Factor Models, Machine Learning, and Asset Pricing

Bryan Kelly
Yale University, AQR Capital Management, and NBER

Dacheng Xiu
Booth School of Business
University of Chicago

October 15, 2021

Abstract

We survey recent methodological contributions in asset pricing using factor models and machine learning. We organize these results based on their primary objectives: estimating expected returns, factors, risk exposures, risk premia, and the stochastic discount factor, as well as model comparison and alpha testing. We also discuss a variety of asymptotic schemes for inference. Our survey is a guide for financial economists interested in harnessing modern tools with rigor, robustness, and power to make new asset pricing discoveries, and it highlights directions for future research and methodological advances.

Key words: asset pricing, machine learning, factor models, stochastic discount factor, risk premium

Contents

1 INTRODUCTION

Factor models are natural workhorses for modeling equity returns because they offer a parsimonious statistical description of returns' cross-sectional dependence structure. Beyond their statistical relevance, the arbitrage pricing theory (APT) of Ross (1976) provides a rigorous economic motivation for factor models. The APT describes how statistical factor representations are directly tied to foundational economic concepts, such as risk exposures and risk premia, which govern the risk-return tradeoff at the core of asset pricing. In light of this, factor models have become the single most widely adopted research paradigm for academics and practitioners alike.

Nevertheless, unobservable asset risk premia are notoriously difficult to pinpoint, because market efficiency forces return variation to be dominated by unforecastable news that obscures expected returns. In addition, the sample size of equity returns is small relative to the predictor count. Structural breaks, regime switches, and non-stationarity in general further diminish the effective sample size. Furthermore, the collection of candidate conditioning variables is large and they are often close cousins and highly correlated. Further still, complicating the problem is ambiguity regarding functional forms through which the high-dimensional predictor set enter into the expected returns. All these issues result in a low signal-to-noise ratio environment that is in stark contrast to prediction problems in computer science and other domains.

Certain aspects of the machine learning paradigms, such as variable selection and dimension reduction, have been part of empirical asset pricing since the very beginning of this research field. In early days, economic theories and parsimonious model specifications were adopted in order to "regularize" learning problems in financial markets. Indeed, we have become accustomed to sorting stocks by their characteristics, forming equal or value weighted portfolios, and selecting a small number of portfolios as factor proxies. These choices have been made, either explicitly or implicitly, to cope with nonlinearity, low signal-to-noise ratios, and the curse of dimensionality that are a difficulty reality when studying asset returns.

Thankfully, recent decades have seen rapid growth of exploratory and predictive techniques proposed by the statistics and machine learning communities. These tools complement economic theory to provide a data-driven solution to the empirical challenges of asset pricing. Embracing these tools enables economists to make rigorous, robust, and powerful empirical discoveries, about which economic theory alone may not be a sufficient guide. Conversely, these new discoveries can offer new insights from data that in turn lead to improved economic theories.

Our objectives in this article are twofold. First, we survey recent methodological contributions in empirical asset pricing. We categorize these based on the methodologies' primary purpose, which range from estimating expected returns, factors and assets' exposures to them, risk premia, and stochastic discount factors, to comparing asset pricing models and testing alphas. Second, we discuss the accompanying asymptotic theory, broken out by their focus time series asymptotics (large T), cross-sectional asymptotics (large N), or two-dimensional panel asymptotics (large T and N), to help guide financial economists to methods most appropriate for their specific research needs. Along the way we compare methodologies, highlight their strengths and limitations, and point out future

directions for improvement.

Throughout the survey, we use $(A : B)$ to denote the concatenation (by columns) of two matrices A and B. e_i is a vector with 1 in the ith entry and 0 elsewhere, whose dimension depends on the context. ι_k denotes a k-dimensional vector with all entries being 1, and $\mathbb{I}_K$ denotes the $K \times K$ identity matrix. For any time series of vectors $\{a_t\}_{t=1}^T$, we denote $\bar{a} = \frac{1}{T}\sum_{t=1}^T a_t$. In addition, we write $\bar{a}_t = a_t - \bar{a}$. We use the capital letter A to denote the matrix $(a_1 : a_2 : \ldots : a_T)$, and write $\bar{A} = A - \bar{a}\iota_T^\intercal$ correspondingly. We denote $\mathbb{P}_A = A(A^\intercal A)^{-1}A^\intercal$ and $\mathbb{M}_A = \mathbb{I}_K - \mathbb{P}_A$, for some $K \times T$ matrix A. We use $a \vee b$ to denote the max of a and b, and $a \wedge b$ as their min, for any scalars a and b. We also use the notation $a \lesssim b$ to denote $a \leq Cb$ for some constant $C > 0$. Similarly, we use $x \lesssim_{\mathrm{P}} y$ to denote $x = O_{\mathrm{P}}(y)$ for two random variables x and y.

We use $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ to denote the minimum and maximum eigenvalues of A, and use $\lambda_i(A)$ to denote the i-th largest eigenvalue of A. Similarly, we use $\sigma_i(A)$ to denote the ith singular value of A. We use $\|A\|_1$, $\|A\|_\infty$, $\|A\|$, and $\|A\|_{\mathrm{F}}$ to denote the $\mathbb{L}_1$ norm, the $\mathbb{L}_\infty$ norm, the operator norm (or $\mathbb{L}_2$ norm), and the Frobenius norm of a matrix $A = (a_{ij})$, that is, $\max_j \sum_i |a_{ij}|$, $\max_i \sum_j |a_{ij}|$, $\sqrt{\lambda_{\max}(A^\intercal A)}$, and $\sqrt{\mathrm{Tr}(A^\intercal A)}$, respectively. We also use $\|A\|_{\mathrm{MAX}} = \max_{i,j}|a_{ij}|$ to denote the $\mathbb{L}_\infty$ norm of A on the vector space. Finally, we use $Diag(A)$ to denote the diagonal matrix of A and $A_{[I]}$ a submatrix of A whose rows are indexed in I.

2 MODEL SPECIFICATIONS

We start by introducing a static factor model, which serves as a benchmark throughout the survey.

2.1 Static Factor Models

In its simplest form, a static factor model can be written as

$$r_t = \mathrm{E}(r_t) + \beta v_t + u_t, \tag{1}$$

where r_t is an $N \times 1$ vector of excess returns of test assets (e.g., size-value double-sorted portfolios), β is an $N \times K$ matrix of factor exposures, v_t is a $K \times 1$ vector of factor innovations, and u_t is an $N \times 1$ vector of idiosyncratic errors.

The expected return can (always) be decomposed as:

$$\mathrm{E}(r_t) = \alpha + \beta\gamma, \tag{2}$$

where γ is a $K \times 1$ vector of risk premia and α is an $N \times 1$ vector of pricing errors. This projection always holds as the right-hand-side has more degrees of freedom than the left, but the APT of Ross (1976) and follow up work by Huberman (1982), Ingersoll (1984), and Chamberlain & Rothschild (1983) predict that no-asymptotic arbitrage implies $\alpha^\intercal\Sigma_u^{-1}\alpha < \infty$ as N increases, where Σ_u is the covariance matrix of u_t.

The most common framework in academic finance literature assumes that factors are known and

observable (an example would be industrial production growth, as in Chen et al., 1986). That is,

$$f_t = \mu + v_t, \tag{3}$$

where μ is some unknown parameter of the (population) expectation of f_t. If factors are tradable portfolios (such as the Fama-French factors in Fama & French, 1993, 2015), then $\mu = \gamma$.

A second framework, which has re-gained popularity recently but dates back to as early as Connor & Korajczyk (1986), assumes that all factors and their exposures are latent, which relaxes the somewhat restrictive assumption in the first setting that all factors are known and observable to econometricians.

A third framework assumes factor exposures are observable but the factors are latent. This is arguably the most prevalent framework for practitioners and is rooted in the MSCI Barra model originally proposed by Rosenberg (1974). The popularity of this model stems from the fact that it conveniently accommodates time-varying exposures of individual equity returns. We turn to the topic of time-varying exposures next.

2.2 Conditional Factor Models

One might argue that the static model in 1 is suitable for certain portfolios of assets (though even in this case the static assumption is dubious), but it is clearly inadequate for most individual assets. For example, risk exposures of individual stocks very likely change over time as firms evolve. More pointedly, assets with fixed maturities and nonlinear payoff structures (e.g., options and bonds) experience mechanical variation in their risk exposures as their maturity rolls down or the value of the underlying asset changes. In this case, a factor model should accommodate conditional risk exposures. The conditional factor model can be specified as:

$$\tilde{r}_t = \alpha_{t-1} + \beta_{t-1}\gamma_{t-1} + \beta_{t-1}v_t + \tilde{u}_t, \tag{4}$$

where $\tilde{r}_t$ and $\tilde{u}_t$ are $M \times 1$ vectors of excess returns and idiosyncratic errors of individual stocks.

Obviously, the right-hand side contains too many degrees of freedom and the model cannot be identified without additional restrictions. Rosenberg (1974) imposes that $\beta_{t-1} = b_{t-1}\beta$, where b_{t-1} is an $M \times N$ matrix of observable characteristics and β is an $N \times K$ vector of parameters. Consequently, the model becomes

$$\tilde{r}_t = b_{t-1}\tilde{f}_t + \tilde{\varepsilon}_t, \tag{5}$$

where $\tilde{f}_t := \beta(\gamma_{t-1} + v_t)$ is a new $N \times 1$ vector of latent factors, and $\tilde{\varepsilon}_t := \alpha_{t-1} + \tilde{u}_t$.[1] This is the MSCI Barra model prototype which has been embraced by practitioners for its simplicity and versatility in modeling individual equity returns.

Barra's model includes several dozens of characteristics and industrial variables in b_{t-1}. Their ad hoc selection procedure is opaque and evidence suggests it is heavily overparameterized. When the

[1] This model also allows for additional approximation error, if any, of β_{t-1} using $b_{t-1}\beta$, since such error can be absorbed into $\tilde{\varepsilon}_t$ as well.

number of firm characteristics N is large, the number of free parameters is $\{\tilde{f}_t\}$, $N \times T$, which can be large compared to sample size and thus noisily estimated.

Kelly et al. (2019) suggest a new modeling approach known as instrumented principal components analysis (or IPCA). IPCA inherits Barra's versatility and tractability, yet avoids its statistical inefficiency via a built-in dimension reduction:

$$\tilde{r}_t = b_{t-1}\beta f_t + \tilde{\varepsilon}_t, \tag{6}$$

where β and $\{f_t\}$ have $N \times K$ and $K \times T$ unknown parameters, respectively. This model of individual asset returns has a direct link with the static model for portfolios in 1. If we project b_{t-1} on both sides of 6 at each t, we obtain

$$r_t := (b_{t-1}^\intercal b_{t-1})^{-1} b_{t-1}^\intercal \tilde{r}_t = \beta f_t + u_t, \quad \text{where} \quad u_t := (b_{t-1}^\intercal b_{t-1})^{-1} b_{t-1}^\intercal \tilde{\varepsilon}_t. \tag{7}$$

This echos the static factor model 1 (which is why we use consistent notation, such as β, N, K, for both models). Moreover, $(b_{t-1}^\intercal b_{t-1})^{-1} b_{t-1}^\intercal$ can be interpreted as portfolio weights for characteristics-"sorted" portfolio returns, $(b_{t-1}^\intercal b_{t-1})^{-1} b_{t-1}^\intercal \tilde{r}_t$. This derivation is consistent with the convention of estimating (static) asset pricing models for characteristics-sorted portfolios. Indeed, IPCA incorporates stock-level and portfolio-level asset pricing in a single specification.

In general, the risk premia associated with factors $\gamma_{t-1} := \mathrm{E}_{t-1}(f_t)$ could also be time-varying, but the time-series path of risk premia $\{\gamma_{t-1}\}$ is not identifiable without additional restrictions. Only $\mathrm{E}(\gamma_{t-1})$ can be identified. To recover the path of risk premia, Gagliardini et al. (2016) employ a parametric model of risk premia, as suggested by Harvey & Ferson (1999), $\gamma_{t-1} = z_{t-1}^\intercal \theta$, where z includes macro time-series such as the term spread and θ is an unknown parameter. Combined with the assumption that factor loadings are linear functions of observed characteristics and macro time series, one can rewrite the dynamics of individual stock returns as

$$\tilde{r}_{i,t} = x_{i,t}\tilde{\beta}_i + \tilde{\varepsilon}_t, \tag{8}$$

where $\{x_{i,t}\}$ are multi-dimensional regressors that depend on observable factors, macro variables, and firm characteristics, and $\{\tilde{\beta}_i\}$ contain (functions of) unknown parameters.

In essence, IPCA and related model employ a linear approximation for risk exposures based on observable characteristics data. But there are no obvious theoretical or intuitive justifications for the linearity assumption beyond tractability. To the contrary, there are many reasons to expect that this assumption is violated. Essentially all leading theoretical asset pricing models predict nonlinearities in return dynamics as a function of state variables; Campbell & Cochrane (1999), Santos & Veronesi (2004), Bansal & Yaron (2004), and He & Krishnamurthy (2013) are prominent examples.

To overcome this limitation, Connor et al. (2012) and Fan et al. (2016) replace the assumption that factor betas are linear in characteristics with an assumption that factor betas are nonparametric functions of characteristics (although these characteristics are assumed to not vary over time for theoretical tractability). Kim et al. (2020) adopt this farmework to construct arbitrage portfolios.

Gu et al. (2021) extend the Barra and IPCA models to a nonlinear setting using a conditional autoencoder model, augmented with additional explanatory variables. It replaces the linear beta specification in 6 with a more realistic and flexible beta function. The Gu et al. (2021) autoencoder model is the first deep learning model of equity returns that explicitly accounts for the risk-return tradeoff. Thanks to recent progress in algorithms and computing power, deep learning models like this are readily available and increasingly popular among practitioners. Nevertheless, deep learning models are often criticized for their black-box nature. Although these models are comprised of simple composite functions (not much more complicated than a regression model), training them can be tedious and is sometimes more art than science. Rigorous theoretical justification still lags far behind the evolution of model architectures and training algorithms.

Continuous-time factor models can sometimes be preferable for modeling time-varying dynamics of asset returns, particularly when high-frequency returns data are available. Return factors have complex dynamics such as stochastic volatility and jumps, and individual asset returns respond to these factors with time-varying risk exposures. High-frequency modeling is well suited for tackling such complexities, though details are beyond the scope of this review. We refer interested readers to Aït-Sahalia et al. (2021).

3 METHODOLOGIES

The conventional methodologies for statistical inference of asset pricing models are designed for low-dimensional settings, e.g., 25 test assets with a handful of factors over tens of years. Recently, the set of explanatory variables (potentially) associated with equity returns has expanded rapidly (e.g. Harvey et al., 2016), and researchers have begun using individual securities as test assets (e.g. Kelly et al., 2019). With the transition to large scale sets of factors and test assets, high-dimensional statistical methods are increasingly relevant for empirical asset pricing analysis. Our survey starts with classical methods but places particular emphasis on statistical methodologies designed to cope with a high-dimensional setting.

3.1 Measuring Expected Returns

A central objective of asset pricing is to understand the behavior of expected returns. But expected returns are shrouded in noise in the form of unforecastable news that moves asset prices. This makes expected returns difficult to measure. If we can improve measurement to see expected returns more clearly, we can better devise economic theories to explain their behavior. In other words, return prediction (i.e., the measurement of expected returns) is critical to developing a clearer understanding of financial markets.

The empirical literature on stock return prediction has three basic strands. The first models differences in expected returns across stocks as a function of a small list of stock-level characteristics, and is exemplified by Fama & French (2008) and Lewellen (2015). It mostly approaches estimation via cross-sectional regression of future returns on lagged stock characteristics.[2] The second strand

[2]In addition to least squares regression, the literature often sorts assets into portfolios on the basis of characteristics

estimates expected returns via time series regression of portfolio returns on a small number of predictor variables (see surveys by Welch & Goyal, 2007; Koijen & Nieuwerburgh, 2011; Rapach & Zhou, 2013).

These traditional methods have potentially severe limitations that more advanced statistical tools in machine learning can help overcome. Most important is that regressions and portfolio sorts are ill-suited to handle the large numbers of predictor variables that the literature has accumulated over five decades. The challenge is how to assess the incremental predictive content of a newly proposed predictor while jointly controlling for the gamut of extant signals (or, relatedly, handling the problems of overfit and multiple comparisons).

The third strand of stock return predictions is newly emerging and is rooted in the methods of machine learning. With an emphasis on variable selection and dimension reduction techniques, machine learning is well suited for such challenging prediction problems by reducing degrees of freedom and condensing redundant variation among predictors. A first wave of high-dimensional models used linear methods such as partial least squares (e.g. Kelly & Pruitt, 2013; Rapach et al., 2013) and lasso (Chinco et al., 2017; Freyberger et al., 2020).

More recently, Gu et al. (2020) conduct a wide-ranging analysis of machine learning methods for return prediction, considering not only regularized linear methods but also more cutting edge nonlinear methods including random forest, boosted regression trees, and deep learning. Their research illustrates the substantial gains of incorporating machine learning when estimating expected returns. This translates into improvements in out-of-sample predictive R^2, as well as large gains for investment strategies that leverage machine learning predictions. The empirical analysis also identifies the most informative predictor variables, which helps facilitate deeper investigation into economic mechanisms of asset pricing.

Machine learning also makes it possible to improve expected return estimates using predictive information in complex and unstructured data sets. For example, Ke et al. (2019) propose a new supervised topic model for constructing return predictions from raw news text and demonstrate its prowess for out-of-sample forecasting. Jiang et al. (2021) and Obaid & Pukthuanthong (2021) demonstrate how to tap return predictive information in image data using machine learning models from the computer vision literature. Both text and image data confer particularly strong return forecasting gains at short horizons of days and weeks, and are likely underpinned by comparatively fast-moving market sentiments, rather than fundamental information that arguably plays a dominant role at forecast horizons of quarters or years. Indeed, sentiment and related behavioral economic driving forces are becoming a core aspect of financial markets research. These are subtle phenomena with circuitous transmission and feedback effects. As such, they are fertile ground for machine learning methods, which offer an ability to capture approximate complex nonlinear associations by exploiting rich and unwieldy data sets.

In general, the return prediction literature delves little into understanding the economic mechanisms (such as risk-return tradeoffs, market frictions, or behavioral biases) that may be responsible for observed predictability. Distinguishing, for example, between risk premia and mispricing requires

and studies portfolio averages—a form of nonparametric regression.

a more structured modeling approach, and factor models are the dominant tool researchers have used in this pursuit.

3.2 Estimating Factors and Exposures

In a factor model, the total variance of an asset can be decomposed into a systematic risk component driven by covariances with the factors and a component that is idiosyncratic to the asset. There are many factor modeling strategies available that differ in their assumptions about whether or not factors and their exposures are assumed known, and whether the model uses a conditional or unconditional risk decomposition.

3.2.1 TSR and CSR

For a static factor model given by 1, if factors are known, we can estimate factor exposures via asset-by-asset time series regressions, which, in matrix form, can be written as:

$$\text{TSR}: \qquad \widehat{\beta} = \bar{R}\bar{F}^{\intercal}(\bar{F}\bar{F}^{\intercal})^{-1}. \tag{9}$$

If asset returns are assumed to follow 8 as in Gagliardini et al. (2016), then asset-by-asset time-series regressions yield estimates for $\tilde{\beta}_i$, which in turn leads to estimates of parametrized factor loadings.

If factors are latent, but exposures are observable, then we can estimate factors by cross-sectional regressions at each time point. In matrix form, we can write the estimator as:

$$\text{CSR}: \qquad \widehat{F} = (\beta^{\intercal}\beta)^{-1}\beta^{\intercal}R. \tag{10}$$

This approach is most commonly used for individual stocks, for which their loadings can be proxied by firm characteristics. It is convenient for the CSR regression to accommodate time-varying characteristics as in 5, in which case we can rewrite 10 accordingly, for each t, as

$$\widehat{\tilde{f}}_t = (b_{t-1}^{\intercal}b_{t-1})^{-1}b_{t-1}^{\intercal}\tilde{r}_t. \tag{11}$$

3.2.2 PCA

If neither factors nor loadings are known, we can resort to PCA to extract latent factors and their loadings. The use of PCA in asset pricing dates back to as early as Chamberlain & Rothschild (1983) and Connor & Korajczyk (1986), and has become increasingly popular, see, e.g., Kozak et al. (2018), Pukthuanthong et al. (2019), Kelly et al. (2019), and Giglio & Xiu (2021). For a static factor model, 1, PCA can identify factors and their loadings up to some unknown linear transformation. It is more convenient to implement this via a singular value decomposition (SVD) of $\bar{R}$:

$$\bar{R} = \sum_{j=1}^{\widehat{K}} \sigma_j \varsigma_j \xi_j^{\intercal} + \widehat{U}, \tag{12}$$

where $\{\sigma_j\}$, $\{\varsigma_j\}$, and $\{\xi_j\}$ correspond to the first $\widehat{K}$ singular values, left and right singular vectors of $\bar{R}$, and $\widehat{K}$ can be any consistent estimator (e.g., Bai & Ng (2002)) of the number of factors in r_t. This decomposition yields a pair of estimates of factor innovations and exposures as

$$\widehat{V} = T^{1/2}(\xi_1 : \xi_2 : \ldots \xi_{\widehat{K}})^\intercal, \quad \widehat{\beta} = T^{-1/2}(\sigma_1\varsigma_1 : \sigma_2\varsigma_2 : \ldots : \sigma_{\widehat{K}}\varsigma_{\widehat{K}}). \tag{13}$$

Because of the fundamental indeterminacy of latent factor models, it is equivalent to use $(\widehat{\beta}H^{-1}, H\widehat{V})$ as alternative estimates, for any invertible matrix H. Said differently, a rotation of factors and an inverse rotation of betas leaves model fits exactly unchanged. While allowing for latent factors and exposures can add great flexibility to a research project, this rotation indeterminacy makes it difficult to interpret the factors in a latent factor model.

The PCA approach is also applicable if some but not all factors are observable. In such a case, Giglio et al. (2021a) suggest conducting PCA on residuals from time-series regressions of returns onto observable factors. They show that the estimated betas of observable and latent factors are, again, consistent with respect to the true betas only up to some unknown linear transformation.

3.2.3 Risk Premia PCA

The PCA-based approach extracts information about latent factors solely from realized return covariances. To see this, the SVD in 12 is applied to $\bar{R}$, which eliminates the average return from each column of R. In fact, if we assume $\alpha = 0$ in 2, the expected return is also spanned by β, so that it is possible to exploit the information in average returns ($\bar{r}$) for more efficient recovery of factors.

Lettau & Pelger (2020b) exploit this intuition and propose a so-called risk-premia PCA estimator for factors. Instead of using $T^{-1}\bar{R}\bar{R}^\intercal = T^{-1}RR^\intercal - \bar{r}\bar{r}^\intercal$, they conduct PCA on $T^{-1}RR^\intercal + \lambda\bar{r}\bar{r}^\intercal$, where λ is a tuning parameter. Risk-premia PCA generalizes the proposal of Connor & Korajczyk (1986), which corresponds to the special case of $\lambda = 0$. Lettau & Pelger (2020a) further prove that the risk-premia PCA could achieve a smaller asymptotic variance for factor loadings than the standard PCA if all factors are pervasive; and they outperform PCA empirically when factors are weak. We defer more detailed discussion on weak factors to Section 3.3.4.

3.2.4 Instrumented PCA

A limitation of PCA is that it only applies to static factor models. It also lacks the flexibility to incorporate other data beyond returns. To address both issues, Kelly et al. (2019) estimate the conditional factor model, 6, by solving the optimization problem $\min_{\beta,\{f_t\}} \sum_{t=2}^{T} \|\tilde{r}_t - b_{t-1}\beta f_t\|^2$. The estimates satisfy first-order conditions:

$$\widehat{f}_t = \left(\widehat{\beta}^\intercal b_{t-1}^\intercal b_{t-1}\widehat{\beta}\right)^{-1}\widehat{\beta}^\intercal b_{t-1}^\intercal \tilde{r}_t, \tag{14}$$

$$\mathrm{vec}(\widehat{\beta}^\intercal) = \left(\sum_{t=2}^{T} b_{t-1}^\intercal b_{t-1} \otimes \widehat{f}_t\widehat{f}_t^\intercal\right)^{-1}\left(\sum_{t=2}^{T}(b_{t-1} \otimes \widehat{f}_t^\intercal)^\intercal \tilde{r}_t\right). \tag{15}$$

Consistent with the discussion in Section 3.2.1, equation 14 shows that, given conditional betas, factors are estimated from cross section regressions of returns on betas. Equation 14 resembles 11, but the former accommodates a potentially large number of characteristics because of the built-in dimension reduction assumption. Equation 15 shows that conditional betas can be recovered from panel regressions of returns onto characteristics interacted with factors. The authors recommend an ALS algorithm to iteratively update $\widehat{\beta}$ and $\widehat{f}_t$ until convergence. Kelly et al. (2020) develop the accompanying asymptotic inference for the extracted factors and loadings.

Kelly et al. (2021) apply this IPCA framework to explain momentum and long-term reversal phenomenon in equity returns. The general framework of IPCA also extends beyond equity into other asset classes, such as corporate bonds (Kelly et al., Forthcoming) and options (Büchner & Kelly, Forthcoming).

3.2.5 Autoencoder Learning

Gu et al. (2021) introduce deep learning to return factor models by proposing a conditional autoencoder to explicitly account for the risk-return tradeoff. The machine learning literature has long recognized the close connection between autoencoders and PCA (e.g., Baldi & Hornik, 1989). However, unlike PCA, Gu et al. (2021) introduce additional conditioning information into the autoencoder specification (in the spirit of IPCA). The autoencoder allows betas to depend on stock characteristics in a more realistic, nonlinear way than IPCA's linear beta specification. Figure 1 illustrates the model's basic structure. At a high level, the mathematical representation of the model

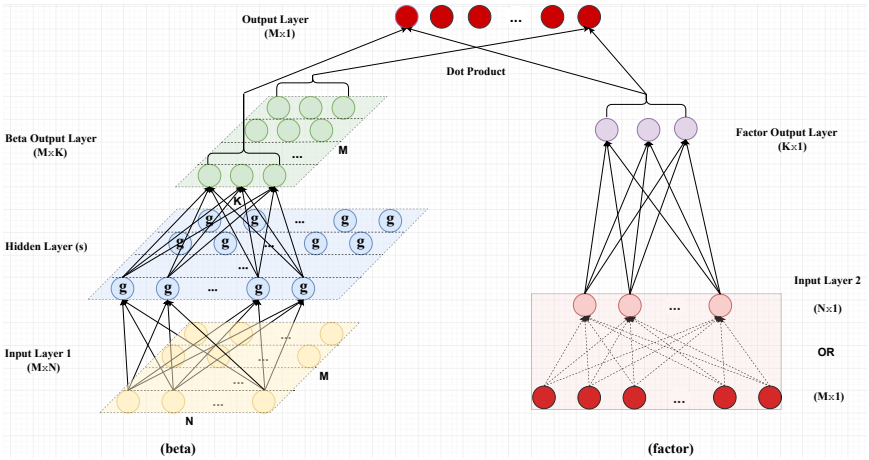

Figure 1: Conditional Autoencoder Model

is identical to equation 4. On the left side of the network, factor loadings are a nonlinear function of covariates (e.g., firm characteristics), while the right side of the network models factors as portfolios of individual stock returns.

In particular, the $K \times 1$ vector $\beta_{i,t-1}$ is specified as a neural network model of lagged firm characteristics, $b_{i,t-1}$. The recursive formulation for the nonlinear beta function is:

$$b_{i,t-1}^{(0)} = b_{i,t-1}, \tag{16}$$

$$b_{i,t-1}^{(l)} = g\left(b^{(l-1)} + W^{(l-1)} b_{i,t-1}^{(l-1)}\right), \quad l = 1, ..., L_\beta, \tag{17}$$

$$\beta_{i,t-1} = b^{(L_\beta)} + W^{(L_\beta)} b_{i,t-1}^{(L_\beta)}. \tag{18}$$

Equation 16 initializes the network as a function of the baseline characteristic data, $b_{i,t-1}$. The equations in 17 describe the nonlinear (and interactive) transformation of characteristics as they propagate through hidden layer neurons. Equation 18 describes how a set of K-dimensional factor betas emerge from the terminal output layer.

On the right side of Figure 1, we see an otherwise standard autoencoder for the factor specification. The recursive mathematical formulation of the factors is:

$$r_t^{(0)} = (b_{t-1}^\intercal b_{t-1})^{-1} b_{t-1}^\intercal r_t, \tag{19}$$

$$r_t^{(l)} = \widetilde{g}\left(\widetilde{b}^{(l-1)} + \widetilde{W}^{(l-1)} r_t^{(l-1)}\right), \quad l = 1, ..., L_f, \tag{20}$$

$$f_t = \widetilde{b}^{(L_f)} + \widetilde{W}^{(L_f)} r_t^{(L_f)}. \tag{21}$$

Equation 19 initializes the network with characteristics-sorted portfolios of individual asset returns, as defined by 7. This sidesteps the incompleteness issue of the panel of individual stock returns and in the meantime performs a preliminary reduction of data. Equations in 20 transform and compress the dimensionality of returns as they propagate through hidden layers. Equation 21 describes the final set of K factors at the output layer. If a single linear layer is included on the factor network, that is, $L_f = 1$, this structure maintains the economic interpretation of factors: they are themselves portfolios (linear combination of returns).

At last, the "dotted operation" multiplies the $M \times K$ matrix output from the beta network with the $K \times 1$ output from the factor network to produce the final model fit for each individual asset return.

When the autoencoder has one hidden layer and a linear activation function, it is equivalent to the PCA estimator for linear factor models described above. Just like the autoencoder model nests the static linear factor model, the augmented autoencoder nests the IPCA factor model as a special case. The high capacity of a neural network model enhances its flexibility to construct the most informative features from data. With enhanced flexibility, however, comes a higher propensity to overfit. We discuss several generic algorithms likely applicable to any deep learning models.

Training, Validation, and Testing To curb overfitting, the entire sample is typically divided into three disjoint subsamples that maintain the temporal ordering of the data. The first, or "training," subsample is used to estimate the model subject to a specific set of tuning hyperparameter values.

The second, or "validation," subsample is used for tuning the hyperparameters. Fitted values are constructed for data points in the validation sample based on the estimated model from the training sample. Next, the objective function is calculated based on errors from the validation sample, and hyperparameters are then selected to optimize the validation objective.

The validation sample fits are of course not truly out-of-sample because they are used for tuning,

which is in turn an input to the estimation. Thus the third, or "testing," subsample is used for neither estimation nor tuning. It is thus used to evaluate a method's out-of-sample performance.

Regularization Techniques The most common machine learning device for guarding against overfitting is to append a penalty to the objective function in order to favor more parsimonious specifications. This regularization approach mechanically deteriorates a model's in-sample performance in the hope of improving its stability out-of-sample. This will be the case when penalization manages to reduce the model's fit of noise while preserving its fit of the signal.

Gu et al. (2021) define the estimation objective to be

$$\mathcal{L}(\theta;\cdot) = \frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\left\|\tilde{r}_{i,t} - \beta_{i,t-1}' f_t\right\|^2 + \phi(\theta;\cdot), \tag{22}$$

where θ summarizes the weight parameters in the loading and factor networks of 16 through 21, $\phi(\theta)$ is a penalty function, such as lasso (or l_1) penalization, which takes the form $\phi(\theta;\lambda) = \lambda\sum_j |\theta_j|$.

In addition to l_1-penalization, Gu et al. (2021) employ a second machine learning regularization tool known as "early stopping." By ending the parameter search early (as soon as the validation sample error begins to increase), parameters are shrunken toward the initial guess, for which parsimonious parameterization is often imposed. It is a popular substitute to "l_2"-penalization of θ parameters because of its convenience in implementation and effectiveness in combatting overfit.

As a third regularization technique, Gu et al. (2021) adopt an ensemble approach in training neural networks. In particular, they use multiple random seeds to initialize neural network estimation and construct model predictions by averaging estimates from all networks. This enhances the stability of the results because the stochastic nature of the optimization can cause different seeds to settle at different optima.

Optimization Algorithms The high degree of nonlinearity and nonconvexity in neural networks, together with their rich parameterization, make brute force optimization highly computationally intensive (often to the point of infeasibility). Gu et al. (2021) adopt the adaptive moment estimation algorithm (Adam), an efficient version of stochastic gradient descent introduced by Kingma & Ba (2014) that computes adaptive learning rates for individual parameters using estimates of first and second moments of the gradients.

Gu et al. (2021) also adopt "batch normalization" (Ioffe & Szegedy, 2015), to control the variability of predictors across different regions of the network and across different datasets. It is motivated by the phenomenon of internal covariate shift in which inputs of hidden layers follow different distributions than their counterparts in the validation sample.

3.2.6 Matrix Completion

It is not uncommon in finance applications to deal with unbalanced panels. Giglio et al. (2021a) adopt a matrix completion algorithm to handle missing data when extracting factors and loadings of a factor model.

The matrix completion approach relies on an assumption that the full matrix can be written as a noisy low-rank matrix. This assumption is naturally justified for 1 (assuming $\alpha = 0$) which, in matrix form, can be rewritten as $R = \beta(V + \gamma\iota_T^\intercal) + U$ and thus clearly satisfies the assumption.

The goal is to recover an $N \times T$ low-rank matrix $X := \beta(V + \gamma\iota_T^\intercal)$. Suppose R (the "noisy version" of X) is not fully observed and Ω is an $N \times T$ matrix whose (i,t)-th element $\omega_{it} = 1_{\{r_{it} \text{ is observed}\}}$. Using this notation, econometricians can only observe $R \circ \Omega$ and Ω, where $\circ$ represents the element-wise matrix product.

The following nuclear-norm penalized regression approach can be employed to recover X:

$$\widehat{X} = \arg\min_X \|(R - X) \circ \Omega\|^2 + \lambda_{NT}\|X\|_n, \text{[3]} \tag{23}$$

where $\|X\|_n$ denotes the matrix nuclear norm and $\lambda_{NT} > 0$ is a tuning parameter. By penalizing the singular values of X, the algorithm achieves a low-rank matrix as the output. The latent factors and betas can then be estimated via the corresponding singular vectors of $\widehat{X}$.

3.3 Estimating Risk Premia

The risk premium of a factor is informative about the equilibrium compensation investors demand to hold risk associated with that factor. One of the central predictions of asset pricing models is that some risk factors for example, intermediary capital or aggregate liquidity should command a risk premium: investors should be compensated for their exposure to those factors, holding constant their exposure to all other sources of risk.

For tradable factors—such as the market portfolio in the CAPM—estimating risk premia reduces to calculating the sample average return of the factor. This estimate is simple, robust, and requires minimal modeling assumptions.

However, many theoretical models are formulated with regard to non-tradable factors—factors that are not themselves portfolios—such as consumption, inflation, liquidity, and so on. To estimate risk premia of such factors it is necessary to construct their tradable incarnations. Such a tradable factor is a portfolio that isolates the non-tradable factor while holding all other risks constant. There are two standard approaches to constructing tradable counterparts of a non-tradable factors: two-pass regressions and factor mimicking portfolios.

3.3.1 Classical Two-pass Regressions

The classical two-pass (or Fama-MacBeth) regressions requires a model like 1 with all factors observable. The first time-series (TS) pass yields estimates of β using regressions in 9. Then the second cross-sectional (CS) pass estimates risk premia via an ordinary least squares (OLS) regression of average returns on the estimated β:

$$\widehat{\gamma} = (\widehat{\beta}^\intercal\widehat{\beta})^{-1}\widehat{\beta}^\intercal\bar{r}. \tag{24}$$

[3]The *nuclear-norm* $\|X\|_n := \sum_{i=1}^{\min\{N,T\}} \psi_i(X)$, where $\psi_1(X) \geq \psi_2(X) \geq ...$ are the sorted singular values of X.

The generalized least squares (GLS) version of 24 replaces the OLS in the CS pass by

$$\widehat{\gamma}_{\text{GLS}} = (\widehat{\beta}^{\intercal}\widehat{\Sigma}_u^{-1}\widehat{\beta})^{-1}\widehat{\beta}^{\intercal}\widehat{\Sigma}_u^{-1}\bar{r}, \tag{25}$$

where $\widehat{\Sigma}_u = T^{-1}\bar{R}\mathbb{M}_{\bar{V}}\bar{R}^{\intercal}$ is the sample covariance matrix of the residuals.

Lewellen et al. (2010) advocate the GLS approach and suggest reporting GLS R^2, partially because their simulations suggest obtaining a high GLS R^2 appears to be a more rigorous hurdle than obtaining a high OLS R^2. In our view, this benefit of GLS is perhaps overshadowed by its disastrous finite sample performance due to the poor estimates of $\widehat{\Sigma}_u$, in particular when N is large. In Section 4.2, we will also show that the OLS and the infeasible GLS (which assumes perfect knowledge of Σ_u) are asymptotically equivalent when both N and T are large, so that there is no asymptotic efficiency gain from using GLS.

3.3.2 Factor Mimicking Portfolios

In contrast to equation 24, Fama & Macbeth (1973) propose an inference procedure that regresses realized returns at *each* time t onto $\widehat{\beta}$:

$$\widehat{\gamma}_t = (\widehat{\beta}^{\intercal}\widehat{\beta})^{-1}\widehat{\beta}^{\intercal}r_t. \tag{26}$$

Note that $\widehat{\gamma}_t$ is itself a portfolio return, corresponding the portfolio weights $(\widehat{\beta}^{\intercal}\widehat{\beta})^{-1}\widehat{\beta}^{\intercal}$. In other words, Fama-MacBeth regression is itself an approach to building factor-mimicking portfolios. Fama & Macbeth (1973) then recommend estimating the risk premium as the time series average of $\widehat{\gamma}_t$. Regularity conditions in Giglio & Xiu (2021) imply

$$\widehat{\gamma}_t = (\widehat{\beta}^{\intercal}\widehat{\beta})^{-1}\widehat{\beta}^{\intercal}(\beta(\gamma + v_t) + u_t) \approx \gamma + v_t.$$

and thus the Fama-MacBeth procedure is an effective approach to estimating risk premia.

Now, suppose we are interested in mimicking a risk proxy, g_t, say, a climate risk measure that is not tradable, that satisfies

$$g_t = \xi + \eta v_t + z_t. \tag{27}$$

Obviously, the risk premium of g_t is $\gamma_g = \eta\gamma$. The Fama-MacBeth procedure readily yields its factor mimicking portfolio, $\widehat{\eta}\widehat{\gamma}_t$, where $\widehat{\eta}$ is simply the vector of coefficients of a time-series regression 27, whose risk premium is precisely $\widehat{\eta}\widehat{\gamma}$.

Another standard approach to tracking a non-traded factor is the maximal-correlation factor-mimicking approach (e.g. Huberman et al., 1987; Lamont, 2001). This directly projects g_t onto a set of basis assets, y_t, which yields weights of the mimicking portfolio:

$$w_g = \text{Var}(y_t)^{-1}\text{Cov}(y_t, g_t),$$

whose returns and expected returns are given by $w_g^{\intercal}y_t$ and $w_g^{\intercal}\text{E}(y_t)$.

How do we reconcile these two approaches? Is γ_g the same as $w_g^{\intercal}\text{E}(y_t)$, for some choice of y_t? How

do we select such y_t? Under data generating processes given by 1, 2, and 27, if we select $y_t = Af_t$ (recall that $f_t = \mu + v_t$), for any invertible matrix A, then $w_g = (A^\intercal)^{-1}\eta$, and $w_g^\intercal \mathrm{E}(y) = \eta\gamma$. In this scenario, both approaches are equivalent, suggesting that we should use the same factors we use in Fama-MacBeth regressions to build a mimicking portfolio for g_t. Obviously, this is only possible if f_t is a vector of tradable portfolios. If not, we can use mimicking portfolios of f_t, $\widehat{\gamma}_t$, but this goes back to the Fama-MacBeth approach described above.

Perhaps a more interesting proposal is to use r_t directly as basis assets when building a mimicking portfolio for g_t. In this case, $w_g = \Sigma^{-1}\beta\Sigma_v\eta^\intercal$, and $w_g^\intercal \mathrm{E}(r_t) = \eta\Sigma_v\beta^\intercal(\Sigma)^{-1}\beta\gamma$. This appears to give a different risk premium parameter in population. Nevertheless, Giglio & Xiu (2021) prove that as $N \to \infty$, $w_g^\intercal \mathrm{E}(r_t)$ converges to $\eta\gamma$. This result suggests that in the limit the maximal-correlation mimicking portfolio approach targets the same risk premium parameter as Fama-MacBeth regressions.

What makes the second approach more appealing is the fact that it does not require a fully specified factor model, since all it needs are the factor of interest, g_t, and the cross-section of test assets, r_t, suggesting that estimating a factor's risk premium does not require knowledge about identities of factors driving asset returns, so it avoids the omitted factor problem. The potential issue with this approach is the curse of dimensionality—it requires a large cross-section of assets ($N \to \infty$), which may exceed the sample size, T, to the extent that a projection of g_t on r_t becomes infeasible. We now turn to a three-pass estimator that adopts PCA regression to resolve this high-dimensionality issue.

3.3.3 Three-pass Regressions

Giglio & Xiu (2021) suggest a principal component regression (PCR) when building the factor mimicking portfolios for g_t using r_t as basis assets. The PCR approach is a natural choice among high-dimensional regressions in that r_t follows a factor model according to 1. The three-pass method proceeds as follows:

1. The first-pass is an SVD of $\bar{R}$ to obtain $\widehat{\beta}$ and $\widehat{V}$ as in 13.
2. The second pass runs a cross-sectional OLS, 24, to obtain risk premia of $\widehat{V}$.
3. Finally, the third pass projects g_t onto $\widehat{V}$:

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

   thus recovering the weights of the mimicking portfolio.

The three-pass estimator of the risk premium of g_t is then obtained by multiplying the portfolio weights $\widehat{\eta}$ by the risk premia of these portfolios $\widehat{\gamma}$. In a compact form, it is given by:

$$\widehat{\gamma}_g = \bar{G}\widehat{V}^\intercal(\widehat{V}\widehat{V}^\intercal)^{-1}(\widehat{\beta}^\intercal\widehat{\beta})^{-1}\widehat{\beta}^\intercal\bar{r}. \tag{28}$$

As we discuss in Section 4.2, this estimator has asymptotic guarantees in the large N, large T setting, but only if all factors are pervasive. Since this estimator does not rely on any pre-specified asset pricing model for risk premia estimation, it provides a solution to the omitted variable bias. Relatedly, Gagliardini et al. (2019) propose a diagnostic criterion for detecting the number of omitted factors.

3.3.4 Weak Factors

Beside the omitted factor bias, another severe issue that plagues the classical two-pass regression is weak identification. Kan & Zhang (1999) first note that the inference on risk premia from two-pass regression becomes distorted when a "useless" factor – a factor to which test assets have zero exposure – is included in the model. Kleibergen (2009) further points out that standard inference fails if betas are relatively small. This issue is quite relevant in practice because many test assets are not very sensitive to macroeconomic shocks. Moreover, the same rank-deficiency problem arises when betas are collinear albeit strong, that is, some factors are redundant in terms of explaining the variation of expected returns. This is again a relevant issue in practice due to the existence of hundreds of factors discovered in the literature, many of which are close cousins and do not add any explanatory power.

When β is close to zero, its estimation error dominates the true signal, resulting in an error-in-variables problem. Kleibergen (2009) proposes several test statistics for risk premia that are uniformly valid over all values of β. The beneficial robustness of these tests come at the cost of a lack of power when weak factors exist. These tests are also designed for testing risk premia of all factors jointly, but are often not informative about the risk premium on any particular factor. Bryzgalova (2015) suggests eliminating weak factors via a penalized two-pass regression, so as to improve the power for detecting strong factors. However, eliminating weaker factors can lead to invalid inference and potentially large biases in the risk premia estimates of the remaining factors.

Jegadeesh et al. (2019) propose a sample-splitting and instrumental variable estimator to correct the error-in-variables bias. In the same spirit and perhaps more rigorously, Anatolyev & Mikusheva (2021) propose a four-split approach that addresses the issues of weak factors and omitted factors. They assume that part of v_t in equation 1, call it v_{1t}, is observable though potentially weak. Also assume that its beta, namely β_1, fully spans the space of expected returns. The other part of v_t, call it v_{2t}, is latent (hence omitted by econometricians) and unpriced. The four-split estimator aims for valid inference on the risk premia of v_{1t}. Note that omitted factors in their setup must be unpriced to achieve valid inference. But in reality it is the omitted priced factors that are most concerning.

Giglio et al. (2021b) argue that the weak factor problem is fundamentally an issue of test asset selection. They argue that factor strength is not an inherent property of a factor, instead it is dictated by the selection of test assets. Weaker factors may still be priced, so just eliminating them is an undesirable solution. Instead, Giglio et al. (2021b) suggest actively selecting test assets to guarantee that the selected assets have sufficient exposure to the factors of interest. To also address the omitted factor problem, they propose an iterative supervised PCA procedure that integrates correlation screening with the three-pass estimator. This estimator is robust to both omitted variable

bias and the weak factor problem, as well as to measurement error in observed factors.

Which Test Assets? Test assets are an important component of empirical asset pricing, yet little work has been dedicated to rigorously and systematically investigating how they should be selected. When a model is comprised of tradable factors, many important asset pricing analyses are independent of the test assets. For example, risk premia of tradable factors are best calculated as simple averages of factor returns, the maximum Sharpe ratio portfolio in the model economy can be inferred from the tradable factors alone, and model comparison can be conducted without test assets (Barillas & Shanken, 2017). In contrast, test assets are central to the study of non-tradable factors because they are used to construct the necessary factor-mimicking portfolios that in turn are inputs to most asset pricing analyses.

The choice of test assets in the literature has mainly followed one of three approaches. The first approach, adopted by the vast majority of the literature, uses a "standard" set of portfolios sorted on a few characteristics, such as size and value, following the seminal work of Fama & French (1993). Lewellen et al. (2010) argue that this approach sets a rather low hurdle for a factor pricing model. They suggest augmenting the set of test assets with industry portfolios. Giglio et al. (2021b) argue that using the standard cross-section likely creates a weak factor problem, because these assets may not have exposure to the factor of interest. Alternatively, Ahn et al. (2009) suggest forming portfolios as test assets by clustering individual securities based on their correlations so that securities within clusters are similar and those across clusters are different. There is not a clear theoretical rationale behind this proposal however.

A second approach that has gained traction more recently expands the set of test assets to include portfolios sorted on a much larger set of characteristics discovered in the last decades, on the order of hundreds of portfolios (e.g. Kozak et al., 2020; Bryzgalova et al., 2020b). Along these lines, an attractive property of IPCA is that it can be viewed and assessed from the perspective of individual stocks *or* characteristic-managed portfolios as test assets. Kelly et al. (2019) argue that this has the attractive property of reducing researcher discretion over test asset selection.

A third approach curates test assets that are targeted for a specific factor of interest (see e.g. Ang et al., 2006). A common approach is to estimate stock-level betas on a given factor, then sort assets into portfolios based on the estimated exposure. A small cross section of these sorted portfolios is expected to be particularly informative about the factor of interest, but it is affected by the omitted factor problem since it tends to focus only on univariate exposures.

3.4 Estimating the SDF and its Loadings

A factor's risk premium is equal to its (negative) covariance with the stochastic discount factor (SDF). In the setup of 1, an SDF can be written as:

$$m_t = 1 - b^\intercal v_t, \tag{29}$$

where $b = \Sigma_v^{-1}\gamma$ and Σ_v is the covariance matrix of factor innovations. The SDF is central to the field of asset pricing because, in the absence of arbitrage, covariances with the SDF unilaterally explain cross-sectional differences in expected returns.

As shown in 29, the vector of SDF loadings, b, is related to mean-variance optimal portfolio weights. SDF loadings b and risk premia γ are directly related through the covariance matrix of the factors, but they differ substantially in their interpretation. The SDF loading of a factor tells us whether that factor is useful in pricing the cross section of returns. For example, a factor could command a nonzero risk premium without appearing in the SDF simply because it is correlated with the true factors driving the SDF. It is thereby not surprising to see many factors with significant risk premia. For this reason, it makes more sense to tame the factor zoo by testing if a new factor has a non-zero SDF loading (or has a non-zero weight in the mean-variance efficient portfolio), rather than testing if it has a significant risk premium.

3.4.1 Generalized Method of Moments

The classical approach to estimating SDF loadings is the generalized method of moments (GMM). In light of 3 and 29 and the definition of the SDF, we can formulate a set of moment conditions:

$$\mathrm{E}(m_t r_t) = 0_{N\times 1}, \quad \mathrm{E}(v_t) = 0_{K\times 1}.$$

Since there are in total $K + N$ moments with $2K$ parameters (μ and b) in general, we need $N \geq K$ to ensure the system is identified.

The GMM estimator is thereby defined as the solution to the optimization problem:

$$\min_{b,\mu} \widehat{g}_T(b,\mu)^\intercal \widehat{W} \widehat{g}_T(b,\mu), \tag{30}$$

where the sample moments are given by

$$\widehat{g}_T(b,\mu) = \begin{pmatrix} \frac{1}{T}\sum_{t=1}^T r_t(1 - b^\intercal(f_t - \mu)) \\ \frac{1}{T}\sum_{t=1}^T f_t - \mu \end{pmatrix}_{(N+K)\times 1}.$$

The inference procedure follows the usual GMM formulation (Hansen, 1982). For efficiency reasons, it is customary to choose the optimal weighting matrix, $\widehat{W}_{\text{opt}} = \widehat{\Omega}^{-1}$, where $\widehat{\Omega}$ is a consistent estimator of Ω in Section 4.1. As an alternative, we discuss a special class of weighting matrices for which a closed form solution to 30 is available:

$$\widehat{b} = (\widehat{C}^\intercal \widehat{W}_{11} \widehat{C})^{-1}(\widehat{C}^\intercal \widehat{W}_{11} \bar{r}), \quad \widehat{\mu} = \bar{f}, \tag{31}$$

where $\widehat{W}_{11}$ is the top $N \times N$ sub-matrix of $\widehat{W}$, and $\widehat{C}$ is the $N \times K$ sample covariance matrix between r_t and v_t. Recall that $b = \Sigma_v^{-1}\gamma$ and note also that $\widehat{\beta}\widehat{\Sigma}_v = \widehat{C}$. It follows that $\widehat{b} = \widehat{\Sigma}_v^{-1}\widehat{\gamma}$, where $\widehat{\gamma}$ is given by 24 ($\widehat{W}_{11} = \mathbb{I}_N$) or 25 ($\widehat{W}_{11} = \widehat{\Sigma}_u^{-1}$). In other words, 31 amounts to running two-pass cross-sectional regressions with (univariate) covariances in place of $\widehat{\beta}$. This is not surprising, because

according to 2 we have

$$\mathrm{E}(r_t) = \alpha + \beta\gamma = \alpha + Cb, \tag{32}$$

where $C = \beta\Sigma_v$ is the covariance between r_t and v_t. The two-pass procedure thereby achieves an estimate of b.

3.4.2 PCA-based Methods

Kozak et al. (2018) argue that the absence of near-arbitrage opportunities forces expected returns to (approximately) align with common factor covariances, even in a world where belief distortions can affect asset prices. The strong covariation among asset returns suggests that the SDF can be represented as a function of a few dominant sources of return variation. PCA of asset returns recovers the common components that dominate return variation. Specifically, the first two passes of the three-pass procedure in Section 3.3.3 yields an SDF estimator without relying on knowledge of factor identities:

$$\widehat{m}_t = 1 - \widehat{\gamma}^\intercal \widehat{v}_t, \tag{33}$$

where $\widehat{v}_t$ is the t-th column of $\widehat{V}$.

3.4.3 Penalized Regressions

In the PCA approach, the SDF is essentially parametrized as a small number of linear combinations of factors, as shown in 29. Kozak et al. (2020) consider an SDF represented in terms of a set of tradable test asset returns:

$$\underline{m}_t = 1 - \underline{b}^\intercal (r_t - \mathrm{E}(r_t)), \tag{34}$$

where $\underline{b}$ satisfies $\mathrm{E}(r_t) = \Sigma\underline{b}$, and Σ is the covariance matrix of r_t. Giglio et al. (2021b) show that the relationship between the two SDFs (29 and 34) depends on the degree of completeness of markets. Assuming that r_t follows 1 and some regularity conditions hold, these two forms of SDF are asymptotically equivalent as $N \to \infty$ in the sense that

$$\frac{1}{T}\sum_{t=1}^{T} |m_t - \underline{m}_t|^2 \lesssim \frac{1}{\lambda_{\min}(\beta^\intercal\beta)}.$$

Since the right-hand side diminishes as $N \to \infty$ even for relatively weak factors, there is generally no theoretical difference between estimands.

To estimate the SDF (34), Kozak et al. (2020) suggest solving an optimization problem, which amounts to a regression of $\bar{r}$ onto $\widehat{\Sigma}$:

$$\widehat{\underline{b}} = \arg\min_{\underline{b}} \left\{ (\bar{r} - \widehat{\Sigma}\underline{b})^\intercal \widehat{\Sigma}^{-1} (\bar{r} - \widehat{\Sigma}\underline{b}) + p_\lambda(\underline{b}) \right\}, \tag{35}$$

with which the estimated pricing kernel is given by

$$\widehat{\underline{m}}_t = 1 - \widehat{\underline{b}}^\intercal (r_t - \bar{r}). \tag{36}$$

In the above, $\widehat{\Sigma}$ is the sample covariance matrix of r_t and $p_\lambda(\underline{b})$ is a penalty term (such as ridge, lasso, or elastic net) through which economic priors are imposed. Relatedly, Korsaye et al. (2019) provide a rigorous framework for regularization techniques in the recovery of the SDF in economies with frictions or ambiguity.

The objective function in 35 appears to require the inverse of the sample covariance matrix $\widehat{\Sigma}^{-1}$, which is not well-defined when $N > T$. Instead, it is equivalent to optimizing a different form of 35:

$$\widehat{\underline{b}} = \arg\min_{\underline{b}} \left\{ \underline{b}^\intercal \widehat{\Sigma} \underline{b} - 2\underline{b}^\intercal \bar{r} + \underline{b}^\intercal \widehat{\Sigma} \underline{b} + p_\lambda(\underline{b}) \right\}, \tag{37}$$

which avoids calculating $\widehat{\Sigma}^{-1}$.

3.4.4 Double Machine Learning

A fundamental task facing the asset pricing field today is to bring more discipline to the proliferation of factors. In particular, a question that remains open is: how to judge whether a new factor adds explanatory power for asset pricing, relative to the hundreds of factors the literature has so far produced? Feng et al. (2020) attempt to address this question by systematically evaluating the contribution of individual factors relative to existing factors as well as for conducting appropriate statistical inference in this high-dimensional setting. While machine learning methods discussed in the previous section perform well by employing regularization to tradeoff bias with variance, both regularization and overfitting cause a bias that distorts inference. Chernozhukov et al. (2018) introduce a general double machine learning (DML) framework to mitigate bias and restore valid inference on a low-dimensional parameter of interest in the presence of high-dimensional nuisance parameters. Feng et al. (2020) make use of this framework to test the SDF loading of a newly proposed factor.

Suppose that g_t is the factor of interest and h_t a vector of potentially confounding factors such that $v_t = (g_t^\intercal : h_t^\intercal)^\intercal$. To test if g_t contributes to expected returns beyond variables in h_t, we should conduct inference on b_g, while controlling b_h, where $b = (b_g : b_h)$ satisfies $\mathrm{E}(r_t) = Cb = C_g b_g + C_h b_h$ and $C = \beta \Sigma_v$ is the covariance between r_t and v_t. If the number of factors in v_t, K, is finite, then the GMM approach introduced in Section 3.4.1 is adequate. When it comes to a large K setting, the classical inference procedure is no longer valid. This is certainly a relevant case in practice, since T is typically in the hundreds, roughly of the same scale as the number of factors studied.

In the spirit of DML, Feng et al. (2020) select controls from $\{\widehat{C}_h\}$ via two respective lasso regressions: $\bar{r}$ onto $\widehat{C}_h$ and $\widehat{C}_g$ onto $\widehat{C}_h$. The selected controls, denoted by $\widehat{C}_{h[I]}$, along with $\widehat{C}_g$, serve as regressors in another cross-sectional regression of $\bar{r}$. The resulting estimator of b_g,

$$\widehat{b}_g = (\widehat{C}_g^\intercal \mathbb{M}_{\widehat{C}_{h[I]}} \widehat{C}_g)^{-1} (\widehat{C}_g^\intercal \mathbb{M}_{\widehat{C}_{h[I]}} \bar{r}),$$

is a desirable candidate for inference because the regularization biases in lasso diminish at a faster rate than $\sqrt{T}$ after partialing out the effect of $\widehat{C}_h$ from $\widehat{C}_g$.

3.4.5 Parametric Portfolios and Deep Learning SDFs

Since the SDF (when projected onto tradable assets) is spanned by optimal portfolio returns, estimating the SDF is effectively a problem of optimal portfolio formation. A fundamental obstacle to the conventional mean-variance analysis is the low signal-to-noise ratio: Expected returns and covariances of a large cross-section of investable assets cannot be learned with high precision. In the previous sections, we have discussed factor-based approaches that either exploit economic intuition and theory, or rely on statistical machine learning methods, to "regularize" this learning problem. With better estimates of expected returns and covariances come improved portfolio performance.

Brandt et al. (2009) propose an innovative solution to the portfolio optimization problem by directly parametrizing portfolio weights as functions of asset characteristics, then estimate the parameters by solving a utility optimization problem:

$$\max_{\theta} \frac{1}{T} \sum_{t=2}^{T} \mathrm{U}\left(\sum_{i=1}^{N_t} w(\theta, b_{i,t-1}) \tilde{r}_{i,t} \right),$$

where $w(\theta, b_{i,t-1})$ is a parametric function of stock characteristics, and $\mathrm{U}(\cdot)$ is some pre-specified utility function. Cong et al. (2021) extend this framework to a more flexible neural network model and optimize the Sharpe ratio of the portfolio (SDF) via reinforcement learning, with more than 50 features plus their lagged values. Chen et al. (2019) parametrize the SDF loadings and weights of test asset portfolios as two separate neural networks, and adopt an adversarial minimax approach to estimate the SDF. Both adopt Long-Short-Term-Memory (LSTM) models to incorporate lagged time series information from macro variables, firm characteristics, or past returns.

3.5 Model Specification Tests and Model Comparison

Although financial economists suggest using economic theory to winnow the best model, their efforts, unfortunately, have led to a zoo of factors and numerous candidate models. Some recent and prominent models with observable portfolios as factors include Fama & French (2015), Hou et al. (2015), Stambaugh & Yuan (2017), He et al. (2017), and Daniel et al. (2020). On the other hand, purely statistical tests are often powerless because the sample size is too limited to tease out the true model.

3.5.1 GRS Test and Extensions

Specifically, assessments of factor pricing models can be formalized as statistical hypothesis testing problems. Such tests most commonly focus on the zero alpha condition: If the factor model reflects the true SDF, then it should price all test assets with zero alpha (up to sampling variation). A

standard formulation for the null hypothesis is

$$\mathbb{H}_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_N = 0. \tag{38}$$

In a simple setting where all factors are observable and tradable, the model given by 1 and 2 can be written as $r_t = \alpha + \beta f_t + u_t$, so that alphas can be estimated via asset-wise time-series regressions:

$$\widehat{\alpha}_{\text{TS}} = (R\mathbb{M}_F\iota_T)(\iota_T^\intercal\mathbb{M}_F\iota_T)^{-1}. \tag{39}$$

Gibbons et al. (1989) construct a quadratic test statistic:

$$\widehat{F} = \frac{T-N-K}{N}\frac{\widehat{\alpha}_{\text{TS}}^\intercal\widehat{\Sigma}_u^{-1}\widehat{\alpha}_{\text{TS}}}{(1+\bar{f}^\intercal\widehat{\Sigma}_v^{-1}\bar{f})}, \tag{40}$$

and developed its exact finite sample distribution, a non-central F-distribution, under the assumption of Gaussian errors.

An important limitation of this result is that it requires that $T > N + K$. In practice, N can be much larger than T. Even in the case of $N < T$, the power of the GRS test may be compromised because it employs an unrestricted sample covariance matrix, $\widehat{\Sigma}_u$, that is known to perform badly even for moderate N. When asset returns follow an approximate factor model (Chamberlain & Rothschild, 1983), the idiosyncratic errors may be weakly correlated and thus it is possible to enhance the power of the GRS test by imposing structure on Σ_u.

Pesaran & Yamagata (2017) suggest a simple quadratic test statistic that ignores off-diagonal elements of Σ_u:[4]

$$\widehat{J}_1 = \frac{T\widehat{\alpha}_{\text{TS}}^\intercal Diag(\widehat{\Sigma}_u)^{-1}\widehat{\alpha}_{\text{TS}}(1+\bar{f}^\intercal\widehat{\Sigma}_v^{-1}\bar{f})^{-1} - N}{\sqrt{2N(1+(N-1)\widehat{\rho}_{N,T}^2)}},$$

where $\widehat{\rho}_{N,T}$ is a correction term related to the sparsity of Σ_u. This term can be omitted if Σ_u is assumed diagonal which in turn leads to a simpler test statistic.

Alternatively, Fan et al. (2015) suggest imposing a sparsity structure on Σ_u. They exploit this sparsity to achieve a consistent estimator of Σ_u, denoted as $\widehat{\Sigma}_u^{\mathcal{T}}$, following Fan et al. (2011). The resulting test statistic is

$$\widehat{J}_2 = \frac{T\widehat{\alpha}_{\text{TS}}^\intercal(\widehat{\Sigma}_u^{\mathcal{T}})^{-1}\widehat{\alpha}_{\text{TS}}(1+\bar{f}^\intercal\widehat{\Sigma}_v^{-1}\bar{f})^{-1} - N}{\sqrt{2N}}.$$

They propose other enhancements to improve the power of their test against sparse alternatives.

These extensions remedy some of the drawbacks of GRS and have asymptotic guarantees as $N, T \to \infty$, which represent an important step forward for tests of asset pricing models. Tests in this section all rely on models entirely composed of tradable factors, but in light of 46 below, the same test statistics and asymptotic inference should be directly applicable to models with non-tradable

[4] We omit finite-sample adjustment terms from the original construction of their test statistic for simplicity and clarity.

and latent factors via 44.

3.5.2 Model Comparison Tests

Testing models is perhaps less informative than comparing models. After all, all models are wrong, but some are more useful than others. As Gibbons et al. (1989) emphasize, the factor model given in 1 directly implies the following equality for the GRS test statistic:

$$\alpha^\intercal \Sigma_u^{-1} \alpha \equiv \begin{pmatrix} (\alpha+\beta\gamma)^\intercal, \gamma^\intercal \end{pmatrix} \begin{pmatrix} \beta\Sigma_v\beta^\intercal + \Sigma_u & \beta\Sigma_v \\ \Sigma_v\beta^\intercal & \Sigma_v \end{pmatrix}^{-1} \begin{pmatrix} \alpha+\beta\gamma \\ \gamma \end{pmatrix} - \gamma^\intercal \Sigma_v^{-1} \gamma. \tag{41}$$

Going one step further, $\alpha^\intercal \Sigma_u^{-1} \alpha = \text{SR}^2(\{r_t, v_t + \gamma\}) - \text{SR}^2(\{v_t + \gamma\})$, where $\text{SR}(\{a_t\})$ denotes the optimal Sharpe ratio of a portfolio using assets a_t. In other words, the classical GRS test statistic can be interpreted as a test of whether the factors achieve the maximal Sharpe ratio, or whether one can improve on that Sharpe ratio by trading the test assets in addition to the factors. Intuitively, if $\{v_t + \gamma\}$ already span the optimal portfolio (i.e., the asset pricing model is correctly specified), the Sharpe ratio gains from augmenting this portfolio with additional test assets r_t should be zero.

Indeed we can compare models using the left-hand side of 41 as a criterion. Specifically, consider two models with tradable factor sets $\{f_t^{(1)}\}$ and $\{f_t^{(2)}\}$ respectively. Barillas & Shanken (2017) advocate comparing these models on their ability to price *all* returns, both test assets and traded factors. With this perspective comes an insight that test assets tell us nothing about model comparison beyond what we learn from each model's ability to price factors of the other models! This observation is verified from 41, since

$$\alpha^{(1)\intercal} \left(\Sigma_u^{(1)}\right)^{-1} \alpha^{(1)} \;<\; \alpha^{(2)\intercal} \left(\Sigma_u^{(2)}\right)^{-1} \alpha^{(2)} \iff \text{SR}^2(\{f_t^{(1)}\}) > \text{SR}^2(\{f_t^{(2)}\}). \tag{42}$$

Barillas et al. (2020) exploit this insight and build asymptotically valid tests of model comparison using differences of squared Sharpe ratios. Their analysis allows for pairwise comparison between non-nested models and accounts for estimation error in factor-mimicking portfolio weights for non-tradable factors. Alternative criteria for model comparison also include the Hansen-Jagannathan distance by Hansen & Jagannathan (1997) (see, e.g. Kan & Robotti, 2009; Gospodinov et al., 2013) and the cross-sectional R^2 (see Kan et al., 2013).

3.5.3 Bayesian Approach

As the set of candidate models expands, model comparison via pairwise asymptotic tests becomes a daunting task. And pairwise model comparison may not unambiguously isolate the best performing model. Moreover, multiple testing issues can arise. To find the best factor pricing model, Barillas & Shanken (2018) develop a Bayesian procedure that computes model probabilities for a collection of asset pricing models with tradable factors. They adopt off-the-shelf Jeffreys prior on betas and

residual covariances, following the earlier work of Harvey & Zhou (1990):

$$\mathrm{P}(\beta, \Sigma_u) \propto |\Sigma_u|^{-(N+1)/2}.$$

Under the null hypothesis of no alpha, alpha follows a delta function concentrated at 0. Under the alternative, alpha is distributed as

$$\mathrm{P}(\alpha|\beta, \Sigma_u) = \mathcal{N}(0, k\Sigma_u), \quad \text{for some} \quad k > 0.$$

The benefit of this prior is its convenience and economic sensibility: It imposes that the expected Sharpe ratio of the "arbitrage portfolio", $\alpha^\intercal \Sigma^{-1} \alpha$, is kN, which does not take implausibly large values. Having an otherwise diffuse prior on α would force the Bayes factor to favor the null (Kass & Raftery (1995)). Barillas & Shanken (2018) provide closed-form expressions of the Bayes factor for testing zero alpha and, more importantly, of the marginal likelihood of each model. In light of Barillas & Shanken (2017), the model comparison in Barillas & Shanken (2018) is based on an aggregation of evidence from all possible multivariate regressions of excluded factors on factor subsets—i.e., it takes test assets out of the picture. Chib et al. (2020) show that the use of the standard Jeffreys priors on model-specific nuisance parameters is unsound for Bayes factors and propose a new class of improper priors for nuisance parameters based on invertible maps, which lead to valid marginal likelihoods and model comparisons.

Bryzgalova et al. (2020a) further extend the Bayesian framework for model selection in the presence of potentially weak and non-tradable factors. They re-parametrize the expected returns using equation 32, and propose a spike-and-slab prior on b to encourage model selection and ensure the validity of Bayes factors (because a flat prior would otherwise inflate Bayes factors for models that contain weak factors).

More specifically, they introduce a vector of binary latent variables $\delta = (\delta_1, \delta_2, \ldots, \delta_K)^\intercal$, where $\delta_j \in \{0, 1\}$. δ indexes 2^K possible models. The jth variable, b_j, (with associated loadings C_j) is included if and only if $\delta_j = 1$. Their prior on b has the following spike-and-slab form:

$$\begin{aligned} \mathrm{P}(b|\delta, \sigma^2) &= \prod_{j=1}^{K} (1-\delta_j)\mathrm{Dirac}(b_j) + \delta_j \mathrm{P}(b_j|\sigma^2), \quad \mathrm{P}(b_j|\sigma^2) \sim \mathcal{N}(0, \sigma^2 \psi_j); \\ \mathrm{P}(\delta|w) &= \prod_{j=1}^{K} w^{\delta_j}(1-w)^{1-\delta_j}, \quad w \sim \mathrm{P}(w); \quad \mathrm{P}(\sigma^2) \sim \sigma^{-2}. \end{aligned}$$

The Gaussian prior is used to model the non-negligible entries (the slab), and the Dirac mass at zero is used to model the negligible entries (the spike), which could be replaced by a continuous density heavily concentrated around zero. This prior, originally proposed by Mitchell & Beauchamp (1988), is known to favor parsimonious models in high dimensions, avoiding the curse of dimensionality. Another crucial component of this prior lies in their choice of ψ_j:

$$\psi_j = \psi \rho_j^\intercal \rho_j,$$

where ρ_j is an $N \times 1$ vector of correlation coefficients between factor j and the test assets, and $\psi > 0$ is a tuning parameter that controls the degree of shrinkage over all factors. If ρ_j is close to zero, the prior discourages it from being selected. This prior, however, does not seem to guard against models with highly correlated factors that cause a similar rank deficiency issue of weak factors.

Taking test assets as given, Bryzgalova et al. (2020a) aim for selecting an SDF that does not contain weak factors. The weak factors are defined in terms of C, which is similar but distinct from the definition in Section 3.3.4, in which the weak factor problem is with respect to β.[5]

3.6 Alphas and Multiple Testing

Alphas are the portion of expected returns that cannot be explained by risk exposures. Thus, a portfolio with significant alpha relative to a status quo model (e.g., the CAPM or Fama-French three-factor model) is dubbed an "anomaly." Harvey et al. (2016) collate more than 300 anomalies and argues that many of these anomalies are statistical artifacts due to data snooping or multiple testing (MT).

The literature in asset pricing has long been aware of data-snooping concerns and MT issues in alpha tests, and has taken various approaches to address it over the years. Leading examples include Lo & MacKinlay (1990) and Sullivan et al. (1999), among many others.

Early proposals suggest replacing a multitude of null hypotheses with one single null hypothesis $\mathbb{H}_0 : \max_i \alpha_i \leq 0$ or $\mathbb{H}_0 : \mathrm{E}(\alpha_i) = 0$ (see e.g. White, 2000; Kosowski et al., 2006; Fama & French, 2010). While these are interesting null hypotheses for testing, more relevant and informative hypotheses for alpha testing are perhaps

$$\mathbb{H}_0^i : \alpha_i = 0, \quad i = 1, \ldots, N. \tag{43}$$

This collection of hypotheses is fundamentally different from the single null hypothesis of GRS in 38. Suppose t_i is a test statistic for the null $\mathbb{H}_0^i$ (often taken as the t-statistic) and that a corresponding test rejects the null whenever $t_i > c_i$ for some pre-specified cutoff c_i. Let $\mathcal{H}_0 \subset \{1, ..., N\}$ denote the set of indices for which the corresponding null hypotheses are true. In addition, let $\mathcal{R}$ be the total number of rejections in a sample, and let $\mathcal{F}$ be the number of false rejections in that sample:

$$\mathcal{F} = \sum_{i=1}^{N} 1\{i \leq N : t_i > c_i \text{ and } i \in \mathcal{H}_0\}, \quad \mathcal{R} = \sum_{i=1}^{N} 1\{i \leq N : t_i > c_i\}.$$

Both $\mathcal{F}$ and $\mathcal{R}$ are random variables. Note that, in a specific sample, we can obviously observe $\mathcal{R}$, but we cannot observe $\mathcal{F}$. Nonetheless, we can design procedures to effectively limit $\mathcal{F}$ relative to $\mathcal{R}$ in expectation.

More formally, the MT literature often works with false discoveries proportion (defined as FDP $= \mathcal{F}/\max\{\mathcal{R}, 1\}$) and seeks procedures to control its expectation, known as the false discover rate (defined as FDR $= \mathrm{E}(\mathrm{FDP})$). Other objects that some MT approaches seek to control are the per-test error rate, $\mathbb{E}(\mathcal{F})/N$, and the family-wise error rate (defined as FWER $= \mathbb{P}(\mathcal{F} \geq 1)$).

[5] In related work, Gospodinov et al. (2014) take a frequentist approach to this problem.

A naïve procedure that tests each individual hypothesis at a predetermined level $\tau \in (0,1)$ guarantees that $\mathbb{E}(\mathcal{F})/N \leq \tau$. Alternatively, the Bonferroni procedure tests each hypothesis at a level τ/N, which translates into a higher t-statistic hurdle. This guarantees that $\mathbb{P}(\mathcal{F} \geq 1) \leq \tau$ and keeps the FDR below τ. Naturally, raising the hurdle for a discovery reduces the incidence of false discovery, but this also mechanically reduces the rate of true positives. In other words, false discovery control sacrifices power. FDR control procedures of Benjamini & Hochberg (1995) and Benjamini & Yekutieli (2001) attempt to strike a better balance between false discovery and power. By accepting a certain number of false discoveries, we pay a lesser price in power and thus have fewer missed discoveries. Barras et al. (2010), Bajgrowicz & Scaillet (2012), and Harvey et al. (2016) are among the first to import these statistical methods into asset pricing contexts.[6]

More recently, to obtain valid p-values and t-statistics for alphas in this context, Giglio et al. (2021a) develop a rigorous framework with asymptotic guarantees to conduct inference on alphas in linear factor models, accounting for high-dimensionality of test assets, missing data, and potentially omitted factors. Factor model presentations up to this point have imposed that alphas are zero, which makes risk premia identifiable. Giglio et al. (2021a) relax the zero-alpha assumption and impose an assumption that α is cross-sectionally independent of β (and accompany this with a large N asymptotic scheme). Their alpha estimator is given by:

$$\widehat{\alpha} = \bar{r} - \widehat{\beta}\widehat{\gamma}^{\star}, \quad \widehat{\gamma}^{\star} = (\widehat{\beta}^{\intercal}\mathbb{M}_{\iota_N}\widehat{\beta})^{-1}(\widehat{\beta}^{\intercal}\mathbb{M}_{\iota_N}\bar{r}), \tag{44}$$

where $\widehat{\beta}$ is given by 9 if all factors are observable, or 13 if factors are latent. Including an intercept term in the cross-sectional regression 44 allows for a possibly non-zero cross-sectional mean for alpha. Then p-values of $\widehat{\alpha}$ can be constructed using 46 below, which serve as inputs for FDR control.

The aforementioned frequentist MT corrections tend to be very conservative to limit false discoveries. Generally speaking, they widen confidence intervals and raise p-values, but do not alter the underlying point estimate. Jensen et al. (2021) take an empirical Bayes approach to understanding alphas in the high-dimensional context of the factor zoo, including addressing concerns about false anomaly discoveries. They propose a Bayesian hierarchical model to accomplish their MT correction, which leverages two key model attributes. First is a zero alpha prior, which imposes statistical conservatism in analogy to frequentist MT methods. It anchors alpha estimates to a sensible null in case the data are insufficiently informative about the parameters of interest. Bayesian false discovery control comes from shrinking estimates toward this prior. A benefit of the Bayesian approach, however, is the degree of FDR control decreases as data accumulates. Eventually, with enough data, the prior gets zero weight and there is no MT correction. This is justified: In the large data limit there are no false discoveries! In other words, Bayesian modeling flexibly decides on the severity of MT correction based on how much information there is in the data.

Second, the hierarchical structure in the Jensen et al. (2021) model leverages joint behavior of factors, allowing factors' alpha estimates to borrow strength from one another. As a result, alphas for different factors are shrunk not only toward zero, but also toward each other. The frequentist

[6] Harvey & Liu (2020) also propose an innovative double-bootstrap method to control FDR, while also considering false negative rate and odds ratio.

corrections above typically treat factors in isolation, making those corrections even more conservative in some cases, and those corrections always widen confidence intervals and reduce discoveries. A fascinating feature of the Bayesian hierarchical model is that jointly modeling factors can in some cases *narrow* confidence intervals. If increased precision of alphas estimates from joint estimation overshadows the discovery-reducing effect of shrinkage, the Bayesian MT approach can in fact *enhance* statistical power. In fact, Jensen et al. (2021) show in global factor return data, conservative shrinkage to the prior and improved alpha estimate precision almost exactly net out, and the number of discoveries is roughly the same as in the frequentist analysis without an MT correction.

In related work, Chen (2021) argues that it would require an absurd amount of hacking attempts for p-hacking to explain the anomaly alpha discoveries documented in the literature. More explicitly, these anomalies are broadly speaking replicable, as demonstrated by Chen & Zimmermann (2021) and Jensen et al. (2021).

4 ASYMPTOTIC THEORY

Three main asymptotic schemes have emerged in the literature for characterizing the statistical properties of factor models, risk premia, and alphas. Classical inference relies on the usual large T fixed N asymptotics. This remains the most common setup in asset pricing. The second scheme allows both N and T to increase to ∞ (with some rate restrictions). The third scheme adopts a large N fixed T design. There are pros and cons with each scheme that should be considered when conducting inference. We will illustrate this point with several examples below.

4.1 Fixed N, Large T

Under the classical scheme, Shanken (1992) developed the central limit theorem of the two-pass estimator (9 and 24). The asymptotic variances of the OLS and GLS two-pass risk premia estimators are given by

$$\text{OLS}: \quad \text{Avar}(\widehat{\gamma}) = \frac{1}{T}\left[(\beta^\intercal\beta)^{-1}\beta^\intercal\Sigma_u\beta(\beta^\intercal\beta)^{-1} \underbrace{\left(1+\gamma^\intercal(\Sigma^v)^{-1}\gamma\right)}_{\text{Shanken adjustment for } \widehat{\beta}} + \Sigma_v\right],$$

$$\text{GLS}: \quad \text{Avar}(\widehat{\gamma}) = \frac{1}{T}\left[(\beta^\intercal(\Sigma_u)^{-1}\beta)^{-1} \underbrace{\left(1+\gamma^\intercal(\Sigma_v)^{-1}\gamma\right)}_{\text{Shanken adjustment for } \widehat{\beta}} + \Sigma_v\right].$$

In the same vein, the GMM estimator of SDF loadings, $\widehat{b}$, given by 30, has an asymptotic variance:

$$\text{Avar}(\widehat{b}) = \frac{1}{T}(G^\intercal WG)^{-1}G^\intercal W\Omega WG(G^\intercal WG)^{-1},$$

where

$$W = \text{plim}_{T\to\infty}\widehat{W}, \quad G = \text{plim}_{T\to\infty}\nabla_{(b,\mu)}\widehat{g}_T(b,\mu), \quad \Omega = \lim_{T\to\infty}\text{Var}\left(\sqrt{T}\widehat{g}_T(b,\mu)\right).$$

4.2 Large N, Large T

Suppose that N is allowed to increase with T. Additionally, suppose that betas satisfy a "pervasiveness" assumption $\left\|N^{-1}\beta^\intercal\beta - \Sigma_\beta\right\| = o_{\text{P}}(1)$ for some $\Sigma_\beta > 0$ as well as a bounded eigenvalue assumption $\|\Sigma_u\| \lesssim 1$. Then we have

$$\|(\beta^\intercal\beta)^{-1}\| \lesssim N^{-1}, \quad \|\beta^\intercal\Sigma_u\beta\| \lesssim N, \quad \|(\beta^\intercal(\Sigma_u)^{-1}\beta)^{-1}\| \lesssim N^{-1}.$$

As a result, it is straightforward to show that the asymptotic variances of both OLS and (infeasible) GLS share the form:

$$\text{Avar}(\widehat{\gamma}) = T^{-1}\Sigma_v + O\left(N^{-1}T^{-1}\right). \tag{45}$$

Heuristically, we see that when N is large, there is no need to worry about estimating a large covariance matrix Σ_u or making a Shaken adjustment. Moreover, both OLS and infeasible GLS are asymptotically equivalent to the sample mean estimator $\bar{f}$ regardless of wether f is tradable or not. All these estimators achieve the same asymptotic variance, Σ_v/T. In this regard, adopting the large N large T scheme greatly simplifies the inference on γ!

Similarly, in light of the aforementioned relationship between $\widehat{\gamma}$ (equation 24) and $\widehat{b}$ (equation 31, so that $\widehat{b} = \widehat{\Sigma}_v^{-1}\widehat{\gamma}$), as well as 45, we can heuristically derive the asymptotic variance of $\widehat{b}$ for both OLS and (infeasible) GLS in the large N large T setting. Simply applying "Delta" method to the joint distribution of $\bar{v}$ and Σ_v, we have

$$\text{Avar}(\widehat{b}) = \frac{1}{T}\left[(\Sigma_v)^{-1} - 2\text{E}\left(\left((\Sigma_v)^{-1}v_t v_t^\intercal(\Sigma_v)^{-1}\right)\left(\gamma^\intercal(\Sigma_v)^{-1}v_t\right)\right) + \text{Var}\left((\Sigma_v)^{-1}v_t v_t^\intercal(\Sigma_v)^{-1}\gamma\right)\right].$$

Again, both infeasible GLS and OLS estimates of $\widehat{b}$ are asymptotically equivalent when N and T are large. But OLS is simpler, since GLS requires $(\widehat{\Sigma}_u)^{-1}$, which would be poorly estimated without additional restrictions on Σ_u.

Another blessing of high dimensionality ($N \to \infty$) is that econometricians need not know the factors' identities. Latent factors and factor exposures can be consistently recovered via SVD in 12, up to some invertible matrix H. Consequently, factor risk premia, γ, are also recoverable up to this transformation. Formally, Giglio & Xiu (2021) establish that

$$\widehat{\gamma} - H\gamma = H\bar{v} + O_{\text{P}}(N^{-1} + T^{-1}).$$

Even though these estimated factors cannot be interpreted, which is a major drawback of any latent factor model, Giglio & Xiu (2021) show that these factors serve as "controls", which facilitate the inference on $\gamma_g = \eta\gamma$, which can be identified and hence interpreted, for any factor of interest, g_t.

With respect to alphas, Giglio et al. (2021a) show that alpha estimates satisfy

$$\begin{aligned} \sigma_{i,NT}^{-1}(\widehat{\alpha}_i - \alpha_i) &\xrightarrow{d} \mathcal{N}(0,1), \\ \sigma_{i,NT}^2 &= \frac{1}{T}\mathrm{Var}(u_{it}(1 - v_t^\intercal \Sigma_v^{-1}\gamma)) + \frac{1}{N}\mathrm{Var}(\alpha_i)\frac{1}{N}\beta_i^\intercal S_\beta^{-1}\beta_i, \end{aligned} \tag{46}$$

for each $i \leq N$ as $N, T \to \infty$. Here we have $S_\beta = \frac{1}{N}\beta^\intercal \mathbb{M}_{1_N}\beta$. The second term is $O_{\mathrm{P}}(N^{-1})$, suggesting that $\widehat{\alpha}$ is inconsistent if N is finite. This formula holds whether factors are observable or latent. If $T \log N = o(N)$, the second term diminishes sufficiently fast that one only needs the first term in (46) to construct p-values for each individual alpha.

A critical assumption behind the above analysis is that all factors are pervasive. While this assumption is widely adopted in modern factor analysis (e.g. Bai, 2003) due to its simplicity and convenience, it is often in conflict with empirical evidence. If this assumption is violated, factors and their risk exposures may not be discovered by PCA.

There is a growing strand of econometrics literature on weak factor models. Bai & Ng (2008) argue that the properties of idiosyncratic errors should be considered when constructing principal components. Dropping some data, if they are noisy, may improve the forecasting. They compare the empirical performance of hard thresholding, lasso, elastic net, and least angle regressions for the selection of subsets for factor estimation (without theoretical analysis). Huang et al. (2021) propose a scaled PCA approach which incorporates information from the forecasting target into the factor extraction procedure. Bailey et al. (2020) assume a sparse structure on the loading matrix of factor exposure. Under this assumption, they propose a measure of factor strength. Freyaldenhoven (2019) proposes an estimator of the number of factors in the presence of weak factors, though the notion of "weak" factors is somewhat strong because PCA in that setting can still recover such "weak" factors consistently. Pesaran & Smith (2019) investigate the impact of factor strength and pricing error on risk premium estimation. They point out that the conventional two-pass risk premium estimator converges at a lower rate as the factors become weaker.

Lettau & Pelger (2020a) compare their risk premia PCA with the standard PCA estimator in a setting where all factors are extremely weak, so much so that they are not statistically distinguishable from idiosyncratic noise. In that case, no estimator can be consistent for either risk premia or the SDF. They show that risk premia PCA does not consistently recover the SDF, but it correlates with the SDF more so than the SDF obtained from standard PCA. Rather than focusing on this extreme case of weak factors, Giglio et al. (2021b) develop asymptotic theory covering a whole range of factor weaknesses, which permits consistent estimation of factors, risk premia, and the SDF. Formally, they allow for the case where the minimum eigenvalues of the factor component in the covariance matrix of returns diverges whereas the largest eigenvalue due to the idiosyncratic errors is bounded. In this general setup, a weak factor problem arises if and only if $N/(\lambda_{\min}(\beta^\intercal\beta)T) \nrightarrow 0$, in which case, the three-pass estimator, ridge, or PLS estimators, and the risk premia PCA estimator all give a biased risk premium estimate, but supervised PCA still works.

4.3 Large N, Fixed T

Raponi et al. (2020) propose a different asymptotic framework to estimate and test linear asset pricing models. In their setup, T is fixed yet N increases. As explained by Shanken (1992), when T is fixed, it is impossible to have a consistent estimator of risk premia. Raponi et al. (2020) therefore focuses on the so-called ex-post risk premia, defined as, $\gamma^p = \gamma + \bar{f} - \mathrm{E}(f_t)$, and establishes that the two-pass OLS estimator, after some bias-correction, converges to γ^p at the rate of $N^{-1/2}$.

Not surprisingly, their CLT provides a more accurate finite sample description of the two-pass estimator when T is small. The caveat, nonetheless, is that the estimand is dominated by factor innovations. This is because $\bar{f} - \mathrm{E}(f_t) \sim O_{\mathrm{P}}(T^{-1/2})$ is typically large relative to γ, as $\gamma/\mathrm{std}(\bar{f} - \mathrm{E}(f_t))$ is effectively the t-statistic for testing factor risk premium, which is small or insignificant unless T is large.

Zaffaroni (2019) extends this framework to allow for latent factors, providing new asymptotic analysis on PCA-based estimators of ex-post risk premia and the associated ex-post SDF. The strength of this set up is that it naturally handles time-varying factor models, where every feature is allowed to be time-varying including loadings, idiosyncratic risk and the number of risk factors.

5 CONCLUSION

Factor models have historically been the workhorse framework for empirical analysis in asset pricing. In this review, we survey the next generation of factor models with an emphasis on high-dimensional settings and the concomitant statistical tools of machine learning. Our recapitulation highlights a recent revival of (highly sophisticated) methodological research into factor modeling in asset markets. The advances and insights that have come with this revival ensure that factor models will continue to be central to empirical asset pricing in coming years.

Machine learning is neither an empirical panacea nor a substitute for economic theory and the structure it lends to empirical work. In other words, finance domain knowledge remains an indispensable component of statistical learning problems in asset markets. Indeed, our view is that the most promising direction for future empirical asset pricing research is developing a genuine fusion of economic theory and machine learning. It is a natural marriage, as asset pricing theory revolves around price formation through aggregation of investor beliefs, which undoubtedly enter prices in subtle, complex, and sometimes surprising ways. At the same time, machine learning constitutes a sophisticated quiver of statistical models that flexibly adapt to settings with rich and complex information sets.

Machine learning factor models are one such example of this fusion. Almost all leading theoretical asset pricing models predict a low-dimensional factor structure in asset prices. Where these models differ is in their predictions regarding the identity of the common factors. Much of the frontier work in empirical asset pricing can be viewed as using the (widely agreed upon) factor structure skeleton as a theory-based construct within which various machine learning schemes are injected to conduct an open-minded investigation into the economic nature of the common factors.

Our survey is inevitably selective and disproportionally influenced by our own research on these

topics. We have mainly focused on methodological contributions, leaving a detailed review of empirical discoveries via these methodologies for future work. This area of research is evolving quickly and there are myriad opportunities for improvements and new directions.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

References

Ahn DH, Conrad J, Dittmar RF. 2009. Basis assets. *The Review of Financial Studies* 22(12):5133–5174

Aït-Sahalia Y, Jacod J, Xiu D. 2021. Inference on risk premia in continuous-time asset pricing models. Tech. rep., Princeton University and the University of Chicago

Anatolyev S, Mikusheva A. 2021. Factor models with many assets: strong factors, weak factors, and the two-pass procedure. *Journal of Econometrics, forthcoming*

Ang A, Hodrick R, Xing Y, Zhang X. 2006. The cross-section of volatility and expected returns. *Journal of Finance* 61:259–299

Bai J. 2003. Inferential Theory for Factor Models of Large Dimensions. *Econometrica* 71(1):135–171

Bai J, Ng S. 2002. Determining the number of factors in approximate factor models. *Econometrica* 70:191–221

Bai J, Ng S. 2008. Forecasting economic time series using targeted predictors. *Journal of Econometrics* 146(2):304–317

Bailey N, Kapetanios G, Pesaran MH. 2020. Measurement of factor strenght: Theory and practice

Bajgrowicz P, Scaillet O. 2012. Technical trading revisited: False discoveries, persistence tests, and transaction costs. *Journal of Financial Economics* 106(3):473–491

Baldi P, Hornik K. 1989. Neural networks and principal component analysis: Learning from examples without local minima. *Neural networks* 2(1):53–58

Bansal R, Yaron A. 2004. Risks for the long run: A potential resolution of asset pricing puzzles. *The journal of Finance* 59(4):1481–1509

Barillas F, Kan R, Robotti C, Shanken J. 2020. Model comparison with sharpe ratios. *Journal of Financial and Quantitative Analysis* 55(6):1840–1874

Barillas F, Shanken J. 2017. Which alpha? *Review of Financial Studies* 30(4):1316–1338

Barillas F, Shanken J. 2018. Comparing asset pricing models. *Journal of Finance* 73(2):715–754

Barras L, Scaillet O, Wermers R. 2010. False discoveries in mutual fund performance: Measuring luck in estimated alphas. *Journal of Finance* 65(1):179–216

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 57(1):289–300

Benjamini Y, Yekutieli D. 2001. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics* :1165–1188

Brandt MW, Santa-Clara P, Valkanov R. 2009. Parametric portfolio policies: Exploiting characteristics in the cross-section of equity returns. *Review of Financial Studies* 22(9):3411–3447

Bryzgalova S. 2015. Spurious Factors in Linear Asset Pricing Models. Tech. rep., Stanford University

Bryzgalova S, Huang J, Julliard C. 2020a. Bayesian solutions for the factor zoo: We just ran two quadrillion models. Tech. rep., London Business School and London School of Economics & Political Science

Bryzgalova S, Pelger M, Zhu J. 2020b. Forest through the trees: Building cross-sections of asset returns. Tech. rep., London School of Business and Stanford University

Büchner M, Kelly BT. Forthcoming. A factor model for option returns. *Journal of Financial Economics*

Campbell JY, Cochrane JH. 1999. By force of habit: A consumption-based explanation of aggregate stock market behavior. *Journal of political Economy* 107(2):205–251

Chamberlain G, Rothschild M. 1983. Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica* 51:1281–1304

Chen AY. 2021. The limits of p-hacking: Some thoughts experiments. *Journal of Finance, forthcoming*

Chen AY, Zimmermann T. 2021. Open source cross-sectional asset pricing. *Critical Finance Review, Forthcoming*

Chen L, Pelger M, Zhu J. 2019. Deep learning in asset pricing. Tech. rep., Stanford University

Chen NF, Roll R, Ross SA. 1986. Economic forces and the stock market. *Journal of Business* 59(3):383–403

Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, et al. 2018. Double/debiased machine learning for treatment and structure parameters. *The Econometrics Journal* 21(1):C1–C68

Chib S, Zeng X, Zhao L. 2020. On comparing asset pricing models. *Journal of Finance* 75(1):551–577

Chinco A, Clark-Joseph AD, Ye M. 2017. Sparse signals in the cross-section of returns. Tech. rep., University of Illinois at Urbana-Champaign

Cong LW, Tang K, Wang J, Zhang Y. 2021. Alphaportfolio: Direct construction through deep reinforcement learning and interpretable ai. Tech. rep., Cornell University

Connor G, Hagmann M, Linton O. 2012. Efficient semiparametric estimation of the fama–french model and extensions. *Econometrica* 80(2):713–754

Connor G, Korajczyk RA. 1986. Performance measurement with the arbitrage pricing theory: A new framework for analysis. *Journal of Financial Economics* 15(3):373–394

Daniel K, Hirshleifer D, Sun L. 2020. Short- and long-horizon behavioral factors. *Review of Financial Studies* 33(4):1673–1736

Fama EF, French KR. 1993. Common risk factors in the returns on stocks and bonds. *Journal of financial economics* 33(1):3–56

Fama EF, French KR. 2008. Dissecting anomalies. *The Journal of Finance* 63(4):1653–1678

Fama EF, French KR. 2010. Luck versus skill in the cross-section of mutual fund returns. *The Journal of Finance* 65(5):1915–1947

Fama EF, French KR. 2015. A five-factor asset pricing model. *Journal of Financial Economics* 116(1):1–22

Fama EF, Macbeth JD. 1973. Risk, Return, and Equilibrium: Empirical Tests. *Journal of Political Economy* 81(3):607–636

Fan J, Liao Y, Mincheva M. 2011. High-dimensional covariance matrix estimation in approximate factor models. *Annals of Statistics* 39(6):3320–3356

Fan J, Liao Y, Wang W. 2016. Projected principal component analysis in factor models. *Annals of Statistics* 44(1):219

Fan J, Liao Y, Yao J. 2015. Power enhancement in high-dimensional cross-sectional tests. *Econometrica* 83(4):14977–1541

Feng G, Giglio S, Xiu D. 2020. Taming the factor zoo: A test of new factors. *Journal of Finance* 75(3):1327–1370

Freyaldenhoven S. 2019. A generalized factor model with local factors

Freyberger J, Neuhierl A, Weber M. 2020. Dissecting characteristics nonparametrically. *Review of Financial Studies* 33(5):2326–2377

Gagliardini P, Ossola E, Scaillet O. 2016. Time-varying risk premium in large cross-sectional equity datasets. *Econometrica* 84(3):985–1046

Gagliardini P, Ossola E, Scaillet O. 2019. A diagnostic criterion for approximate factor structure. *Journal of Econometrics* 212(2):503–521

Gibbons M, Ross SA, Shanken J. 1989. A test of the efficiency of a given portfolio. *Econometrica* 57(5):1121–1152

Giglio S, Liao Y, Xiu D. 2021a. Thousands of alpha tests. *Review of Financial Studies* 34(7):3456–3496

Giglio S, Xiu D, Zhang D. 2021b. Test assets and weak factors. Tech. rep., Yale University and University of Chicago

Giglio SW, Xiu D. 2021. Asset pricing with omitted factors. *Journal of Political Economy* 129(7):1947–1990

Gospodinov N, Kan R, Robotti C. 2013. Chi-squared tests for evaluation and comparison of asset pricing models. *Journal of Econometrics* 173(1):108–125

Gospodinov N, Kan R, Robotti C. 2014. Misspecification-Robust Inference in Linear Asset-Pricing Models with Irrelevant Risk Factors. *The Review of Financial Studies* 27(7):2139–2170

Gu S, Kelly B, Xiu D. 2020. Empirical asset pricing via machine learning. *Review of Financial Studies* 33(5):2223–2273

Gu S, Kelly BT, Xiu D. 2021. Autoencoder asset pricing models. *Journal of Econometrics* 222:429–450

Hansen LP. 1982. Large sample properties of generalized method of moments estimators. *Econometrica* 50:1029–1054

Hansen LP, Jagannathan R. 1997. Assessing specification errors in stochastic discount factor models. *Journal of Finance* 52:557–590

Harvey CR, Ferson WE. 1999. Conditioning variables and the cross-section of stock returns. *Journal of Finance* 54:1325–1360

Harvey CR, Liu Y. 2020. False (and missed) discoveries in financial economics. *Journal of Finance* 75(5):2503–2553

Harvey CR, Liu Y, Zhu H. 2016. ... and the cross-section of expected returns. *Review of Financial Studies* 29(1):5–68

Harvey CR, Zhou G. 1990. Bayesian inference in asset pricing tests. *Journal of Financial Economics* 26(2):221–254

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

He Z, Krishnamurthy A. 2013. Intermediary asset pricing. *American Economic Review* 103(2):732–70

Hou K, Xue C, Zhang L. 2015. Digesting anomalies: An investment approach. *Review of Financial Studies* 28(3):650–705

Huang D, Jiang F, Li K, Tong G, Zhou G. 2021. Scaled pca: A new approach to dimension reduction. *Management Science, forthcoming*

Huberman G. 1982. A simple approach to arbitrage pricing theory. *Journal of Economic Thoery* 28(1):183–191

Huberman G, Kandel S, Stambaugh RF. 1987. Mimicking portfolios and exact arbitrage pricing. *The Journal of Finance* 42(1):1–9

Ingersoll JE. 1984. Some results in the theory of arbitrage pricing. *Journal of Finance* 39(4):1021–1039

Ioffe S, Szegedy C. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *International Conference on Machine Learning* :448–456

Jegadeesh N, Noh J, Pukthuanthong K, Roll R, Wang J. 2019. Empirical tests of asset pricing models with individual assets: Resolving the errors-in-variable bias in risk premium estimation. *Journal of Financial Economics* 133(2):273–298

Jensen TI, Kelly B, Pedersen LH. 2021. Is there a replication crisis in finance? *Journal of Finance, forthcoming*

Jiang J, Kelly B, Xiu D. 2021. (re-)imag(in)ing price trends. Tech. rep., University of Chicago and Yale University

Kan R, Robotti C. 2009. Model comparison using the hansen-jagannathan distance. *Review of Financial Studies* 22(9):3449–3490

Kan R, Robotti C, Shanken J. 2013. Pricing model performance and the two-pass cross-sectional regression methodology. *Journal of Finance* 68(6):2617–2649

Kan R, Zhang C. 1999. Two-Pass Tests of Asset Pricing Models with Useless Factors. *The Journal of Finance* 54(1):203–235

Kass RE, Raftery AE. 1995. Bayes factors. *Journal of the American Statistical Association* 90(430):773–795

Ke T, Kelly B, Xiu D. 2019. Predicting returns with text data. Tech. rep., Harvard University and Yale University and the University of Chicago

Kelly B, Moskowitz T, Pruitt S. 2021. Understanding momentum and reversal. *Journal of Financial Economics* 140(3):726–743

Kelly B, Palhares D, Pruitt S. Forthcoming. Modeling corporate bond returns. *Journal of Finance*

Kelly B, Pruitt S. 2013. Market expectations in the cross-section of present values. *The Journal of Finance* 68(5):1721–1756

Kelly B, Pruitt S, Su Y. 2019. Characteristics are covariances: A unified model of risk and return. *Journal of Financial Economics* 134(3):501–524

Kelly B, Pruitt S, Su Y. 2020. Instrumented principal component analysis. Tech. rep., Yale University and Arizona State University and Johns Hopkins University

Kim S, Korajczyk R, Neuhierl A. 2020. Arbitrage portfolios. *Review of Financial Studies, forthcoming*

Kingma D, Ba J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*

Kleibergen F. 2009. Tests of risk premia in linear factor models. *Journal of Econometrics* 149(2):149–173

Koijen R, Nieuwerburgh SV. 2011. Predictability of returns and cash flows. *Annual Review of Financial Economics* 3:467–491

Korsaye SA, Quaini A, Trojani F. 2019. Smart sdfs. Tech. rep., University of Geneva

Kosowski R, Timmermann A, Wermers R, White H. 2006. Can mutual fund "stars" really pick stocks? New evidence from a bootstrap analysis. *The Journal of Finance* 61(6):2551–2595

Kozak S, Nagel S, Santosh S. 2018. Interpreting factor models. *Journal of Finance* 73(3):1183–1223

Kozak S, Nagel S, Santosh S. 2020. Shrinking the cross section. *Journal of Financial Economics* 135(2):271–292

Lamont OA. 2001. Economic tracking portfolios. *Journal of Econometrics* 105(1):161–184

Lettau M, Pelger M. 2020a. Estimating latent asset-pricing factors. *Journal of Econometrics* 218:1–31

Lettau M, Pelger M. 2020b. Factors that fit the time series and cross-section of stock returns. *Review of Financial Studies* 33(5):2274–2325

Lewellen J. 2015. The cross-section of expected stock returns. *Critical Finance Review* 4(1):1–44

Lewellen J, Nagel S, Shanken J. 2010. A skeptical appraisal of asset pricing tests. *Journal of Financial economics* 96(2):175–194

Lo AW, MacKinlay AC. 1990. Data-snooping biases in tests of financial asset pricing models. *Review of financial studies* 3(3):431–467

Mitchell TJ, Beauchamp JJ. 1988. Bayesian variable selection in linear regression. *Journal of American Statistical Association* 83(404):1023–1032

Obaid K, Pukthuanthong K. 2021. A picture is worth a thousand words: Measuring investor sentiment by combining machine learning and photos from news. *Journal of Financial Economics, forthcoming*

Pesaran H, Yamagata T. 2017. Testing for alpha in linear factor pricing models with a large number of securities. Tech. rep.

Pesaran MH, Smith R. 2019. The role of factor strength and pricing errors for estimation and inference in asset pricing models

Pukthuanthong K, Roll R, Subrahmanyam A. 2019. A protocol for factor identification. *Review of Financial Studies* 32(4):1573–1607

Rapach D, Zhou G. 2013. Forecasting stock returns. vol. 2 of handbook of economic forecasting, 328 383

Rapach DE, Strauss JK, Zhou G. 2013. International stock return predictability: what is the role of the united states? *The Journal of Finance* 68(4):1633–1662

Raponi V, Robotti C, Zaffaroni P. 2020. Testing beta-pricing models using large cross-sections. *Review of Financial Studies* 33:2796–2842

Rosenberg B. 1974. Extra-market components of covariance in security returns. *Journal of Financial and Quantitative Analysis* 9(2):263–274

Ross SA. 1976. The arbitrage theory of capital asset pricing. *Journal of economic theory* 13(3):341–360

Santos T, Veronesi P. 2004. Conditional betas. Tech. rep., National Bureau of Economic Research

Shanken J. 1992. On the Estimation of Beta Pricing Models. *The Review of Financial Studies* 5(1):1–33

Stambaugh RF, Yuan Y. 2017. Mispricing factors. *Review of Financial Studies* 30(4):12700–1315

Sullivan R, Timmermann A, White H. 1999. Data-snooping, technical trading rule performance, and the bootstrap. *The journal of Finance* 54(5):1647–1691

Welch I, Goyal A. 2007. A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies* 21(4):1455–1508

White H. 2000. A reality check for data snooping. *Econometrica* 68(5):1097–1126

Zaffaroni P. 2019. Factor models for asset pricing. Tech. rep., Imperial College London