may underestimate the true significance of Γ_α. The asymptotic variable of $\hat{W}_\alpha$ using the delta method and Slustky's theorem, is given as $4(\hat{\Gamma}_\alpha' W_\alpha^* \hat{\Gamma}_\alpha)$. We can contrast the standardized quantity, $\hat{W}_\alpha^2/4(\hat{\Gamma}_\alpha' W_\alpha^* \hat{\Gamma}_\alpha)$ with the WS statistic given in (33). A mathematical relationship from the inequality below (see (Marshall et al., 1979, p. 659-660)) is useful in this regard:
Lyapunov's Inequality: If W_α^* is a positive definite matrix, x is a unit vector, $||x|| = 1$, then for $a \geq b \geq c$,

$$(x' W_\alpha^{*b} x)^{a-c} \leq (x' W_\alpha^{*c} x)^{a-b} \cdot (x' W_\alpha^{*a} x)^{b-c}$$

This follows from the fact that $(x' W_\alpha^{*r} x)$ can be expressed as the r^{th} moment of a distribution of the eigenvalues of W_α^*. Note if we set $a = 1$, $b = 0$ and $c = -1$, (12) reduces to $(x'x)^2 \leq (x' W_\alpha^{*-1} x)(x' W_\alpha^* x)$ and by setting $x = \frac{\hat{\Gamma}_\alpha}{\sqrt{\hat{\Gamma}_\alpha' \hat{\Gamma}_\alpha}}$, we obtain $\frac{(\hat{\Gamma}_\alpha' \hat{\Gamma}_\alpha)^2}{4(\hat{\Gamma}_\alpha' \hat{W}_\alpha^* \hat{\Gamma}_\alpha)} \leq \frac{(\hat{\Gamma}_\alpha' \hat{W}_\alpha^{*-1} \hat{\Gamma}_\alpha)}{4}$. Thus the standardized test statistic $\frac{(\hat{\Gamma}_\alpha' \hat{\Gamma}_\alpha)^2}{4(\hat{\Gamma}_\alpha' W_\alpha^* \hat{\Gamma}_\alpha)}$ will grossly underestimate the significance of Γ_α.

next section.

5 Sparseness analysis: Which characteristics matter?

The focus of the RRR setup is mainly on the dimension reduction via a low rank structure. The resulting linear combinations $Z_t'\Gamma_\beta$ still require keeping all the asset characteristics (Z_t) in the model. It does not result in any specific variable selection and therefore no asset characteristic can be discarded. It is possible that many of the entries in Γ_β may not be significant. While the RRR exploits the simultaneous dependence among the asset characteristics and returns to achieve parsimony in dimension reduction, the sparseness helps in the selection of key variables in this dependence. We use a sparseness method where both dimension reduction and variable selection are possible. The integration of RRR and the sparsity structure can greatly advance the interpretation of empirical results (Reinsel et al., 2022, Ch. 13). Specifically, imposing Γ_β to be sparse would allow us to isolate the key characteristics behind the extracted factors and thus better understand their significance. The spirit of this analysis is to address Cochrane (2011)'s critical question, 'Which characteristics do really matter?'.

The procedure consists of two steps: rank determination first and then the rank-constrained sparse estimation of the component matrices. As given in Bunea et al. (2012), this approach provides strong theoretical and computational guarantees and is superior to the approach where the two steps are reversed. Specifically, we modify the GLS criterion (25) by incorporating a weighted Lasso penalty term as follows:

$$\begin{aligned} S(\theta) =& \frac{1}{2T}\sum_{t=2}^{T+1}[r_{t+1} - Z_t(\Gamma_\alpha + \Gamma_\beta f_{t+1})]'\Gamma_t[r_{t+1} - Z_t(\Gamma_\alpha + \Gamma_\beta f_{t+1})] + \delta \left\|\Gamma_\beta \Lambda\right\|_1 \\ &\propto \sum_{t=2}^{T+1}(\tilde{\beta}_{t+1}^{GLS} - \Gamma_\alpha - \Gamma_\beta f_{t+1})' Z_t' \Gamma_t Z_t (\tilde{\beta}_{t+1}^{GLS} - \Gamma_\alpha - \Gamma_\beta f_{t+1}) + \delta \|\Gamma_\beta \Lambda\|_1. \qquad (34) \\ \text{s.t.} \qquad & \Gamma_\beta' \Gamma_\beta = I_K, FF' = diag(\lambda_1, \cdots, \lambda_K) = \Lambda \end{aligned}$$

Directly penalizing Γ_β is not appropriate because the rescaled versions of Γ_β and F are obtained via singular value decomposition. Singular vectors corresponding to factors with larger variances, i.e, those with larger singular values (e.g. λ_1), can be estimated more precisely and should therefore contribute more to the regularization, whereas those corresponding to smaller singular values are not easily identifiable and should not influence the regularization. Therefore, we introduce the weighted LASSO penalty term, $\Gamma_\beta\Lambda$.

The solution to (34) in its present form is hard to obtain because of the time-varying weight matrix $Z_t'\Gamma_t Z_t$. To get the sparseness results, we simply replace $Z_t'\Gamma_t Z_t$ by $W \equiv \frac{1}{T}\sum_{t=2}^{T+1} Z_t'\Gamma_t Z_t$, then the sparseness target function (34) can be written as,

$$\begin{aligned}
&\sum_{t=2}^{T+1}(W^{1/2}(\tilde{\beta}_{t+1}^{GLS}-\hat{\Gamma}_\alpha)-W^{1/2}\Gamma_\beta f_{t+1})'(W^{1/2}(\tilde{\beta}_{t+1}^{GLS}-\hat{\Gamma}_\alpha)-W^{1/2}\Gamma_\beta f_{t+1})+\delta\|\Gamma_\beta\Lambda\|_1 \\
=&\|W^{1/2}(\tilde{\beta}^{GLS'}-\hat{\Gamma}_\alpha 1_T')-W^{1/2}\Gamma_\beta F\|_F^2+\delta\|\Gamma_\beta\Lambda\|_1 = S_W(\theta) \\
&\text{s.t.} \qquad \Gamma_\beta'\Gamma_\beta = I_K, FF' = diag(\lambda_1,\cdots,\lambda_K)=\Lambda
\end{aligned} \tag{35}$$

Problem (35) fits the sparse reduced-rank regression framework studied by Chen and Huang (2012). Their algorithm can determine the set of significant firm characteristics uniquely, but the weights assigned to the elements among the corresponding rows in Γ_β are all non-zero. This makes it challenging to determine which firm characteristics mainly drive specific subsets of factors. Hence, we define $\underset{L\times T}{Y} = W^{1/2}(\tilde{\beta}^{GLS'}-\hat{\Gamma}_\alpha 1_T')$ and $\underset{L\times L}{X} = W^{1/2}$, and apply the SOFAR (Sparse orthogonal factor regression) algorithm proposed by Uematsu et al. (2019) to estimate the sparse matrix Γ_β and the non-sparse matrix F containing the time series of factors. Uematsu et al. (2019) suggest a two-step approach to obtain the SOFAR estimator. In the first step, they minimize a L_1-penalized squared loss for $Y = XC$ where $C = \Gamma_\beta F$ to obtain an initial estimator. Because it is theoretically guaranteed that this initial estimator is not far away from the true coefficient matrix C^*, in the second step, they minimize the objective function (35) in an asymptotically shrinking neighborhood of the initial estimator. The details of the algorithm are described in Appendix C. The SOFAR estimator not only can help us to identify the subsets of significant firm characteristics, but also help us to understand which aspects of firm characteristics information mainly drive the estimated latent factors. Our empirical sparseness analysis results, presented in Section 7.4, provide valuable economic insights concerning the structure of the RRR GLS factors.

6 Empirical Design

6.1 Data

We use the same data that was studied in Kelly et al. (2019)[13] but eliminate stocks with less then 25 observations. This data is a subset of (36 out of 62 variables) data originally studied by Freyberger et al. (2020). The asset characteristics are broadly classified into past, returns(4), investment(3), profitability(10), intangibles(4), value(7), and trading frictions(8). The nature of these variables suggest that they are likely to be time-varying. The data is not balanced as the number of assets vary over the duration of the study, $N_t \in [233, 3680]$. Our analysis can accommodate different number of stocks available at different points in time (as new firms get listed, some existing firms get de-listed, dropped or merged etc..). To summarize, we have $N = \sum_{t=1}^{T} N_t = 9890$ different stocks for the duration of the study (July 1962 − May 2014). The descriptive statistics of the 36 variables are given in Table 1. In order to show the time-varying nature of the variables, we compute $\Delta Z_{i,t} = Z_{i,t} - Z_{i,t-1}$ and then the coefficient of variation, a standardized measure, which is the ratio of the standard deviation of $\Delta Z_{i,t}$ and its mean. If the characteristics are not time-varying, the coefficient of variation is expected to be close to zero. The values reported in Table 1 clearly indicate that most of the variables vary over time, some with higher volatility than others.

6.2 Performance metrics

In the empirical analysis we assess the performance of the models introduced in Section 2.1 through their mean-variance spanning abilities, measured by the maximal Sharpe ratio,

[13]For more details refer to section 4.1 of Kelly et al. (2019).

and the goodness of fit, measured via the BIC and R^2_{GLS} metrics.[14] These metrics broadly cover both the economic as well as the statistical spectrum.

Maximal Sharpe ratio: As standard in the literature, we evaluate the economic performance of our models by their ability to generate factors that can span the unconditional mean-variance frontier. We do so by examining the in-sample Sharpe ratio of the tangency portfolio formed by the factors generated by a given model. To justify the adoption of the unconditional spanning of the mean-variance frontier, as in Kozak and Nagel (2023) (Assumption 1 and Corollary 2) we assume that stock risk premia μ_t are linear in the firm characteristics Z_t, or a lower-dimensional linear combination, $Z_t\Gamma_\beta$. This assumption implies that factor premia are time invariant and thus they are the risk premia for the tangency portfolios as well (the numerators of the adopted in-sample maximal Sharpe ratios).[15] Following Barillas and Shanken (2017), we test the relative spanning abilities of any two given models by examining the difference between the in-sample squared Maximal Sharpe ratios using the tests in Barillas et al. (2020).[16]

In the RRR framework when $\Gamma_\alpha = 0$, the tangency portfolio is formed exclusively from the time series estimates of the latent factors f_{t+1}. In the presence of mispricing (i.e. when $\Gamma_\alpha \neq 0$) the latent factors are inefficient, therefore the tangency portfolio is constructed using the latent factors and the returns of an additional pure-alpha portfolio. This pure-alpha portfolio is a function of Γ_α, and is orthogonal to the extracted factors, and such that when added as an additional factor to the model, the time-varying intercept vanishes (for a related discussion of such orthogonal portfolios see section 6.6 of Campbell et al. (1998)). A suitable pure-alpha portfolio for the case of the RRR model (12) then is

$$f^h_{t+1} = \left[\Gamma'_\alpha (Z'_t \Gamma_t Z_t)^{-1} Z'_t \Gamma_t\right] r_{t+1} = \Gamma'_\alpha \tilde{\beta}^{GLS}_{t+1}. \tag{36}$$

[14]In Appendix D we also provide the analysis with respect to the AIC metric. The insights are the same.

[15]As already mentioned before and in contrast with Kozak and Nagel (2023), we do not need the linearity assumption to proof the mean-variance efficiency of the factors. We invoke it here solely to enhance the interpretation of the unconditional spanning.

[16]Barillas et al. (2020) provide statistical tests when factors are exogenous (or known), endogenous (or latent), nontraded and when the tested models have overlapping factors.

Note that f^h_{t+1} is a linear combination of the GLS regression coefficients $\tilde{\beta}^{GLS}_{t+1}$ which are factors themselves. When $\Gamma_t = \frac{1}{\sigma^2} I_{N_t}$, (36) reduces to the IPCA "arbitrage" portfolio described in section 4.5.3. of Kelly et al. (2019).

Goodness of fit: To evaluate the statistical performance of our analyzed models, we employ two metrics: the BIC and the R^2_{GLS}. While R^2_{GLS} is a standardized metric, BIC is a comparative metric that penalizes over-fitting, making it suitable for comparing different models using the same data. Results under the AIC metric, a popular alternative to BIC, although not discussed in the main text are presented in Appendix D.

BIC: The BIC metric is calculated based on the log-likelihood function of the analyzed model with a penalty term. Given a set of estimates for the vector of model parameters θ we can compute the log-likelihood as $lnL(\theta) \equiv \sum_t^T lnf(r_{t+1}|Z_t;\theta) = -\frac{1}{2}\left[ln(2\pi)\sum_t^T N_t + \sum_t^T ln|\Gamma_t^{-1}|\right] - TS(\theta)$, where $S(\theta)$ is defined in equation (25).[17] Then,

$$BIC = s \cdot ln(T) - 2lnL(\theta) \tag{37}$$

where s represents the total number of estimated parameters,[18] and $ln(T)$ is the natural logarithm of the total number of observations.

$\mathbf{R^2_{GLS}}$: A readily interpretable criterion as an alternative to BIC is the R^2_{GLS} goodness of fit measure:

$$R^2_{GLS} = 1 - \frac{\sum_t \left(\Gamma_t^{\frac{1}{2}} r_{t+1} - \Gamma_t^{\frac{1}{2}} Z_t(\hat{\Gamma}_\alpha + \hat{\Gamma}_\beta \hat{f}_{t+1})\right)' \left(\Gamma_t^{\frac{1}{2}} r_{t+1} - \Gamma_t^{\frac{1}{2}} Z_t(\hat{\Gamma}_\alpha + \hat{\Gamma}_\beta \hat{f}_{t+1})\right)}{\sum_t \left(\Gamma_t^{\frac{1}{2}} r_{t+1}\right)' \left(\Gamma_t^{\frac{1}{2}} r_{t+1}\right)}. \tag{38}$$

[17] For the case of the IPCA model it reduces to $lnL(\theta) = -\frac{1}{2}ln(2\pi\sigma^2)\sum_t^T N_t - TS(\theta)$ where σ^2 represents the average residual variance. For the fit of model (12) for GLS and OLS factors set $\Gamma_\alpha = 0$ and $K = T$ while defining Γ_β and f_{t+1}.

[18] $s = L + K(T+L) + (N + N_I(N_I-1)/2) - (K+K^2)$ for the RRR model with $\Gamma_\alpha \neq 0$ and N_I industries. The terms appearing with a minus sign are the degree of freedom adjustments accounting for the constraints in (26). When $\Gamma_\alpha = 0$ only K^2 is subtracted. The number of estimated parameters for the IPCA model is the same as that for the RRR model except that we need to replace the number of independent parameters in the error covariance matrix, $N + N_I(N_I-1)/2$ by unity, which accounts for σ^2. For the case of the GLS factors from model (12) $s = LT + N + N_I(N_I-1)/2$. Finally, for the case of OLS factors $s = LT + N + 1$.

Equation (38) extends the R^2_{OLS} used in Kelly et al. (2019) by weighing the returns r_{t+1} and characteristics Z_t with Γ_t, the inverse of the error-covariance matrix. Observe the equation (38) represents the ratio of re-scaled return variances explained by both the dynamic behavior of the estimated conditional loadings $Z_t\hat{\Gamma}_\beta$ (and estimated time-varying intercept $Z_t\hat{\Gamma}_\alpha$ in the unrestricted model), as well as by the contemporaneous factor realizations $\hat{f}_{t+1}$, aggregated over all assets and all time periods. Kandel and Stambaugh (1995) and Lewellen et al. (2010) argue that the R^2_{GLS} is a more meaningful statistic to assess the mean-variance spanning abilities of the factors provided by the regression model. Specifically, if the tangency portfolio formed by a model's factors is nearly mean-variance efficient, the R^2_{GLS} is close to one but the R^2_{OLS} does not enjoy such nice property because it is not appropriately standardized.

6.3 Specification and estimation of the error covariance matrix

As discussed in Section 3, if covariances of the returns in the error terms are used in the model, the GLS(RRR) factors achieve the best possible mean-variance spanning. However, estimating the covariance matrix, Σ_t in a framework like ours and providing a robust time series of estimates $\{\Gamma_t^{-1}\}_t$ of large and time-varying dimensions $N_t \times N_t$ with $N_t \in [233, 3680]$, entails a daunting task. For an in-depth survey of this topic, refer to Fan et al. (2016a).

A central contribution of this paper is to propose a viable procedure to estimate the covariances. Popular approaches generally impose some sort of structure to reduce the number of parameters to estimate. The commonality literature suggests PCA-based structures (e.g. Hasbrouck and Seppi (2001)), where the covariances are accounted by a few linear combinations of asset returns. Ledoit and Wolf (2017, 2020) provide shrinkage methods resulting in a covariance matrix that is a linear combination of a diagonal matrix and a rank one matrix. Engle and Kelly (2012) focus on covariance matrices with equi-correlated block structures. Recent developments include thresholding the principal components via shrinkage methods (see Fan et al. (2013)). These are computationally intensive procedures and to apply them in our analysis with time-varying sample sizes, it is even more difficult. So to keep the main focus on the estimation of GLS factors, we follow an intuitive approach. We first assume that the covariance matrix of the error term is not time-varying, $\Sigma_t = \Sigma$ and it is positive definite. We further assume that it has a block structure where each block belongs to a different industry. The number of industry types is solely determined by the empirical metrics introduced in Section 6.2. Our design with this approach can reduce the number of parameters to be estimated in Σ by more than 95%. Industry-grouping seems to be a natural way of clustering the universe of stocks. More importantly, this choice is supported by the argument provided in (Daniel et al., 2020, p1931) that industry exposures might represent a source of unpriced common variation in characteristics-based factor models.

In practice, as mentioned earlier, the estimation of Σ is complicated because at different points in time, different stocks are available to be included. We therefore make Σ a function of time merely to keep track of the stock availability, that is, we do not impose any structure other than the industry structures. Specifically, given I industries we compute the entries of Σ_t in two steps by using the T, N_{t+1}-dimensional vectors of OLS residuals from model (12). In the first step we collect the time series of the residuals, use them to form I equally weighted industry portfolios, and then compute the in-sample $I \times I$ covariance matrix of such portfolios. In the second step, for each time t and with available N_{t+1} stocks we compute the $N_{t+1} \times N_{t+1}$ matrix Σ_t by using the in-sample variances and industry portfolio covariances from the previous step. For each (i, j) industry pair we assign their covariance to all off-diagonal entries of Σ_t of stocks belonging to industry i and j, while all the residual variances of the available stocks are assigned to the main diagonal.

Different number of industry classifications can provide different specifications for the time-series of residual covariance matrices Σ_t. We use the industry classifications available from K. French website[19] and we generate for each t in our sample positive definite matrices for all the industries. The number of industry classifications that meet the stated requirements are the 5,10,12 and 17. We also consider $\Sigma_t = \sigma^2 I_N$ as in the IPCA model with

[19] $https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data\_library.html$

no industry distinctions. From the point of view of model fit, Figure 3 clearly shows that the 17 industry structure is the best. It has the lowest BIC and the highest R^2_{GLS} values and maximal Sharpe ratio. Generally, all measures improve with the size of the industry classification. We therefore use the 17 industry structure in our empirical analysis.

7 Empirical Results

The results presented in this section are based on the entire sample, except for out-of-sample testing, where a subset of data is set aside for validation. We conduct both cross-sectional and chronological validations. In order to show that our findings are not just dependent on the data of interest, we also carry out a simulation exercise that can apply to a more general setting.

7.1 L Characteristics-based Factors

As noted earlier, at the population level, the GLS factors should perfectly span the mean-variance frontier both conditionally and unconditionally, while other characteristics-based factors that disregard the information contained in the covariance of stock returns should be suboptimal. Consistently, we can observe from Figure 3 that GLS factors achieve the highest unconditional spanning and the best fits (in terms of R^2_{GLS} and BIC). Table 2 shows how the maximal Sharpe ratio obtained from the GLS factors with the 17-industry structure for the covariance matrix of residuals, $\tilde{\beta}^{GLS}$, compares with the Sharpe ratios obtained from the OLS factors, $\tilde{\beta}^{OLS}$, and with the Sharpe ratios obtained from a set of ten OLS hedged factors, obtained through ten rounds of hedging. The hedging procedure is detailed in Appendix A. The GLS spanning is the highest, with a difference $(\tilde{\beta}^{GLS} - max_{i \neq GLS}\{\tilde{\beta}^i\})$ in maximal Sharpe ratio of at least 0.26. Therefore, for a volatility of 20%, the GLS factors provide investors with an unconditional risk premium that at least is 5.2% higher than what other factors can provide. As reported in the third column of the table, the difference between the performance of the GLS factors and other factors is always statistically significant, with p-values less than 0.025. Notice also how the OLS spanning results in the lowest Sharpe ratio and how sequential hedging does help remove some of the unpriced risk. However, sequential hedging can at best cover only a third of the gap that separates the OLS from the GLS spanning (that is, 0.13 out of a 0.39 gap).

These results show the importance of incorporating information from the covariance matrix of returns and how this might be a preferable option than to engage in a discretionary number of sequential hedging to try recovering such information in models that disregard it.

7.2 $K(\leq L)$ Lower-Dimensional Factors

While we have shown empirically that (under the 17-industry structure) the feasible ($L = 37$)-dimensional GLS characteristics-factors achieve the highest unconditional spanning, it is possible that these characteristics may carry similar, somewhat duplicate information. Therefore, more parsimonious ($K < L$)-dimensional subset or combinations of characteristics may capture the same information and if so, they are of economic interest. Theoretically we know that such factors also span the mean-variance frontier given a proper subset of K linear combinations ($Z_t\Gamma_\beta$) of stock characteristics Z_t. Moreover, models that use redundant information may run into overfitting issues while tested out of sample.

Spanning: Figure 4 plots the maximal Sharpe ratios of the RRR(red line) and IPCA(black line) factors with(solid line) and without(dashed line) the mispricing vector Γ_α as a function of the number of extracted risk factors K. When $K = L$, we get the spanning plotted as horizontal lines of the GLS (in red) and OLS (in black) factors with all variables. The spanning of the RRR(IPCA) factors converges from below to that of the GLS(OLS) factors. Thus the maximal Sharpe ratio of the GLS factors act as an empirical upper bound on the highest achievable unconditional spanning under our design. Note that with the IPCA factors, the bound is not reachable. Further observe, as suggested by theory, for any given $K : K \geq 5$ the spanning of the RRR factors when $\Gamma_\alpha = 0$ is consistently the highest.(As we discuss later in Section 7.3, when $K < 5$ the RRR model with $\Gamma_\alpha = 0$ is likely to be misspecified.) Finally, no matter what Γ_α is and what the numbers of extracted risk factors are, the RRR model always achieves a higher spanning than the IPCA model. This can be attributed to incorporating the covariance information in the model.

In Table 3 for each K, the number of extracted risk factors, we provide the statistical significance in the difference of spanning among the relevant analyzed factors. As reported by the first column, it takes 18 factors for the RRR with $\Gamma_\alpha = 0$ to produce a spanning that is not different from the upper bound achieved by the 37 GLS factors at the 5% level. This finding suggests that these 18 factors capture comparable pricing information, and a more parsimonious model is attainable. In contrast, as confirmed by the second column, the spanning of the IPCA factors are always significantly lower than the spanning that results from the GLS upper bound. This corroborates the finding of DeMiguel et al. (2020) and Kozak and Nagel (2023) regarding the inefficiency of OLS-type factors. The third column confirms that the spanning of RRR factors with $\Gamma_\alpha = 0$ is always higher than the spanning of IPCA factors (with the exception of the range $K \in [8, 14]$ where RRR factors with $\Gamma_\alpha = 0$ are still point-wise higher than the corresponding versions in the IPCA setup as shown in Figure 4). Finally, the last two columns display the impact of mispricing in the RRR and IPCA models. This is discussed in detail below.

Mispricing: In the population, given a subset of K stock characteristics, the RRR model with $\Gamma_\alpha = 0$ produces mean-variance efficient factors. Therefore when Γ_α is not zero, we also know the extracted factors are not efficient. Thus a non-zero Γ_α can be taken as a proxy for mispricing.

Figure 5 plots the p-values for testing the null of $\Gamma_\alpha = 0$ (top graph) and the absolute differences in the mean-variance spanning generated by a non-zero Γ_α (bottom graph) for the RRR and IPCA models as a function of the number of extracted risk factors K. The points in the bottom graph with asterisk refer to quantities that are significant at the 5% level, using the test for the difference in maximal squared Sharpe ratios, developed in Barillas et al. (2020). The actual p-values are reported in the last two columns of Table 3. We can state that it takes at least 28(25) factors for the effect of mispricing to become irrelevant for the RRR(IPCA) model. This is (roughly) confirmed by the p-values from the Wald test developed in Section 4.2 against the null of $\Gamma_\alpha = 0$, which gets higher than 5% only after $K = 30(32)$ for the RRR(IPCA) model.

The availability of a limiting distribution for the regression parameters allows us to better assess the role of mispricing coming from the stock characteristics Z_t. In contrast to Kelly et al. (2019) (due to a bias in resampling), we find mispricing to be non-negligible both in the RRR and the IPCA formulations. As a matter of fact, using the code provided for the IPCA model, we find that we could not reject the null at 1% level in the presence of $K = 5$ factors; moreover from $K > 6$ we could not reject at any conventional level. However, this conclusion, for the most part, is in sharp contrast to the significant impact that Γ_α has in the unconditional mean-variance spanning as reported by the black line in the bottom graph of Figure 5. The reported empirical impact of Γ_α is instead in line with the p-value of our proposed Wald test as displayed by the black line in the top graph of Figure 5.

Goodness of Fit: Figure 6 reports the goodness of fit measures discussed in Section 6.2 for the RRR and the IPCA models in red and black respectively. Measures for models with(out) Γ_α are displayed in solid(dashed) lines. The RRR model has a better fit as per the BIC criterion. Also, after few risk factors K are extracted the impact of Γ_α becomes negligible. This is in line with theory as eventually Γ_α becomes 0 when $K \to L$ as there should be no mispricing.

From the R^2_{GLS} metric, the top graph in the figure, we can translate the detected superior fit of RRR models into their ability to consistently explain more of the total variation in the cross-section of stock returns. Observe how the gap between RRR and IPCA models keep widening as we extract more factors. This also implies that the RRR models are more parsimonious; To give some perspective, notice that in the presence of eight factors the RRR models can roughly explain 21.6% of the total cross-sectional variation, to explain the same variation the IPCA models require all 37 factors. Finally, as for the case of BIC and in line with theory, after few (K) risk factors are extracted the impact of Γ_α becomes negligible.

Efficiency of GLS estimators: In Section 3 we have shown how in the population the GLS estimators for the extracted GLS/RRR factors are more efficient. Here we want to relate this fact to the higher mean-variance spanning that these factors can provide. Recall that the Sharpe ratio of the tangency portfolios formed by the factors consists of two components, the average excess returns as the numerator and the standard deviation of the returns as the denominator. Thus assessing the source of the superior spanning comes down to assessing the marginal contribution of the numerator as opposed to that of the denominator. We find that the main reason for the higher GLS/RRR spanning is the lower volatility coming from their tangency portfolios. Therefore the higher statistical efficiency of the GLS estimators translates to the ability of the GLS/RRR factors to achieve a higher mean-variance spanning. The fourth and fifth column of Table 2 report the decomposition of the maximal Sharpe ratios for the GLS, OLS and hedged-OLS factors when $K = L$ (excess returns and volatilities of the tangency portfolios respectively). Notice how the higher spanning of the GLS factors come from the ability of the GLS factors to achieve a tangency portfolio with the lower level of volatility, although the OLS factors and their hedged adjusted factors have higher excess returns.[20] The hedging adjustments help reduce the volatility of the OLS factors but not to the extent reached from the use of the covariance of returns in the RRR modeling. According to the mean-variance theory the fact that GLS factors provide a higher mean-variance spanning while guaranteeing a lower risk premium than OLS factors implies that investors who have access to a better proxy for risk (through the residual covariance matrix specification of our setup) should become more cautious in taking on risk. This is because the implied aggregate risk aversion from the GLS tangency portfolio is lower than that of from the OLS tangency portfolio.[21] Also as reported in Figure

[20]The only exception is the OLS factors with nine rounds of hedging which have a lower volatility of the GLS factor but also a lower average excess return.

[21]Here is the proof that if the tangency portfolio formed by GLS factors has a higher spanning but a lower risk premium with respect to the tangency portfolio formed by OLS factors then the implied risk aversion from the GLS tangency portfolio must be higher. Let $\bar{\beta}^f$ and $\hat{\Gamma}^f$ be the sample mean vector and precision matrix (inverse of the covariance matrix) for the factors $f \in \{GLS, OLS\}$. Then the maximal square Sharpe ratio from f is $\bar{\beta}^{f\prime}\hat{\Gamma}^f\bar{\beta}^f$ and the risk premium for the tangency portfolio is $\frac{1}{\lambda^f}\bar{\beta}^{f\prime}\hat{\Gamma}^f\bar{\beta}^f$ where $\lambda^f = \frac{1}{1'\hat{\Gamma}^f\bar{\beta}^f}$ represent the tangency portfolio implied risk aversion. Because empirically $\bar{\beta}^{GLS\prime}\hat{\Gamma}^{GLS}\bar{\beta}^{GLS} > \bar{\beta}^{OLS\prime}\hat{\Gamma}^{OLS}\bar{\beta}^{OLS}$ and $\frac{1}{\lambda^{GLS}}\bar{\beta}^{GLS\prime}\hat{\Gamma}^{GLS}\bar{\beta}^{GLS} < \frac{1}{\lambda^{OLS}}\bar{\beta}^{OLS\prime}\hat{\Gamma}^{OLS}\bar{\beta}^{OLS}$ it follows that $\lambda^{GLS} > \lambda^{OLS}$.

7 for the leading case of $\Gamma_\alpha = 0$, the volatility and excess return of the factors' tangency portfolios of the RRR models (dashed red lines) are consistently smaller than those of the IPCA models (dashed black lines).[22] These results highlight the crucial role played by the information from the covariances.

7.3 Optimal Model Choice

Results thus far show how RRR factors can more parsimoniously and efficiently span the mean-variance frontier. Selecting a best model thus comes down to deciding K, the optimal number of factors to extract. Observe that such a choice also depends on the presence of the mispricing vector Γ_α in the model. In this section we argue that the RRR model with five factors and $\Gamma_\alpha = 0$ is reasonably a good candidate.

The RRR setup takes K, the number of factors to extract, as given and provide the closest K-dimensional approximation of the GLS factors according to the GLS criterion (25). A closer look at the structure of $\tilde{\beta}^{GLS}$, the $T \times L$ matrix stacking the GLS factors as rows, can help develop a data-driven estimate for K, namely its empirical rank. Reinsel et al. (2022) suggest estimating such rank by looking at the eigenvalues of a standardized version of $\tilde{\beta}^{GLS}$. A natural standardization for our setup is $\tilde{\beta}_\alpha \Gamma^x \tilde{\beta}_\alpha'$, where $\tilde{\beta}_\alpha = \left(I_T - \frac{1_T 1_T'}{T}\right)\tilde{\beta}^{GLS}$ and $\Gamma^x = \sum_t N_{t+1}^2 (Z_t' \Gamma_t Z_t)^{-1}/T$ represents the inverse of the average covariance matrix of residuals from the RRR model (12) when applied to characteristics-based portfolios returns $X_t = Z_t' \Gamma_t r_{t+1}/N_{t+1}$.

The bottom graph of Figure 6 plots, in order of importance, the eigenvalues of $\tilde{\beta}_\alpha \Gamma^x \tilde{\beta}_\alpha'$. As we can see the first few eigenvalues capture the majority of the variation in $\tilde{\beta}^{GLS}$. Moreover, as shown in the second graph from the top, the BIC metric suggests an RRR model that extracts $K = 7$ as the best model. Thus, overall the analysis indicates that perhaps fewer factors than suggested by BIC need to be extracted, for the RRR model to be correctly specified.

[22]Such patterns are also found for the unreported case of $\Gamma_\alpha \neq 0$

These results give an explanation of why the spanning of the RRR model with $\Gamma_\alpha = 0$ reported in Figure 4 is suboptimal for $K < 5$: insufficient extraction of factors can lead to model mis-specification. More specifically, the significance of a higher spanning for $K < 5$ resulting from mispricing (indicated by the fact that the RRR model with a nonzero Γ_α outperforms the RRR model with $\Gamma_\alpha = 0$ at any conventional level) can be attributed to additional unextracted factors that are actually required to be in the model.

We can now use the spanning results from Figure 4 to select a model with a specification that strikes a desirable balance between parsimonious number of factors to extract and maximal achievable Sharpe ratio. The RRR model with $\Gamma_\alpha = 0$ and $K = 5$ seems to stand.[23] Its maximal Sharpe ratio of 4.61 covers approximately 80% of the maximum achievable spanning of the full $L = 37$-dimensional GLS factors. Adding more factors can only marginally contribute to the spanning as the additional 32 extractable factors would result in an increment of the maximal Sharpe ratio by 20%. On the other hand, extracting one less factor would cause the RRR factors from the model to only cover 25% of the maximum achievable spanning, with a maximal Sharpe ratio of 1.46.

Also notice that the RRR model with $\Gamma_\alpha = 0$ and $K = 5$ has a maximal Sharpe ratio of 0.27 higher than the analogous version of IPCA, the best model according to Kelly et al. (2019). The increase in Sharpe ratio is significant at the 10% level and allows the RRR model to deliver an extra premium of 5.4% given a volatility level of 20%. This makes the RRR model better than the models that are compared with the IPCA model, namely, the CAPM and the Fama and French factors models in their specifications including three through six factors.

7.4 Sparseness Results

We first verify the validity of the sparseness framework before presenting our empirical findings. By employing the sparseness method and utilizing additional information derived

[23]We also test and confirm the stability of the significant number of factors, that is five, under different market conditions (such as during NBER recessions and high and low volatility).

from the covariances of returns, we can obtain a more precise interpretation of the extracted factors. Specifically, the first factor represents the market exposure, the second factor is a proxy for size exposure, and the next two factors serve as proxies for momentum, while the fifth factor is a liquidity factor and plays a marginal role.

Validation of the setup: In Section 5 we replaced $Z_t'\Gamma_t Z_t$ with its temporal average $W \equiv \frac{1}{T}\sum_t Z_t'\Gamma_t Z_t$ to be able to implement the sparseness analysis, that is, to run the SOFAR algorithm. In this subsection we check the impact of such assumption.

Specifically, in Section 7.3 we have concluded that the RRR model with $\Gamma_\alpha = 0$ and $K = 5$ can be reasonably taken as the best model. Therefore, here we check the validity of the sparseness setup for that select model and its equivalent IPCA formulation. We do so by implementing the SOFAR algorithm with no sparse penalty (i.e. $\delta = 0$) and compare its GLS R^2 and unconditional maximal Sharpe ratio with those obtained from the baseline setup (PLS algorithm with $Z_t'\Gamma_t Z_t$). This way differences should only come from the questioned assumption. We find the maximal Sharpe ratio(GLS R^2) of the RRR model with $\Gamma_\alpha = 0$ and $K = 5$ to be 4.68(0.206) under the SOFAR algorithm and 4.61(0.209) in the baseline setup. Similarly, we find the maximal Sharpe ratio(GLS R^2) of the IPCA model with $\Gamma_\alpha = 0$ and $K = 5$ to be 4.34(0.198) under the SOFAR algorithm and 4.33(0.20) in the baseline setup. Therefore, we may conclude that replacing $Z_t'\Gamma_t Z_t$ with its temporal average has at most a second order impact in the sparseness analysis.

Empirical results: Having confirmed that the necessary assumption for running the proposed sparseness algorithm is satisfied, we proceed to apply the SOFAR algorithm to the RRR and IPCA models with $\Gamma_\alpha = 0$ and $K = 5$ to conduct the sparseness analysis. It is worth noting that increasing the LASSO penalty δ in the objective function (35) leads to greater sparsity in Γ_β. We select the highest possible value for δ while ensuring that Γ_β is of rank 5 in both setups.

The RRR model generally selects fewer characteristics. Out of the full set of 36 characteristics (37 including the constant), only six appear to be sufficient (7 counting the constant). These characteristics are marked with asterisks in Table 1 and they belong to either the

"Past Returns" or the "Trading frictions" category. They are: momentum and short term reversal, as well as market capitalization, market beta, turnover and the price relative to the previous 52-weeks high. These characteristics also represent six out of the ten characteristics that are found to be significant at the 1% in Kelly et al. (2019) for the corresponding version of the IPCA model. The sparseness version of the IPCA model selects 5 more characteristics as shown in Figure 8[24].

The RRR model with all characteristics remains the best performer as reported in the top panel of Table 4 under the column named "Baseline". Its maximal Sharpe ratio is the highest (see the maximal Sharpe ratios and the p-values on the differences with respect to the Sharpe ratio of the RRR baseline model in the first and second row respectively). As before, the reason of a higher Sharpe ratio with respect to the IPCA setup comes from the lower volatility of its tangent portfolio. Within the RRR setup instead, reducing the number of characteristics decreases the volatility of the tangency portfolio even further but also more than proportionally reduces its excess return. The RRR model in the baseline setup also has the best goodness of fit with the highest GLS R^2 and the lowest BIC values.

A significant benefits of using a sparse design is that it enables a more direct interpretation of the extracted factors when incorporating information from the covariance of returns. As shown in the lower panel of Table 4, the time-series of the SOFAR factors and the baseline (five) factors f_{t+1} exhibit strong correlations in both the RRR and IPCA models. Specifically, all correlations are generally high, with none smaller than 0.5. In Figure 8, we present heatmaps that depict the nonzero entries of the sparse loading matrices Γ_β. This visualization highlights the importance of incorporating information from the second moments of returns, as the structure of the RRR matrix is notably simpler, enabling an easier interpretation of the key determinants that underlie each factor (i.e., the columns of Γ_β).

Specifically, the heatmap for the RRR model reveals that the first factor, which accounts

[24]We remove characteristics that have all zero coefficients in both the IPCA and RRR setups. Therefore, in Figure 8, we only display the selected characteristics that have at least one non-zero coefficient under either the IPCA or RRR model.

for 49% of the cross-sectional variation, is primarily driven by the constant term and the variation in the stock market betas. The second, the third and the fourth factor, explain 28%, 12% and 11% of the cross-sectional variation respectively and are each driven by a single characteristic: size and two proxies for momentum. Finally the fifth factor explains less then 1% of the cross-sectional variation and shares a similar structure with the first factor, but with the turnover variable in place of the market betas. As all the sparse factors are orthogonal, the fifth factor primarily captures liquidity information, while the first factor primarily captures market exposure. We therefore can interpret the first factor as a proxy for the market exposure, the second as a proxy for the size exposure and the next two as proxies for momentum while the fifth factor is a liquidity factor and plays a marginal role. The extracted factors load on similar exposures to the one typically found in the literature, for example the Fama and French models. However, the exposure of our factors are directly extracted from the stock characteristics while those from the literature typically come from sorted portfolios.

In summary, the sparseness design used in conjunction with the RRR model is particularly more useful for interpreting the nature of the extracted factors when information from the covariance of returns is utilized. The visualization of the nonzero entries of the sparse loading matrices enables us to identify the key determinants that drive each factor, and thus allowing for valuable insights into the structure and interpretation of the factors.

7.5 Out of sample results

This section analyses the out-of-sample performance of the RRR model with $\Gamma_\alpha = 0$ and $K = 5$ along two different cross-sectional and temporal dimensions. In the cross-sectional out-of-sample exercise we randomly leave out 20% of stocks evenly from each industry group that was used to construct Σ_t as explained in Section 6.3. We fit the model using the remaining 80% of the data and evaluate the fit and spanning in the remaining 20%. In the temporal out-of-sample exercise, we recursively estimate the model over a 20-year rolling window (i.e. we compute Σ_t and $\Gamma_{\beta,t}$ using only the data available in the rolling window) and produce the one step ahead tradable factors $f_{t+1}^{oos} = (\hat{\Gamma}'_{\beta,t} Z_t' \Gamma_t Z_t \hat{\Gamma}_{\beta,t})^{-1} \hat{\Gamma}'_{\beta,t} Z_t' \Gamma_t r_{t+1}$. We then use the produced time series of estimates to evaluate the model performance. Table 5 reports the out-of-sample performance of the best models from the RRR and IPCA versions.

The performance of the best model is found stable both in terms of spanning and goodness of fit. The in-sample maximal Sharpe ratio is 4.607, while the out-of-sample cross-sectional(time-series) ratio is 4.517(4.077), indicating consistency. Similarly, while the in-sample GLS R^2 is 0.209, the out-of-sample cross-sectional R^2 is even better at 0.236, while the out-of-sample time-series R^2 of 0.202 is very close to the in-sample one.

Also the best RRR model continues to outperform its IPCA analog both in terms of economic and statistical criteria. The difference in maximal Sharpe ratio is roughly 0.3(1) for the cross-sectional(time-series) case, and it is statistically significant. This superiority is due to the higher efficiency of the RRR estimates, which is reflected in the lower volatility of the tangency portfolio (as shown in the fourth row of Table 5) and the ability of the RRR factors to produce a tangency portfolio with higher excess returns (as shown in the third row of Table 5). As far as the statistical fit is concerned, the RRR model has slightly higher GLS R^2: 0.236(0.202) in the cross-sectional(time-series) setup, compared to 0.228(0.190) for the IPCA analog.

In summary, the best RRR model outperform its IPCA analog both in- and out-of-sample.

7.6 Simulation results

To make sure that the findings are not data-specific, in this section we conduct a simulation exercise to more precisely quantify the performance of the RRR and the IPCA models for the case of $\Gamma_\alpha = 0$ and $K = 5$. We have already shown the RRR model's superiority first theoretically and then empirically using a specific dataset. In this section we assume the data generating process for r_{t+1} is known and measure, in addition to the mean-variance spannings of the RRR and the IPCA factors, how close can the RRR and IPCA estimates get to the true parameters .

Specifically, taking the firm characteristics data $\{Z_t\}_t$ and residual covariance matrices $\{\Sigma_t\}_t$ as given, we assume that the data generating process for the excess returns r_{t+1} follows the same setup as in (12)-(14), and generate 1,000 hypothetical datasets. For each dataset we generate $T = 599$ observations for r_{t+1} using (12) and random draws for β_{t+1} and ϵ_{t+1} from the distributions in (14) where $\tilde{\beta}_{t+1}^{GLS}$ and Σ_t are the in-sample estimates from the analysis in Section 7. We then estimate the RRR and IPCA models with $\Gamma_\alpha = 0$ and $K = 5$ using the generated dataset. We obtain for each model the estimate for the loading matrix $\hat{\Gamma}_\beta$ and the $K \times T$ matrix of estimated factors $\hat{F} = [\hat{f}_2, ..., \hat{f}_{T+1}]$.

To measure in a given generated dataset how close to $\tilde{\beta}_{t+1}^{GLS}$, the true value of $\tilde{\beta}_{t+1}$, the product of $\hat{\Gamma}_\beta$ and $\hat{f}_{t+1}$ is, we compute the distance between $\tilde{C}' = [\tilde{\beta}_2^{GLS}, ..., \tilde{\beta}_{T+1}^{GLS}]$, the $T \times L$ matrix stacking the true values $\{\tilde{\beta}_{t+1}\}_t$ over time, and $\hat{F}' \cdot \hat{\Gamma}_\beta'$; that is $D \equiv ||\tilde{C}' - \hat{F}'\hat{\Gamma}_\beta'||_F$. The average difference across the generated datasets are for RRR, $D_{RRR} = 5.223$ and for IPCA, $D_{IPCA} = 5.273$. The difference between these two estimates is significant with a p-value near zero. A larger D_{IPCA} implies that the IPCA estimators have larger estimation errors, which is consistent with our theoretical results.

We also computed for each generated sample the maximal Sharpe ratio obtainable from the factors of each model. The average maximal Sharpe ratio from the RRR factors is 1.82, which is 8.56% more than the Sharpe ratio from IPCA factors (1.68). The difference is statistically significant. It is also economically significant because the difference of 0.14 in Sharpe ratios can generate a 2.8% extra risk premium in normal market environment (assuming an annual market volatility of 20%, which is the average VIX level from 2000 to 2022). As shown in the empirics and as implied theoretically, the superior spanning of the RRR factors decisively comes from a lower median volatility for the tangency portfolio (0.671 versus 0.689), but the median excess return for the portfolio is higher for the IPCA (1.22 versus 1.15). Similar to the in-sample analysis, a better estimate for the factor risks, achieved through the imposed structure on the residual covariance matrix, can yield several benefits. These benefits include reducing the volatility of the tangency portfolio and increasing its returns by eliminating unpriced risk, which can be considered as a form of estimation error. Furthermore, the RRR model has a significantly smaller standard deviation from the distribution of maximal Sharpe ratios (0.103) than that of the IPCA model (0.132), indicating that the RRR setup produces more stable maximal Sharpe ratios. Therefore the simulations confirm the crucial role played by the more efficient GLS estimators for the RRR factors in reaching a higher and more stable mean-variance spanning. As shown here, this conclusion is not data specific and can apply to more general settings.

8 Conclusions

In this paper we introduce a unifying econometric framework for characteristics-based (conditional) factor models via the method of reduced rank regression. Our setup generalizes the IPCA model of Kelly et al. (2019) and the PPCA model of Fan et al. (2016b) by accommodating the cross-sectional dependence in the error terms via the error covariance matrix. Our setup recovers the properties of the estimators of the GLS factors, also described in Kozak and Nagel (2023) at the population level, as well as the lower dimensional approximations.

The main insight of our work is to document the crucial role played by the information contained in the covariances of returns. As noted in (Daniel et al., 2020, p. 1932), "... the characteristic-sorted portfolios ... can be improved by taking into account historical information on the covariance structure." But there are challenges in estimating large covariance matrices. While economic theory would suggest that including the covariance matrix of return in the construction of characteristics-based factors is a necessary condition for mean-variance efficiency, empirically it arises naturally because of the co-movement of asset returns. This introduces constraints across regression coefficients. If there are no constraints, the dependence across equations is not utilized. Theoretical and empirical arguments not only offer different perspectives but actually complement each other. We find, both empirically and via simulations, that the higher mean-variance spanning of models that use the information from the covariances of returns is driven by the higher efficiency of the GLS estimators.[25] According to the mean-variance theory the fact that GLS estimators provide a higher mean-variance spanning while guaranteeing a lower risk premium implies that investors who have access to a better proxy for risk (through the residual covariance matrix specification of our setup) should become more cautious in taking on risk.

Another interesting insight is the role played by sparseness on top of the constrained regression. By making the factor loadings matrix sparse, we can better understand the nature of the extracted factors. Once again incorporating the stock covariances results in more parsimonious models. For the case of the selected model, that is for the RRR model in the absence of mispricing with five factors, sparsity leads to a straightforward interpretation of the first factor as the market, the second as the size, the third and forth as momentum, and the fifth as illiquidity.

Further, the derivation of the limiting distributions for the estimators of the model parameters enables more direct and less computationally intensive closed form tests. It also highlights a bias in the IPCA inferential setup, that leads to underestimation of the role of mispricing in the data analyzed by Kelly et al. (2019). We do provide a method to correct the bias.

The proposed RRR model can be expanded to handle non-linear characteristics by first extracting the non-linear signals contained in the matrix of characteristics (for example by random Fourier features, see Rahimi et al. (2007,2008)) and then use the expanded matrix of characteristics as the new set of characteristics for the analysis. The model can also be used to rigorously test the value added by the stock characteristics beyond known/classical factors such as Fama and French factors. Although the focus of this paper has been on the RRR models with independent errors, the model can easily include an autocorrelation structure in the error terms. We leave these interesting venues open for future research.

[25]Kozak and Nagel (2023) propose a hedging method that does not require inverting a large error-covariance matrix. The empirical results show that multiple rounds of hedging lead to some improvement over OLS estimators but not nearly as good as the GLS estimators. Therefore the estimation of a large covariance matrix and its inverse is of a great deal of importance.

References

Adrian, Tobias, K. Richard Crump, and Erik Vogt, 2019, Nonlinearity and flight-to-safety in the risk-return trade-off for stocks and bonds, *The Journal of Finance* 74, 1931–1973.

Barillas, Francisco, Raymond Kan, Cesare Robotti, and Jay Shanken, 2020, Model Comparison with Sharpe Ratios, *Journal of Financial and Quantitative Analysis* 55, 1840–1874.

Barillas, Francisco, and Jay Shanken, 2017, Which alpha?, *The Review of Financial Studies* 30, 1316–1338.

Box, George EP, 1954, Some theorems on quadratic forms applied in the study of analysis of variance problems, i. effect of inequality of variance in the one-way classification, *The annals of mathematical statistics* 290–302.

Bunea, Florentina, Yiyuan She, and Marten H. Wegkamp, 2012, Joint variable and rank selection for parsimonious estimation of high-dimensional matrices, *The Annals of Statistics* 40, 2359–2388.

Campbell, John Y, Andrew W Lo, A Craig MacKinlay, and Robert F Whitelaw, 1998, The econometrics of financial markets, volume 2, 559–562 (Cambridge University Press).

Chen, Lisha, and Jianhua Z Huang, 2012, Sparse reduced-rank regression for simultaneous dimension reduction and variable selection, *Journal of the American Statistical Association* 107, 1533–1545.

Clarke, Charles, and Matthew Linn, 2023, Characteristics and the cross-section of covariances, Working paper.

Clarke, Charles, and Morteza Momeni, 2021, Testing asset pricing models on individual stocks, Working paper.

Cochrane, John H, 2005, *Asset Pricing: Revised Edition* (Princeton University Press).

Cochrane, John H, 2011, Presidential address: Discount rates, *The Journal of finance* 66, 1047–1108.

Cochrane, John H., and Monika Piazzesi, 2005, Bond risk premia, *American economic review* 95, 138–160.

Daniel, Kent, Lira Mota, Simon Rottke, and Tano Santos, 2020, The cross-section of risk and return, *Review of Financial Studies* 33, 1927–1979.

DeMiguel, Victor, Alberto Martin-Utrera, Francisco J Nogales, and Raman Uppal, 2020, A transaction-cost perspective on the multitude of firm characteristics, *The Review of Financial Studies* 33, 2180–2222.

DeMiguel, Victor, Francisco J Nogales, and Raman Uppal, 2014, Stock return serial dependence and out-of-sample portfolio performance, *The Review of Financial Studies* 27, 1031–1073.

Engle, Robert, and Bryan Kelly, 2012, Dynamic Equicorrelation, *Journal of Business and Economic Statistics* 30, 212–228.

Fama, Eugene F, and Kenneth R French, 1993, Common risk factors in the returns on stocks and bonds, *Journal of financial economics* 33, 3–56.

Fama, Eugene F., and Kenneth R. French, 2020, Comparing cross-section and time-series factor models, *Review of Financial Studies* 33, 1891–1926.

Fan, Jianqing, Yuan Ke, and Yuan Liao, 2021, Augmented factor models with applications to validating market risk factors and forecasting bond risk premia, *Journal of Econometrics* 222, 269–294.

Fan, Jianqing, Yuan Liao, and Liu Han, 2016a, An overview of the estimation of large covariance and precision matrices, *The Econometrics Journal* 19, C1–C32.

Fan, Jianqing, Yuan Liao, and Martina Mincheva, 2013, Large covariance estimation by thresholding principal orthogonal complements, *Journal of the Royal Statistical Society. Series B, Statistical methodology* 75.

Fan, Jianqing, Yuan Liao, and Weichen Wang, 2016b, Projected principal component analysis in factor models, *Annals of Statistics* 44, 219.

Freyberger, Joachim, Andreas Neuhierl, and Michael Weber, 2020, Dissecting characteristics nonparametrically, *The Review of Financial Studies* 33, 2326–2377.

Fujikoshi, Yasunori, 1974, The likelihood ratio tests for the dimensionality of regression coefficients, *Journal of Multivariate Analysis* 4, 327–340.

Gibbons, Michael R, Stephen A Ross, and Jay Shanken, 1989, A test of the efficiency of a given portfolio, *Econometrica: Journal of the Econometric Society* 1121–1152.

Harvey, Campbell R, and Yan Liu, 2022, A census of the factor zoo, *Available at SSRN 3341728* .

Hasbrouck, Joel, and Duane J. Seppi, 2001, Common factors in prices, order flows, and liquidity, *Journal of financial Economics* 59, 383–411.

Kandel, Shmuel, and Robert F Stambaugh, 1995, Portfolio inefficiency and the cross-section of expected returns, *The Journal of Finance* 50, 157–184.

Kelly, Bryan T, Diogo Palhares, and Seth Pruitt, 2022, Modeling corporate bond returns, *Available at SSRN 3720789* .

Kelly, Bryan T, Seth Pruitt, and Yinan Su, 2019, Characteristics are covariances: A unified model of risk and return, *Journal of Financial Economics* 134, 501–524.

Kim, Soohun, Robert A Korajczyk, and Andreas Neuhierl, 2021, Arbitrage portfolios, *The Review of Financial Studies* 34, 2813–2856.

Kozak, Serhiy, and Stefan Nagel, 2023, When do cross-sectional asset pricing factors span the stochastic discount factor?, Working paper.

Ledoit, Olivier, and Michael Wolf, 2017, Nonlinear Shrinkage of the Covariance Matrix for Portfolio Selection: Markowitz Meets Goldilocks, *Review of Financial Studies* 30, 4349–4388.

Ledoit, Olivier, and Michael Wolf, 2020, Analytical Nonlinear Shrinkage of Large-Dimensional Covariance Matrices, *Annals of Statistics* 48, 3043–3065.

Lewellen, Jonathan, Stefan Nagel, and Jay Shanken, 2010, A skeptical appraisal of asset pricing tests, *Journal of Financial Economics* 96, 175–194.

Marshall, Albert W., Ingram Olkin, and Barry C. Arnold, 1979, *Inequalities: theory of majorization and its applications*, volume 143 (Springer).

Rao, C. Radhakrishna, 1967, Least squares theory using estimated dispersion matrix and its application to measurement of signals, *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* 1, 355–372.

Reinsel, Gregory C, Raja Velu, and Kun Chen, 2022, *Multivariate Reduced-Rank Regression Theory, Methods and Applications*, volume 225 (Springer Science & Business Media).

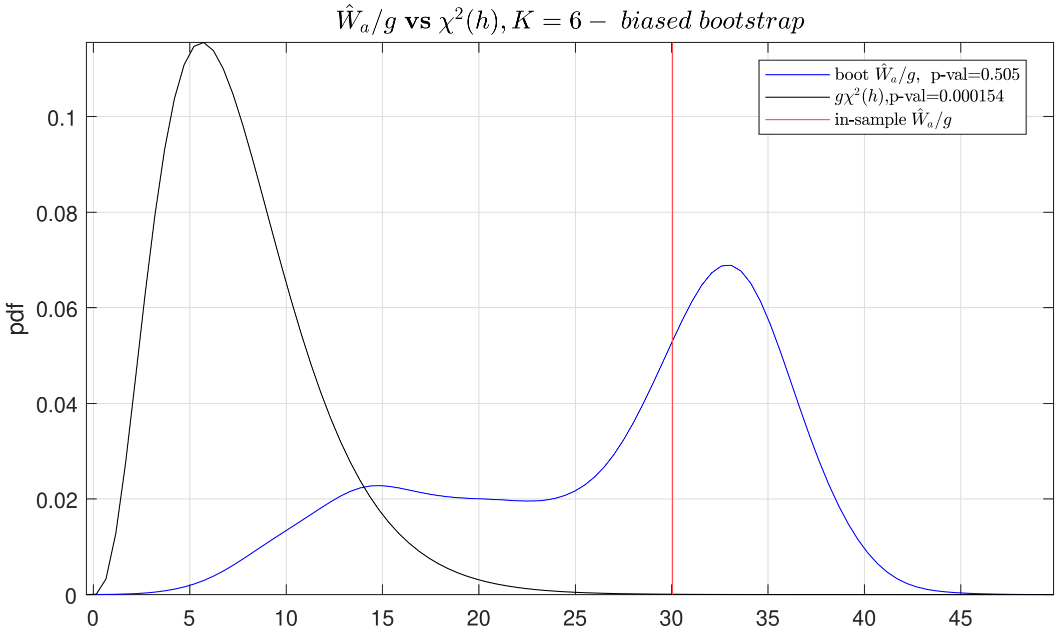
Ross, Stephen, 1976, The arbitrage theory of capital asset pricing, *Journal of Economic Theory* 13, 341–360.

Silvey, Samuel D, 1959, The Lagrangian multiplier test, *The Annals of Mathematical Statistics* 30, 389–407.

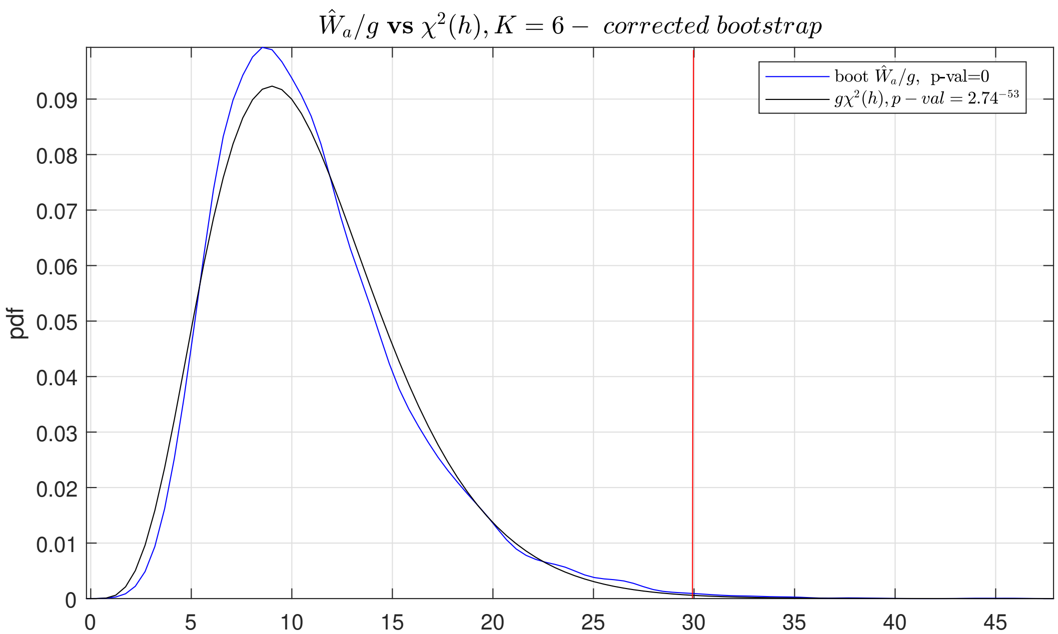
Uematsu, Yoshimasa, Yingying Fan, Kun Chen, Jinchi Lv, and Wei Lin, 2019, Sofar: Large-scale association network learning, *IEEE transactions on information theory* 65, 4924–4939.

Zhang, Chu, 2023, Testing pricing errors of models with latent factors and firm characteristics as covariances, *Management Science* .

Figure 1: **The Bias in the IPCA Inference**
The figure reports the (scaled) bootstrapped distribution from the test developed in Kelly et al. (2019) (blue lines), its approximate limiting distribution (black lines), and the in-sample (scaled) test statistic (red vertical lines) for an IPCA model that extracts six time-varying factors in the presence of mispricing (i.e. with $\Gamma_\alpha \neq 0$). In the top graph (a) the boostrapped distribution is generated using the code from Kelly et al. (2019) while in the bottom graph (b) it is adjusted by assuming that the error term is normal with covariance $\sigma^2 \cdot \frac{Z_t' Z_t}{N_t}$. In the legend, next to each distribution we also report its p-value against the null $\Gamma_\alpha = 0$.



(a)



(b)

Figure 2: **Power functions for the $\hat{W}_\alpha$ and WS test statistics**
Comparison of the power function for the $\hat{W}_\alpha$ and WS test statistics defined in (32) and (33). For demonstration purpose, we set $\Gamma_\alpha = \alpha 1_L$, for $0 < \alpha < 1$, the probability of rejection of $H_0 : \Gamma_\alpha = 0$ is calculated using (32) and (33).

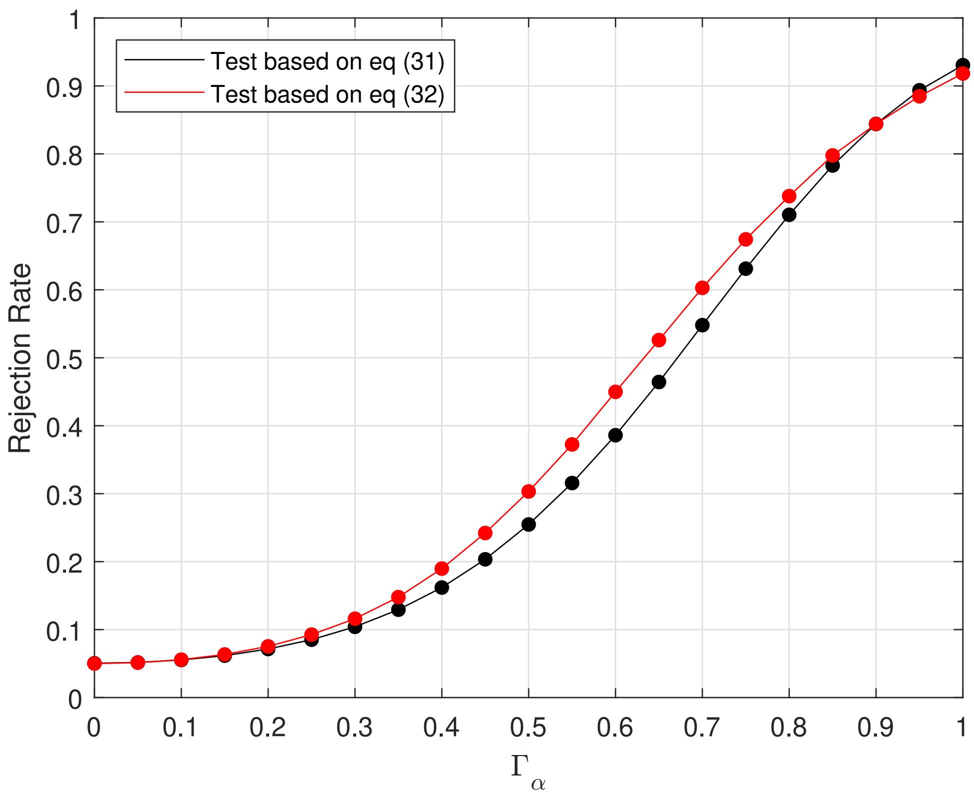


Figure 3: **Performance Criteria for Industry Sizes**
The BIC, R^2_{GLS} and maximal Sharpe ratios (Max.SR) for the covariance structures with 5,10,12 and 17 industries. We refer to the 0-industry structure as a diagonal matrix representing the firm residual variance, $\Sigma = diag(\sigma_i^2)$. When $\Sigma = \sigma^2 I_t$ the GLS factors reduces to the Fama and French (2020) OLS factors.

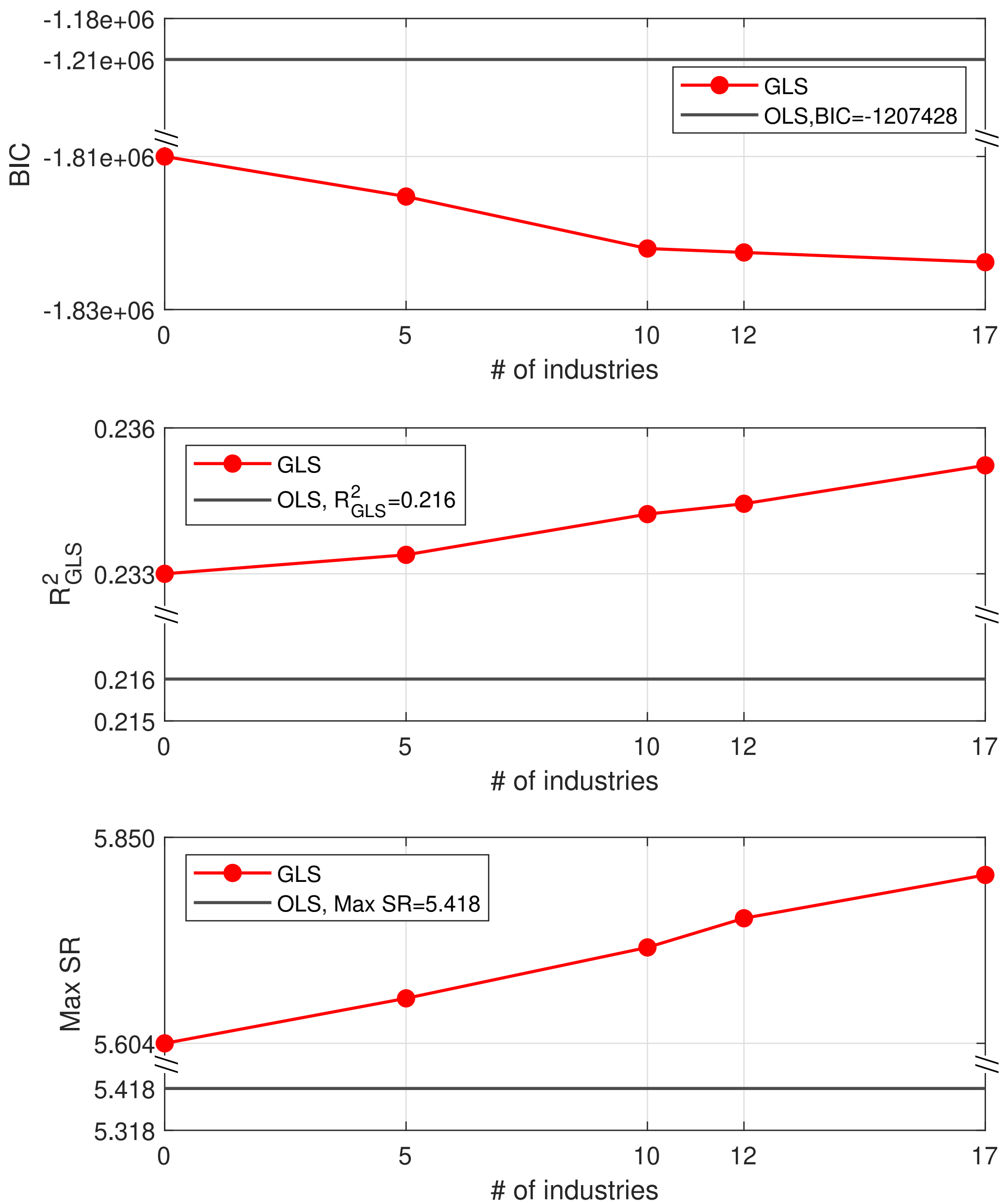


Figure 4: **Maximal Sharpe Ratios**
This figure illustrates the unconditional maximal Sharpe ratios of the RRR_{I17} and IPCA models as well as the maximal Sharpe ratios of the BARRA/OLS factors and the feasible GLS factors. Ratios referring to models with(out) the intercept are plotted using solid(dashed) lines.

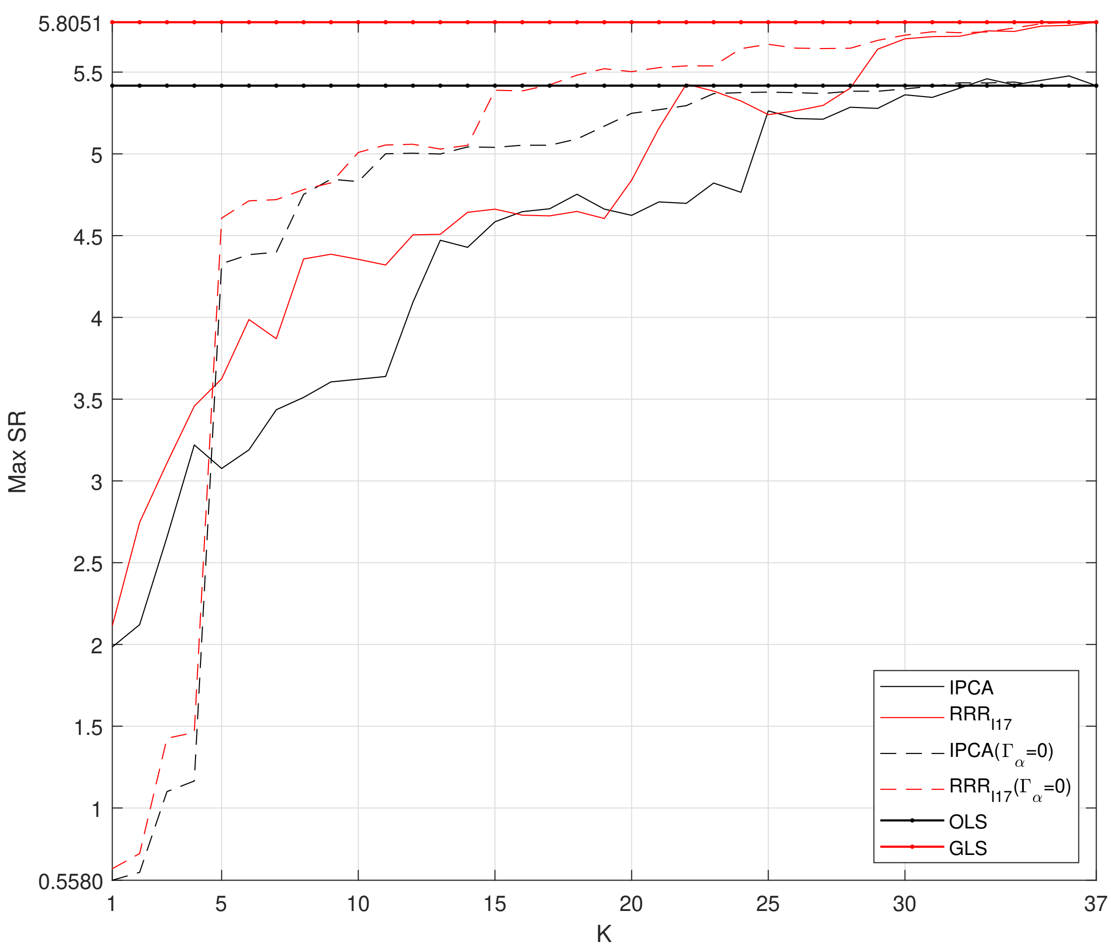


Figure 5: **Test for Mispricing**

The figure plots the p-values for testing, $H_0 : \Gamma_\alpha = 0$ (top graph) and the absolute differences in the maximal Sharpe ratios with and without Γ_α (mispricing) in the RRR and IPCA models for various number of extracted risk factors K. The stars in the bottom graph refer to quantities that are significant at the 5% level using the test for the difference in maximal squared Sharpe ratios developed in Barillas et al. (2020).

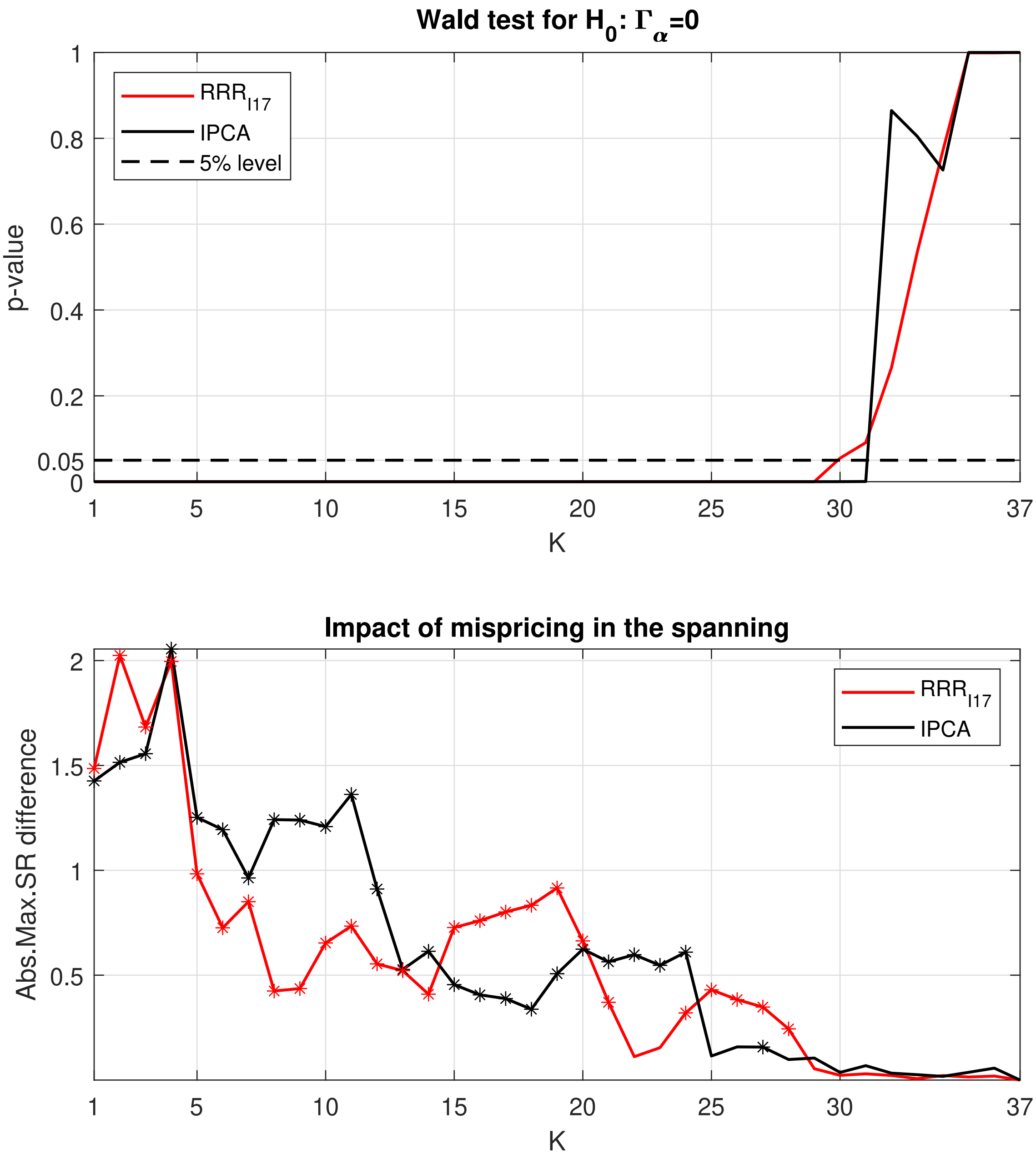


Figure 6: **Goodness of Fit Measures**

The first two plots in the figure reports the goodness of fit measures discussed in section 6.2 for the RRR and the IPCA models in red and black respectively. Quantities of models with(out) Γ_α are displayed in solid(dashed) lines. The last plot displays a RRR diagnostic discussed in section 7.3 mapping the marginal contribution of each additional extracted factor. The first few factors account for the majority of the contribution.

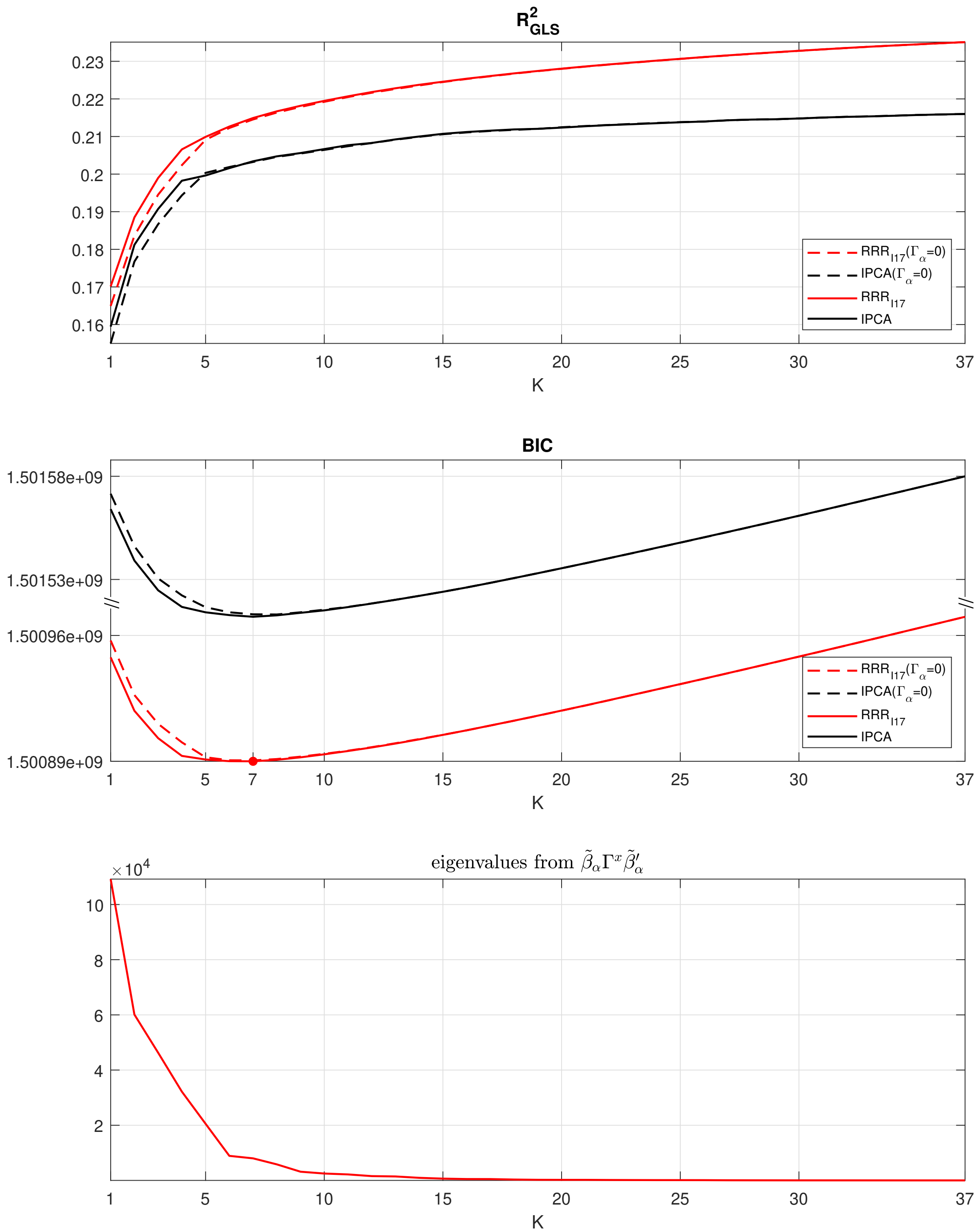


Figure 7: **Decomposition of maximal Sharpe ratios**
The figure reports the decomposition of maximal Sharpe ratios of the RRR and IPCA factors in the absence of mispricing (when $\Gamma_\alpha = 0$).

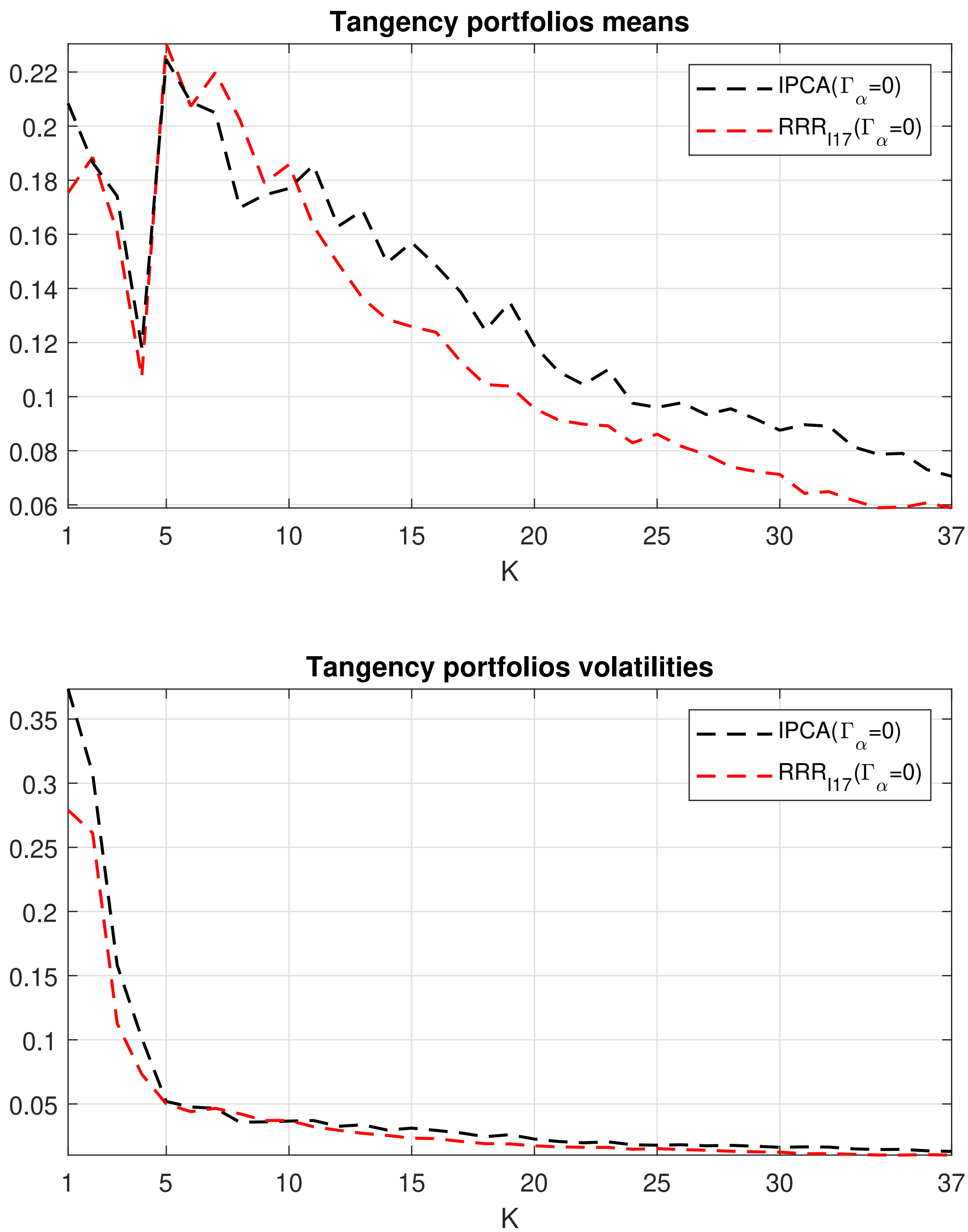


Figure 8: **Sparse matrices in the presence of five factors and no mispricing**
The figure reports the selected firm characteristics and their weights inside the sparse loading matrix Γ_β produced by the SOFAR algorithm for both the RRR and the IPCA models when $K = 5$ and $\Gamma_\alpha = 0$.

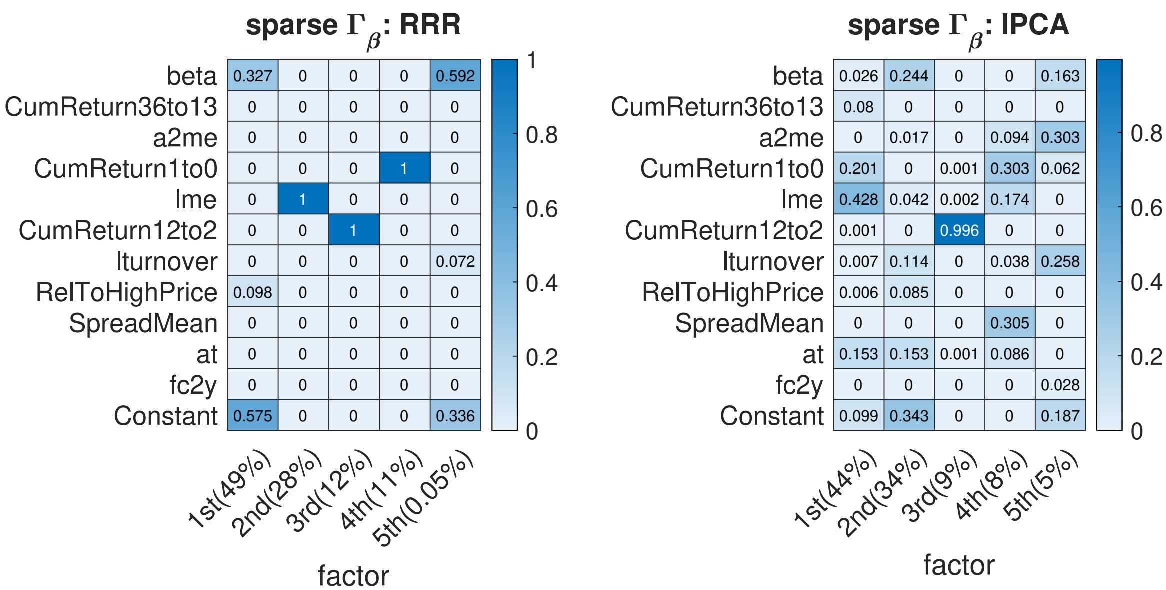


Table 1: **Descriptive statistics for firm characteristics**
Characteristics with asterisks are selected through the sparseness test for a five-factor RRR model in the absence of mispricing ($\Gamma_\alpha = 0$).

Characteriscs	**Mean**	**Median**	**Std**	**Std of ΔZ_t**	**Coefficient of Variation**
Past Returns					
*short_term_reversal**	0.01	0.00	0.17	0.24	17.84
*momentum**	0.15	0.06	0.71	0.35	2.34
intermediate_momentum	0.08	0.03	0.47	0.29	3.63
long_term_reversal	0.35	0.13	1.29	0.52	1.46
Investment					
investment	0.16	0.08	0.67	0.23	1.45
net_operating_asset	0.67	0.68	0.50	0.15	0.23
Δ_{PPE}/Δ_{TA}	0.09	0.05	0.29	0.10	1.14
Profitability					
earning_to_price	-0.02	0.05	0.59	0.21	-11.17
return_to_equitiy	0.05	0.10	1.98	0.71	14.77
capital_turnover	1.41	1.24	1.26	0.21	0.15
sale_to_assets	2.76	2.08	59.66	21.13	7.65
profit_margin	-0.13	0.07	7.95	2.46	-18.27
return_on_net_operating_assets	0.30	0.14	25.97	10.74	36.39
return_to_assets	0.02	0.04	0.19	0.05	2.23
gross_profitability	1.15	0.71	16.05	6.40	5.57
SGA_to_sales	0.56	0.27	1.12	0.25	0.45
price_to_cost_margin	0.36	0.35	1.71	0.61	1.71
Intangibles					
operating_accruals	-0.89	-0.03	582.38	255.71	-288.90
operating_leverage	1.15	1.02	0.95	0.09	0.08
cash_to_short_term_investment	0.13	0.07	0.16	0.02	0.18
fixed_costs_to_sales	0.52	0.24	9.23	2.45	4.72
Value					
assets_to_market	2.83	1.40	7.07	1.61	0.57
book_to_market	0.89	0.65	0.98	0.24	0.27
Tobin's_Q	1.67	1.20	1.73	0.40	0.24
cash_flow_to_book	-0.15	0.04	5.19	1.87	-12.47
leverage	0.32	0.30	0.24	0.03	0.10
sales_to_price	2.78	1.31	5.47	1.13	0.41
capital_intensity	0.04	0.04	0.04	0.01	0.19
Trading frictions					
*market_capitalization** (billion)	1.84	1.17	11.38	1.01	0.55
*market_beta**	1.00	0.93	0.62	0.08	0.08
*turnover**	0.10	0.05	0.21	0.19	1.94
*price_relative_to_52week_high**	0.73	0.78	0.21	0.10	0.13
idiosyncratic_volatility	0.03	0.02	0.03	0.02	0.76
unexplained_volume	0.24	-0.19	3.04	4.25	17.65
bid_ask_spread	0.04	0.02	0.07	0.02	0.56
total_assets (billion)	2.42	0.16	28.54	1.27	0.53

Table 2: **Maximal Sharpe ratios when $K = L$**

The table reports the Maximal Sharpe ratios of the feasible GLS factors $\tilde{\beta}^{GLS}$ under a 17-industry block structure for the covariance matrix of the residuals, the BARRA/OLS factors $\tilde{\beta}^{OLS}$, and ten different versions of hedged OLS factors $\tilde{\beta}^{h,1}, ..., \tilde{\beta}^{h,10}$ obtained using the Kozak and Nagel (2023) procedure detailed in Appendix A. The third column additionally reports the p-values on the difference of a given Maximal Sharpe ratio and that generated from the GLS factors $\tilde{\beta}^{GLS}$. The fourth and the fifth columns report the average excess returns (Sharpe ratio numerators) and the volatilities (Sharpe ratio denominators) of the tangent portfolios generated by the factors.

factors	**Sharpe ratio**	**p − value**	Mean excess return	Volatility
$\tilde{\beta}^{GLS}$	5.81	–	0.296	0.051
$\tilde{\beta}^{OLS}$	5.42	0.005	0.842	0.155
$\tilde{\beta}^{h,1}$	5.54	0.023	0.405	0.073
$\tilde{\beta}^{h,2}$	5.49	0.009	0.389	0.071
$\tilde{\beta}^{h,3}$	5.55	0.025	0.323	0.058
$\tilde{\beta}^{h,4}$	5.51	0.010	0.334	0.061
$\tilde{\beta}^{h,5}$	5.55	0.024	0.296	0.053
$\tilde{\beta}^{h,6}$	5.51	0.010	0.317	0.058
$\tilde{\beta}^{h,7}$	5.55	0.022	0.282	0.051
$\tilde{\beta}^{h,8}$	5.51	0.009	0.305	0.055
$\tilde{\beta}^{h,9}$	5.55	0.021	0.273	0.049
$\tilde{\beta}^{h,10}$	5.51	0.009	0.297	0.054

Table 3: **P-values behind the unconditional spanning displayed in Figure 4**
The table reports the p-values against the null that the differences in the maximal Sharpe ratios generated by factors x and y are zero. Column two through column six report the differences on the tested (x, y) pairs. $\tilde{\beta}^{GLS}$ are the feasible regression GLS factors under a 17-industry block structure for the covariance matrix of the residuals, $\hat{f}^{GLS}_{\Gamma_\alpha=0}(\hat{f}^{IPCA}_{\Gamma_\alpha=0})$ are the RRR(IPCA) factors in the absence of mispricing, and $\hat{f}^{GLS}(\hat{f}^{IPCA})$ are the RRR(IPCA) factors in the presence of mispricing.

	(x, y): Maximal $SR(x)-$ Maximal $SR(y)$				
# of factors	$(\tilde{\beta}^{GLS}, \hat{f}^{GLS}_{\Gamma_\alpha=0})$	$(\tilde{\beta}^{GLS}, \hat{f}^{IPCA}_{\Gamma_\alpha=0})$	$(\hat{f}^{GLS}_{\Gamma_\alpha=0}, \hat{f}^{IPCA}_{\Gamma_\alpha=0})$	$(\hat{f}^{GLS}_{\Gamma_\alpha=0}, \hat{f}^{GLS})$	$(\hat{f}^{IPCA}_{\Gamma_\alpha=0}, \hat{f}^{IPCA})$
1	0	0	0.07	0	0
2	0	0	0.03	0	0
3	0	0	0	0	0
4	0	0	0	0	0
5	0	0	0.08	0	0
6	0	0	0.01	0	0
7	0	0	0.01	0	0
8	0	0	0.86	0	0
9	0	0	0.83	0	0
10	0	0	0.15	0	0
11	0	0	0.62	0	0
12	0	0	0.62	0	0
13	0	0	0.76	0	0
14	0	0	0.91	0.01	0
15	0.03	0	0.02	0	0
16	0.02	0	0.02	0	0
17	0.02	0	0.01	0	0
18	0.07	0	0.01	0	0
19	0.12	0	0.02	0	0
20	0.07	0.01	0.09	0	0
21	0.09	0.01	0.08	0	0
22	0.11	0.01	0.11	0.28	0
23	0.1	0.04	0.23	0.15	0
24	0.48	0.04	0.06	0	0
25	0.62	0.03	0.04	0	0.26
26	0.39	0.03	0.04	0	0.06
27	0.31	0.02	0.03	0	0.02
28	0.27	0.02	0.04	0.01	0.15
29	0.44	0.02	0.02	0.38	0.15
30	0.65	0.02	0.01	0.54	0.52
31	0.85	0.02	0.02	0.47	0.28
32	0.61	0.03	0.02	0.57	0.42
33	0.56	0.02	0.02	0.98	0.67
34	0.76	0.02	0.02	0.54	0.59
35	0.86	0.01	0.01	0.54	1
36	0.78	0.01	0.01	0.44	1
37	1	0.01	0.01	0.01	1

Table 4: **Sparseness analysis with five factors and no mispricing**

The table reports the performance of the sparse RRR and IPCA model, which includes five factors and mispricing term (i.e. $\Gamma_\alpha = 0$). The baseline models include all characteristics and extract five factors, while the sparse models, described in Section 5, are estimated using the SOFAR algorithm.

	RRR		IPCA	
	Baseline	Sparse	Baseline	Sparse
Maximal Sharpe ratio	4.607	4.019	4.328	3.518
$\Delta^{SR}_{RRR_{base}}$ p-value	-	0.000	0.077	0.000
Tan. port. excess return	0.230	0.0944	0.225	0.267
Tan. port. volatility	0.050	0.0235	0.052	0.076
R^2_{GLS}	0.209	0.187	0.20	0.192
Best model: BIC	x			

	RRR	IPCA
	Corr(Sparse,Baseline)	Corr(Sparse,Baseline)
Factor 1	0.97	0.78
Factor 2	0.50	0.95
Factor 3	0.51	0.60
Factor 4	0.63	0.89
Factor 5	0.72	0.82

Table 5: **Validation Results**

The table reports the out-of-sample performance of the RRR and IPCA model with five time-varying factors and no mispricing term (i.e. $\Gamma_\alpha = 0$) under the cross-sectional and time-series designs described in section 7.5. "p-value of $\Delta^{SR}_{RRR-IPCA}$" is, for each design, the p-value under the null that the difference between the maximal Sharpe ratio of the RRR and IPCA model is reported as the same. The third and the forth row report the average excess returns (Sharpe ratio numerators) and the volatilities (Sharpe ratio denominators) of the tangent portfolios generated by the discussed factors. The last row reports the GLS R^2 described in section 6.2 computed using the time series of out-of-sample factors.

	Cross-sectional		Time-series	
	RRR	IPCA	RRR	IPCA
Maximal Sharpe ratio	4.517	4.224	4.077	3.156
p-value of $\Delta^{SR}_{RRR-IPCA}$	0.000	-	0.000	-
Tan. port. excess return	0.232	0.229	0.199	0.142
Tan. port. volatility	0.052	0.054	0.049	0.053
R^2_{GLS}	0.236	0.228	0.202	0.190

Appendix

A OLS Hedged Factors

In this section we report the steps, worked out in Kozak and Nagel (2023), to compute a version of the Fama and French (2020) OLS factors residualized with respect to the zero-expected return hedge portfolios. The implied variance of r_{t+1} in model (12) can be decomposed into two distinct components: priced risk and unpriced risk:

$$\Sigma_t = Z_t \Phi_t Z_t' + U_t \Omega_t U_t' \quad \text{and} \quad U_t' Z_t = 0 \tag{39}$$

with conformable matrices Φ_t and Ω_t.

We define H_t, the $N_t \times K$ matrix, as the weights for the K OLS hedged factors at time t. We expect that $H_t' X_t$ has full rank to ensure that no information about expected returns is lost. Additionally, we impose $U_t' H_t = 0$ to guarantee that the factors do not load on unpriced risk. Although U_t is not directly observable, it still can partially be determined from moments of R_t and Z_t. Kozak and Nagel (2023) propose an advanced approach based on Daniel et al. (2020)'s method for case $U_t' Z_t \neq 0$, which is outlined as follow:

The first step is to construct a hedge stock portfolio that has precisely zero expected return. This is achieved by regressing conditional covariances of individual stocks with factors, $Cov(r_{t+1}, (Z_t' Z_t)^{-1} Z_t' r_{t+1}) = \Sigma_t Z_t (Z_t' Z_t)^{-1}$, on Z_t, and then using the residuals,

$$W_{h,t} = \bar{P}_t \Sigma_t Z_t (Z_t' Z_t)^{-1}, \tag{40}$$

where $\bar{P}_t = I - Z_t (Z_t' Z_t)^{-1} Z_t'$, as portfolio weights for the hedge portfolio.

The second step is to calculate the covariances between stocks' returns and the hedge portfolio returns so that we can remove the component that is correlated with the unpriced risk U_t from factor portfolio weights in the next step:

$$\hat{V}_t = \Sigma_t W_{h,t}. \tag{41}$$

The third step is to regress the factor portfolio weights $Z_t (Z_t' Z_t)^{-1}$ on $\hat{V}_t$ to obtain residual factor portfolio weights $\hat{H}_t$ that have been purged of unpriced risk exposure

$$\hat{H}_t = Z_t (Z_t' Z_t)^{-1} - \hat{V}_t (\hat{V}_t' \hat{V}_t)^{-1} \hat{V}_t' Z_t (Z_t' Z_t)^{-1}. \tag{42}$$

Finally the OLS hedged factors are obtained as $\tilde{\beta}^h_{t+1} = \hat{H}_t' r_{t+1}$.

A.1 Multiple rounds of hedging

Kozak and Nagel (2023) further show that additional improvements on the OLS factors can be achieved by engaging in multiple rounds of hedging. For example, the n-th hedging round is achieved by regressing stocks' conditional covariances with *hedged* factors from the $n-1$ round, i.e. $\Gamma_t^{-1}\hat{H}_t^{n-1}$ on Z_t and collecting the residuals $R_t\Gamma_t^{-1}\hat{H}_t^{n-1}$. Then, calculate stocks' covariances with these residuals, getting $\hat{V}_{n,t}$. Finally, obtain the hedged portfolio weighs after n rounds of hedging $\hat{H}_t^n$ as the residuals from regressing $Z_t(Z_t'Z_t)^{-1}$ on $\hat{V}_t$ through $\hat{V}_{n,t}$.

B Expressions Related to Result 1

The GLS criterion $S(\theta)$ is defined in equation (25). Let $s = L + K(T + L)$ be the dimension of the vector of RRR parameters θ defined in Section 4. The second derivative matrix B_θ with its elements is:

$$\underset{s\times s}{B_\theta} = \frac{\partial^2 S(\theta)}{\partial\theta(\partial\theta)'} = \frac{1}{T}\begin{bmatrix} \frac{\partial^2 S(\theta)}{\partial\Gamma_\alpha(\partial\Gamma_\alpha)'} & \frac{\partial^2 S(\theta)}{\partial\Gamma_\alpha(\partial f)'} & \frac{\partial^2 S(\theta)}{\partial\Gamma_\alpha(\partial vec(\Gamma_\beta'))'} \\ \left(\frac{\partial^2 S(\theta)}{\partial\Gamma_\alpha(\partial f)'}\right)' & \frac{\partial^2 S(\theta)}{\partial f(\partial f)'} & \left(\frac{\partial^2 S(\theta)}{\partial vec(\Gamma_\beta')(\partial f)'}\right)' \\ \left(\frac{\partial^2 S(\theta)}{\partial\Gamma_\alpha(\partial vec(\Gamma_\beta'))'}\right)' & \frac{\partial^2 S(\theta)}{\partial vec(\Gamma_\beta')(\partial f)'} & \frac{\partial^2 S(\theta)}{\partial vec(\Gamma_\beta')(\partial vec(\Gamma_\beta'))'} \end{bmatrix}$$

where

$$\frac{\partial^2 S(\theta)}{\partial\Gamma_\alpha(\partial\Gamma_\alpha)'} = \underset{L\times L}{\sum_t Z_t'\Gamma_{t+1}Z_t},$$

$$\frac{\partial^2 S(\theta)}{\partial\Gamma_\alpha(\partial f)'} = \underset{L\times TK}{[(Z_1'\Gamma_2 Z_1)\Gamma_\beta, ..., (Z_T'\Gamma_{T+1}Z_T)\Gamma_\beta]}$$

$$\frac{\partial^2 S(\theta)}{\partial\Gamma_\alpha(\partial vec(\Gamma_\beta'))'} = \underset{L\times KL}{\sum_t (Z_t'\Gamma_{t+1}Z_t \otimes f_{t+1}')},$$

$$\frac{\partial^2 S(\theta)}{\partial f(\partial f)'} = \underset{TK\times TK}{Diag\left(\Gamma_\beta'(Z_1'\Gamma_2 Z_1)\Gamma_\beta, \cdots, \Gamma_\beta'(Z_T'\Gamma_{T+1}Z_T)\Gamma_\beta\right)}$$

$$\frac{\partial^2 S(\theta)}{\partial vec(\Gamma_\beta')(\partial f)'} = [-(Z_t'\Gamma_{t+1}Z_t)(\beta_{t+1} - \Gamma_\alpha - \Gamma_\beta f_{t+1}) \otimes I_K] + (Z_t'\Gamma_{t+1}Z_t \otimes f_{t+1})\Gamma_\beta$$

$$\frac{\partial^2 S(\theta)}{\partial vec(\Gamma_\beta')(\partial vec(\Gamma_\beta'))'} = \underset{KL\times KL}{\sum_t \left(Z_t'\Gamma_{t+1}Z_t \otimes f_{t+1}f_{t+1}'\right)}.$$

$H_\theta = \{\partial h_j(\theta)/\partial\theta_i\}$ is a $s\times K(K+1)$ matrix containing the partial derivatives of $h(\theta) = 0$, the $K(K+1)\times 1$ vector of normalization conditions described in (26). The first K components of $h(\theta)$ are given by $\Gamma_\beta'\Gamma_\alpha$, the next $\frac{K(K-1)}{2}$ components are of the form $F_i'F_j, i \leq j$ where F_i denotes the i-th column of the matrix F'.

The remaining $\frac{K(K+1)}{2}$ are of the form $\Gamma_{\beta}^{(i)\prime}\Gamma_{\beta}^{(j)} - \delta_{ij}, i \leq j$ where $\Gamma_{\beta}^{(j)}$ denotes the j-th column of Γ_{β} and $\delta_{ij} = 1$ for $i = j$ and 0 otherwise.

C Sparseness Algorithm Based on SOFAR

The discussion in Section 5 leads to the following algorithm:

Input: $\underset{L\times T}{Y} = W^{1/2}(\tilde{\beta}^{GLS\prime} - \hat{\Gamma}_{\alpha}1_T')$ and $\underset{L\times L}{X} = W^{1/2}$

Output: Γ_{β}, F and Λ

Step 1: construct an initial estimator $\tilde{C}$ with $\lambda_0 \geq 0$ from

$$\tilde{C} = \underset{C}{\text{argmin}} \left\{ \tfrac{1}{2T} \|W^{1/2}(\tilde{\beta}^{GLS\prime} - \hat{\Gamma}_{\alpha}1_T') - W^{1/2}C\|_F^2 + \lambda_0\|C\|_1 \right\}$$

Get U_0, D_0 and V_0 from the singular value decomposition (SVD) of $C = U_0 D_0 V_0$.

Step 2: Use the augmented Lagrangian method (ALM) coupled with block coordinate descent (BCD) to search the solution of the following problem

$$(\hat{\Theta}, \hat{\Omega}) = \underset{\Theta,\Omega}{\text{argmin}} \left\{ \tfrac{1}{2T} \|W^{1/2}(\tilde{\beta}^{GLS\prime} - \hat{\Gamma}_{\alpha}1_T') - W^{1/2}UDV'\|_F^2 + \delta\|UD\|_1 \right\}$$
$$s.t. U'U = I, V'V = I, UD = A, VD = B$$

where $\Theta = (D, U, V)$ and $\Omega = (A, B)$.

Get the converge solution D^*, U^* and V^*.

Step 3: Construct the solution for our problem. $\Gamma_{\beta} = U$, $F = DV'$ and $\Lambda = D^2$.

In our case, we choose to make only Γ_{β} sparse while keeping the factor time series non-sparse. As a result, we add only one penalty term to our problem, which is a special case of the SOFAR algorithm. The ALM-BCD algorithm, which we used in step 2, is described in detail in Table 1 of Uematsu et al. (2019).

D AIC Criterion

In our empirical studies, we compute AIC metric as

$$AIC = 2s - 2lnL(\theta) \tag{43}$$

where the log-likelihood $lnL(\theta) \equiv \sum_t^T lnf(r_{t+1}|Z_t;\theta) = -\frac{1}{2}\left[ln(2\pi)\sum_t^T N_t + \sum_t^T ln|\Gamma_t^{-1}|\right] - TS(\theta)$, $S(\theta)$ is defined in equation (25) and s is the total number of estimated parameters. Refer to footnote [18] in the main text for further detail on s.

In this robustness check, we use the AIC metric to select the optimal industry structure and evaluate the goodness of fit of our analyzed models. Figure 9(left) confirms that the GLS model outperforms the OLS model, as the former exhibits a lower AIC. Moreover, the AIC is decreasing in the number of industries, indicating that GLS factors with larger industry block structure performs better. These results confirm that the optimal model is the one based on the 17-industry structure, which is consistent with the findings based on BIC, R^2_{GLS} and maximal Sharpe ratios as presented in Figure 3.

In Figure 9(right), we present the AICs for models with different number of factors. As consistently observed, the RRR model outperforms the IPCA model. Additionally, we find that models without the mispricing term perform better than models with one. The AIC results indicate that the optimal number of factors is 18, which confirms that the importance of a parsimonious model.

Figure 9: Robustness Check Based on AIC Criterion