Culling the Factor Zoo*

Gurdip Bakshi† Timothy Christensen‡ John Crosby§ Xiaohui Gao¶

January 14, 2025

Abstract

We propose a methodology to simultaneously select and estimate the best low-dimensional linear factor model from a high-dimensional set of candidate factors. This method selects the best subset of factors and estimates their stochastic discount factor (SDF) loadings by mixed-integer optimization. The approach is based on a sound theoretical criterion, is supported by statistical guarantees, precludes unreasonably high rewards-for-risk, and enforces positive SDFs. Portfolios based on factors selected with our best subset selection methodology yield superior out-of-sample alphas relative to standard benchmarks, and relative to dense models that use all the candidate factors.

Keywords: Asset pricing, factor models, best subset selection, mixed-integer optimization

*An earlier version of this paper was circulated under the title "Factor Glut in Asset Pricing through a Modern Optimization Lens" (SSRN # 4470335). We are grateful to Caio Almeida, Torben Andersen, Francisco Barillas, Xiaohong Chen, Lars Peter Hansen, Steve Heston, Serhiy Kozak, Pete Kyle, Oliver Linton, Mark Loewenstein, Andreas Neuhierl, Andrew Patton, Tom Schneider, Shrihari Santosh, Mobina Shafaati, Tom Smith, Dacheng Xiu, and Paolo Zaffaroni for discussions and comments. Earlier versions of the paper were presented at ODU, the University of Maryland, Temple University, UNSW, Macquarie, SOFIE 2023, and INFORMS conferences. Julia computer codes are available from the authors.

†Fox School of Business, Temple University. gurdip.bakshi@temple.edu

‡Department of Economics, Yale University. timothy.christensen@yale.edu

§Strome College of Business, Old Dominion University. acrosby@odu.edu

¶Fox School of Business, Temple University. xiaohui.gao.bakshi@temple.edu

1 Introduction

Many risk factors have been used to analyze the cross-section of stock returns (see, e.g., Cochrane (2011), Harvey, Liu, and Zhu (2016), and Harvey and Liu (2019)). Although, many of these factors show economic and statistical relevance, there is considerable debate about whether the majority of return fluctuations can be explained by a few key factors. Nevertheless, economic models that focus on a small set of factors continue to be essential tools for asset pricing, portfolio allocation, and evaluating the performance of asset managers.

Despite methodological advances for navigating the factor zoo, one key challenge remains: How can we select the best set of factors for a k-factor model (e.g., a five-, six-, or seven-factor model) from the hundreds of factors proposed so far? This paper addresses this challenge by introducing a novel methodology for selecting the optimal subset of factors to include in a linear asset pricing model. Our methodology is grounded in economic theory, leverages advances in optimization algorithms, and is backed by statistical guarantees. We support our methodology with out-of-sample empirical evidence showing that factor models chosen using our method outperform, with positive alphas relative to standard benchmarks.

Methodology for choosing the best subset of factors. Our approach is based on an economically motivated criterion. The best subset of factors and their SDF loadings are chosen by minimizing the distance between the candidate SDF and the set of admissible SDFs. We consider two distance measures: (i) a measure based on absolute deviations, which we show can be interpreted as the degree of mispricing of a set of test assets, and (ii) the squared deviation measure of Hansen and Jagannathan (1997), which measures the degree of mispricing of a portfolio with unit second moment. Either criterion results in factors that minimize in-sample pricing errors.

Two key aspects of our methodology are worth highlighting for their economic relevance. First, we incorporate bid-ask spreads to account for the practical impact of transaction costs — which can be significant when working with portfolios — and short-sale constraints.

Second, Cochrane and Saá-Requejo (2000) argue that SDFs with unrealistically high rewards-for-risk, such as very high Sharpe ratios, are unreasonable. We add a regularization to eliminate unreasonably high rewards-for-risk. For reasons of tractability, we implement this regularization through a bound on the gain-loss ratio (Bernardo and Ledoit, 2000). This approach has an additional advantage of enforcing positivity of the

estimated SDF, the importance of which is argued, for example, in Hansen and Richard (1987).

With many factors in the zoo, the number of possible k-factor models is enormous. For instance, with 500 factors there are 250 billion five-factor models and 1.5 quadrillion seven-factor models. We leverage recent advances in mixed-integer optimization (Bertsimas, King, and Mazumder, 2016) to confront the challenge of choosing the best model in this vast set. Our approach simultaneously (i) selects the best k factors to be included in the SDF and (ii) estimates their loadings. This is achieved by solving either a mixed-integer linear program (MILP) for the absolute deviation measure or a mixed-integer quadratic program (MIQP) for the squared deviation measure.

Our method eliminates the need to estimate and evaluate numerous models separately, which is computationally infeasible with many factors. Instead, we replace this exhaustive search with a single mixed-integer optimization problem that can be solved efficiently using reliable algorithms, such as those in Bertsimas and Stellato (2022). Bid-ask spreads, short-sale constraints, and regularized rewards-for-risk are all integrated into our MILP/MIQP formulation of the problem. To the best of our knowledge, ours is the first methodology to account for these aspects in a model-selection framework. Our approach also extends to conditional asset pricing models (as outlined in Appendix A.2).

Solving the MILP/MIQP yields the *best subset selection* of factors and their loadings. The best subset consists of the k factors whose combination minimizes in-sample pricing errors — as quantified through our distance measures — while accounting for bid-ask spreads, short-sale constraints, and regularized rewards-for-risk.

In addition to its economic realism and computational advantages, our methodology is supported by statistical theory. We establish consistency for both best subset selection and SDF loadings, with only weak assumptions on the data-generating process. This guarantees that our method recovers the best subset of factors in large samples.

To address potential robustness concerns inherent to model selection, we develop a *bagged* version of our methodology (Breiman, 1996). This approach maximizes the stability of factor selection by minimizing sensitivity to small data perturbations. It is also optimal for factor selection: no other factor-selection method can outperform this approach asymptotically (Christensen, Moon, and Schorfheide, 2023). Through simulations, we demonstrate that our method exhibits excellent performance in finite samples.

Empirical evidence. To determine how well our methodology can price a diverse range of test portfolios, we implement it on a set of 203 stock portfolios that range across different industries and characteristics. We use a set of 64 factors drawn from common

sources. We select the best subset of factors and estimate their SDF loadings using 26 years of monthly data from January 1972 to December 1997. We then construct mean-variance efficient portfolios from the selected factors and evaluate their performance over the 25-year period of January 1998 to December 2022. As benchmarks, we use mean-variance efficient portfolios constructed from the five-factor model of Fama and French (2015), its six-factor extension, which adds a momentum factor, and the six-factor model by Barillas and Shanken (2018).

The five-, six- and (our preferred) seven-factor models chosen using our method all generate positive alphas against all three benchmarks. These results show that our method effectively captures the cross-section of returns both in- and out-of-sample.

We also contribute to the ongoing debate on sparsity versus nonsparsity of the SDF (e.g., Feng, Giglio, and Xiu (2020), Kozak, Nagel, and Santosh (2020), and Didisheim, Ke, Kelly, and Malamud (2023)). To this end, we propose a method for estimating *dense* models that incorporate all available factors by minimizing our absolute and squared distance measures, accounting for bid-ask spreads and regularizing rewards-for-risk. As we show, these estimations may be implemented as linear and quadratic programs. We find that the out-of-sample alphas obtained from the best subset selection approach outperform those from dense models.

Related work. This paper offers a new viewpoint on factor selection, sparsity versus nonsparsity, and stability of the selected factors. The approach we propose is based on economically motivated criteria, accounts for the realism of bid-ask spreads and short-sale constraints, uses a regularization to eliminate unreasonably high rewards-for-risk, and ensures a positive SDF.

Our methodology represents a conceptual departure from recent works that compare models using Bayesian techniques, which evaluate the likelihood for returns and/or factors at the observed data and integrate against a prior (e.g., Barillas and Shanken (2018), Chib, Zeng, and Zhao (2020), and Bryzgalova, Huang, and Julliard (2023)). These Bayesian approaches do not guarantee the positivity of the SDF and do not appear compatible with regularized rewards-for-risk, bid-ask spreads, and short-sale constraints.

Furthermore, for tractability, these Bayesian approaches assume Gaussian likelihoods with independent observations and constant variance. Such assumptions overlook empirical regularities such as fat tails and persistence in volatility. As a result, these measures favor models that minimize Kullback–Leibler divergence between the assumed i.i.d. Gaussian distribution and the true data distribution (Walker, 2013). As asset returns data are not i.i.d. Gaussian, the model that minimizes Kullback–Leibler divergence may

be different from that which minimizes pricing errors. Our approach is immune from this problem, as it places no parametric assumptions on the data-generating process and is instead based on minimizing pricing errors.

Several studies, for example, Kozak et al. (2020) and Freyberger, Neuhierl, and Weber (2020), have used LASSO-type penalizations for factor selection. Unlike LASSO, our method selects relevant factors without penalizing the magnitude of the SDF loadings. This means that our methodology is robust to the scaling of factors, whereas the model selected by LASSO is sensitive to how all factors are scaled. This raises the possibility that LASSO eliminates factors with meaningful effects. Our approach is also backed by consistency and optimality guarantees for model selection, whereas there is evidence that LASSO makes important mistakes in factor selection (Feng et al., 2020).

It is important to emphasize that our focus is on model selection rather than inference, which is a different statistical problem. It is therefore complementary to recent literature on inference for factor models with many and/or weak factors (e.g., Gospodinov, Kan, and Robotti (2016), Gospodinov, Kan, and Robotti (2019), Feng et al. (2020), Giglio, Liao, and Xiu (2020), and Giglio, Xiu, and Zhang (2024)).

Our research is related to the factor model literature, which has recently received extensive empirical attention. In addition to the works cited previously, other relevant references include Asness and Frazzini (2013), Fama and French (2015), Hou, Xue, and Zhang (2015, 2020), Lettau and Pelger (2020), Daniel, Mota, Rottke, and Santos (2020), Bybee, Kelly, and Su (2023), and Preite, Uppal, Zaffaroni, and Zviadadze (2024), among others.

The remainder of the paper is organized as follows. Section 2 describes the modeling environment, introduces our methodology, and presents theoretical guarantees. In Section 3, we apply our methodology to selecting the best factors empirically. Section 4 concludes. Appendix A presents extensions of our methodology to dense models (Appendix A.1) and time-varying SDF loadings (Appendix A.2). Additionally, it provides further discussion on the regularized rewards-for-risk (Appendix A.3). Appendix B presents technical results and all proofs.

2 Methodology

Our methodology selects the best subset of factors and their corresponding SDF loadings by minimizing pricing errors. Sections 2.1 and 2.2 frame the problem at a population level and provide its economic interpretation. Implementing our methodology is straightforward: the best subset of factors and their factor loadings are estimated together by solving a single mixed-integer optimization problem presented in Section 2.3. Section 2.4 provides a formal theoretical justification for our method. Section 2.5 shows how our methodology eliminates unreasonably high rewards-for-risk. Excellent performance of our method in simulations is demonstrated in Section 2.6. Additionally, Section 2.7 shows that a bagged implementation of our method is optimal for factor selection. Finally, Section 2.8 discusses the advantages of our method relative to existing methods.

2.1 Model and Admissible SDFs

We begin by introducing the modeling environment. Time is discrete and indexed by the nonnegative integers. We use $\mathbb{E}$ to represent unconditional expectations, $\mathbb{E}_t$ to represent the expectation conditional on date-t information, and $\mathbb{P}_t$ to represent the corresponding conditional real-world probability measure. Let $\mathcal{S}$ denote the space of all measurable events, and let $s_{t+1} \in \mathcal{S}$ denote the state at date $t+1$. The state $\{s_t : t \geq 0\}$ is a stationary and ergodic stochastic process.

Consider market transactions that take place at two dates, t and $t+1$. Assets are purchased (sold) at date t, and at date $t+1$ the payoffs are received (paid out). Suppose there are K_{asset} assets whose payoffs are some function of the state s_{t+1}. We collect their payoffs in a K_{asset}-dimensional vector $\mathbf{X}_{t+1}$. The first asset is a risk-free bond with gross return $R_{f,t}$.

We incorporate bid-ask spreads to take into consideration the practicality of transaction costs, which can be significant when working with portfolios. We also allow for short-sale constraints. To incorporate these features, we denote the vector of date-t bid and ask prices by $\mathbf{p}_{t,b}$ and $\mathbf{p}_{t,a}$, respectively. We assume that the risk-free bond can be traded without transaction costs, so the first elements of $\mathbf{p}_{t,b}$ and $\mathbf{p}_{t,a}$ are $R_{f,t}^{-1}$. For assets with short-sale constraints, the corresponding element of $\mathbf{p}_{t,b}$ is zero.

An *admissible* SDF is a nonnegative random variable m_{t+1} such that

$$\mathbf{p}_{t,b} \leq \mathbb{E}_t[m_{t+1}\mathbf{X}_{t+1}] \leq \mathbf{p}_{t,a}. \tag{1}$$

Because the risk-free return is known at date t, the only uncertainty is the assets' payoffs relative to $R_{f,t}$. We therefore scale assets' payoffs by $R_{f,t}$, writing

$$\mathbf{Z}_{t+1} = \frac{\mathbf{X}_{t+1}}{R_{f,t}}, \tag{2}$$

where the first element of $\mathbf{Z}_{t+1}$ is $R_{f,t}^{-1}$. Similarly, we work with Radon–Nikodym derivatives — SDFs scaled by the gross risk-free return so that their expected value is one across all considered models — rather than SDFs. Hence, we rewrite (1) as

$$\mathbf{p}_{t,b} \leq \mathbb{E}_t[n_{t+1}\mathbf{Z}_{t+1}] \leq \mathbf{p}_{t,a}, \tag{3}$$

where

$$n_{t+1} = R_{f,t}\, m_{t+1}.$$

Let L^2 denote the space of all random variables with finite second moments under $\mathbb{E}[\,\cdot\,]$, and let L^2_+ denote all nonnegative random variables in L^2. We assume that each element of $\mathbf{Z}_{t+1}$ is in L^2. We term any $n_{t+1} \in L^2_+$ satisfying (3) an *admissible* date-t Radon–Nikodym derivative. We do not assume that markets are complete, so n_{t+1} need not be unique. We let $\mathcal{N}_{t+1}$ denote the set of all admissible date-t Radon–Nikodym derivatives. We assume that $\mathcal{N}_{t+1}$ is nonempty.

2.2 Distance Measures

Let y_{t+1} be a Radon–Nikodym derivative posited by an asset pricing model. In the next section, we specialize to factor models, but at present we keep the analysis general. An important empirical question is how closely y_{t+1} matches the set $\mathcal{N}_{t+1}$ of admissible Radon–Nikodym derivatives.

Fixing $\mathsf{q} \in \{1, 2\}$, we consider the following minimum distance problem:

$$\mathbb{D}_t^{(\mathsf{q})}(y_{t+1}) \;\equiv\; \inf_{n_{t+1} \in \mathcal{N}_{t+1}} \frac{\mathsf{q}}{2}\, \mathbb{E}_t \left[|n_{t+1} - y_{t+1}|^{\mathsf{q}}\right]^{\frac{1}{\mathsf{q}}}. \tag{4}$$

Both n_{t+1} and y_{t+1} may induce different (conditional) distributions. This leads to the interpretation of

$$\mathbb{E}_t \left[|n_{t+1} - y_{t+1}|^{\mathsf{q}}\right]^{\frac{1}{\mathsf{q}}} \tag{5}$$

as the L^{q} distance between the distributions induced by n_{t+1} and y_{t+1}. Minimizing over all n_{t+1} in $\mathcal{N}_{t+1}$ gives the *minimum* L^{q} *distance* between y_{t+1} and the set of admissible

Radon–Nikodym derivatives.

If y_{t+1} is an admissible Radon–Nikodym derivative, then $\mathbb{D}_t^{(\mathsf{q})}(y_{t+1}) = 0$. However, in general, y_{t+1} may not be admissible. Consequently, there may be pricing errors. The degree to which $\mathbb{D}_t^{(\mathsf{q})}(y_{t+1})$ differs from zero is the degree of mispricing.

We focus on the cases $\mathsf{q} = 1$ and $\mathsf{q} = 2$ for two reasons. First, these distance measures have a clear economic interpretation. Second, we can implement empirical counterparts to this problem by mixed-integer linear and quadratic programming, respectively.

Case of $\mathsf{q} = 1$: The quantity $\frac{1}{2}\mathbb{E}_t\left[|n_{t+1} - y_{t+1}|\right]$ is the *total variation distance* between the probability measures induced by n_{t+1} and y_{t+1}. Scheffé's theorem[1] yields the representation

$$\frac{1}{2}\mathbb{E}_t\left[|n_{t+1} - y_{t+1}|\right] \;=\; \sup_{A \subseteq \mathcal{S}} \Big| \mathbb{E}_t\left[n_{t+1}\,\mathbb{I}[s_{t+1} \in A]\right] - \mathbb{E}_t\left[y_{t+1}\,\mathbb{I}[s_{t+1} \in A]\right] \Big|, \tag{6}$$

where $\mathbb{I}[\,\cdot\,]$ takes the value one if its argument is true and zero otherwise. In view of (6), one may interpret total variation distance as the maximum disagreement between the probabilities induced by n_{t+1} and y_{t+1}. Equation (6) shows that $\mathbb{D}_t^{(1)}(y_{t+1})$ is bounded between 0 and 1.

We can assign a further economic interpretation to $\mathbb{D}_t^{(1)}(y_{t+1})$ using Arrow–Debreu state-prices. Recall that the first asset is a risk-free bond. Suppose the remaining assets are the Arrow–Debreu securities with payoffs $\mathbb{I}[s_{t+1} \in A_j]$, $j = 1, \ldots, K_{\text{asset}}$, where $A_2, \ldots, A_{K_{\text{asset}}}$ correspond to disjoint events. Let $\mathbb{P}_t(A_j) > 0$ for all j and $\mathbb{P}_t(\cup_{j=2}^{K_{\text{asset}}} A_j) < 1$ so that no asset is redundant. Let $\tilde{p}_{t,j} = R_{f,t}^{-1}\,\mathbb{E}_t\left[y_{t+1}\,\mathbb{I}[s_{t+1} \in A_j]\right]$ denote the date-t price of the Arrow–Debreu security implied by y_{t+1}.

We show in Appendix B that

$$\mathbb{D}_t^{(1)}(y_{t+1}) = R_{f,t} \max\left(\sum_{j=2}^{K_{\text{asset}}} \max(\tilde{p}_{t,j} - p_{t,a,j}, 0), \sum_{j=2}^{K_{\text{asset}}} \max(p_{t,b,j} - \tilde{p}_{t,j}, 0) \right). \tag{7}$$

Thus, $\mathbb{D}_t^{(1)}(y_{t+1})$ measures the maximal degree of mispricing of Arrow–Debreu securities.

Case of $\mathsf{q} = 2$: The distance $\mathbb{D}_t^{(2)}(y_{t+1})$ is the conditional Hansen and Jagannathan (1997) distance recast in terms of Radon–Nikodym derivatives. As Hansen and Jagannathan (1997) explain, their distance measure can be interpreted as a degree of mispricing

[1] See, e.g., Lemma 2.1 in Tsybakov (2009).

of a portfolio of test assets with unit second moment.

2.3 Best Subset Selection and Implementation

The objective of our methodology is to choose the best subset of factors entering the SDF and estimate their loadings. This is achieved by solving Problem 1 below.

Best Subset of Factors. We are concerned with asset pricing models whose Radon–Nikodym derivatives are a linear combination of the universe of K_{factor} factors $\mathbf{f}_{t+1}$:

$$y_{t+1}(\boldsymbol{\eta}) = \mathbf{f}_{t+1}'\boldsymbol{\eta}, \tag{8}$$

where $\boldsymbol{\eta}$ are the corresponding SDF loadings. In equation (8), we have constructed $\mathbf{f}_{t+1}$ by normalizing the "raw" factors entering the SDF by $R_{f,t}$ so that $y_{t+1}(\boldsymbol{\eta})$ is a Radon–Nikodym derivative. When the risk-free return $R_{f,t}$ is included as a raw factor, scaling by $R_{f,t}$ converts this to a constant. Henceforth, it should be understood that the constant factor represents the risk-free rate. The quantity K_{factor} denotes the number of possible factors entering the SDF. Low-dimensional models consisting of only a few factors are accommodated in (8) by setting the loadings of the excluded factors to zero.

Although the number K_{factor} of factors proposed to date is large, there are reasons why one might wish to form a parsimonious model consisting of only a few factors. To this end, let $k_{\max}$ denote the maximum number of factors to be included. We refer to the *best subset* of factors as the set of at most $k_{\max}$ factors whose Radon–Nikodym derivative minimizes a version of the distance problem in equation (4). In other words, the best subset of factors is the subset whose SDF produces smaller pricing errors than any other SDF formed from at most $k_{\max}$ factors.

Setting Up the Problem. Our empirical implementation is based on an unconditional version of equation (4). We first recast the problem in terms of returns. Define $\bar{\mathbf{p}}_t = \frac{1}{2}(\mathbf{p}_{t,a} + \mathbf{p}_{t,b})$. We construct scaled gross "returns" $\mathbf{R}_{t+1}$ by dividing $\mathbf{Z}_{t+1}$ element-wise by $\bar{\mathbf{p}}_t$. Thus, equation (3) becomes

$$\mathbf{1} - \mathbf{c}_t \leq \mathbb{E}_t[n_{t+1}\mathbf{R}_{t+1}] \leq \mathbf{1} + \mathbf{c}_t, \tag{9}$$

where $\mathbf{c}_t$ is constructed by dividing $\mathbf{p}_{a,t} - \mathbf{p}_{b,t}$ element-wise by $\mathbf{p}_{a,t} + \mathbf{p}_{b,t}$. That is,

$$\mathbf{c}_t \;\equiv\; (\mathbf{p}_{a,t} - \mathbf{p}_{b,t}) \oslash (\mathbf{p}_{a,t} + \mathbf{p}_{b,t}) \geq 0, \tag{10}$$

where $\oslash$ denotes element-wise division. Note that if any traded asset were to have a zero bid-ask spread, then the corresponding element of $\mathbf{c}_t$ would be zero. We assume that $(\mathbf{f}_t, \mathbf{R}_t, \mathbf{c}_t)$ are jointly stationary and drop time subscripts in what follows.

In view of equation (9), an admissible Radon–Nikodym derivative n must satisfy

$$\mathbf{1} - \mathbb{E}[\mathbf{c}] \leq \mathbb{E}[n\mathbf{R}] \leq \mathbf{1} + \mathbb{E}[\mathbf{c}]. \tag{11}$$

The unconditional version of the distance measure (4) is

$$Q^{(\mathsf{q})}(\boldsymbol{\eta}) \equiv \inf_{n \in L^2_+} \mathbb{E}[|n - \mathbf{f}'\boldsymbol{\eta}|^{\mathsf{q}}] \quad \text{subject to} \quad \mathbf{1} - \mathbb{E}[\mathbf{c}] \leq \mathbb{E}[n\mathbf{R}] \leq \mathbf{1} + \mathbb{E}[\mathbf{c}]. \tag{12}$$

The interpretation of $Q^{(\mathsf{q})}(\boldsymbol{\eta})$ is analogous to $\mathbb{D}_t^{(\mathsf{q})}$. The best subset of $k_{\max}$ factors corresponds to a model whose SDF loadings $\boldsymbol{\eta}$ minimize $Q^{(\mathsf{q})}(\boldsymbol{\eta})$ in equation (12) subject to the sparsity constraint

$$\sum_{k=1}^{K_{\text{factor}}} \mathbb{I}[\eta_k \neq 0] \;\leq\; k_{\max}. \tag{13}$$

This constraint ensures that at most $k_{\max}$ elements of $\boldsymbol{\eta}$ are nonzero. Coefficients that are nonzero correspond to the best subset of factors, whereas coefficients that are zero correspond to discarded factors.

Empirical Implementation. The data consist of a time series of factors and returns $(\mathbf{f}_t, \mathbf{R}_t, \mathbf{c}_t)_{t=1}^T$. We estimate the best subset of factors and their SDF loadings by solving a sample version of the problem in equation (12). This problem can be recast as a mixed-integer optimization.

To do so, we build on the approach of Bertsimas et al. (2016) for regression models. We fold the sparsity constraint (13) into the objective function by introducing binary variables χ_k for each factor. Each χ_k indicates whether the corresponding η_k is nonzero ($\chi_k = 1$), in which case the factor is selected, or zero ($\chi_k = 0$), in which case the factor

is discarded. Let $\bar{\mathbf{c}}_T = \frac{1}{T}\sum_{t=1}^{T}\mathbf{c}_t$ and $\boldsymbol{\chi} = (\chi_k)_{k=1}^{K_{\text{factor}}}$. Define

$$
\begin{aligned}
Q^{(\mathsf{q})}_{T,k_{\max}}(\boldsymbol{\eta}) \equiv \min_{\boldsymbol{\chi},\{n_t\}_{t=1}^{T}} \; & \frac{1}{T}\sum_{t=1}^{T}|n_t - \mathbf{f}_t'\boldsymbol{\eta}|^{\mathsf{q}} \\
\text{subject to} \quad & \mathbf{1} - \bar{\mathbf{c}}_T \leq \tfrac{1}{T}\textstyle\sum_{t=1}^{T} n_t\mathbf{R}_t \leq \mathbf{1} + \bar{\mathbf{c}}_T, \\
& n_t \geq 0, && t = 1,\ldots,T, \\
& \chi_k \in \{0,1\}, && k = 1,\ldots,K_{\text{factor}}, \\
& -\mathsf{b}\chi_k \leq \eta_k \leq \mathsf{b}\chi_k, && k = 1,\ldots,K_{\text{factor}}, \\
& \textstyle\sum_{k=1}^{K_{\text{factor}}}\chi_k \leq k_{\max}.
\end{aligned}
\tag{14}
$$

The constraint $-\mathsf{b}\chi_k \leq \eta_k \leq \mathsf{b}\chi_k$ reduces to either $\eta_k \in [-\mathsf{b}, \mathsf{b}]$ when $\chi_k = 1$ or $\eta_k = 0$ when $\chi_k = 0$, where b is a positive constant. In the latter case of $\chi_k = 0$, it has the effect of discarding the kth factor. For values of $\boldsymbol{\eta}$ that violate the sparsity constraint, the problem in equation (14) has no solution and the objective value is $Q^{(\mathsf{q})}_{T,k_{\max}}(\boldsymbol{\eta}) = +\infty$. The constant b can be set to be arbitrarily large but finite.

Implementation is operationalized by augmenting the problem in (14) with a vector of slackness variables $(a_t)_{t=1}^{T}$. Each a_t measures the size of the deviation $n_t - \mathbf{f}_t'\boldsymbol{\eta}$. To simultaneously select the best subset of factors and estimate their SDF loadings, we solve the following problem:

Problem 1 (Best Subset Selection and Estimation of SDF Loadings)

$$
\begin{aligned}
\min_{\boldsymbol{\eta},\boldsymbol{\chi},\{n_t,a_t\}_{t=1}^{T}} \; & \frac{1}{T}\sum_{t=1}^{T} a_t^{\mathsf{q}} \\
\text{subject to} \quad & \mathbf{1} - \bar{\mathbf{c}}_T \leq \tfrac{1}{T}\textstyle\sum_{t=1}^{T} n_t\mathbf{R}_t \leq \mathbf{1} + \bar{\mathbf{c}}_T, \\
& n_t \geq 0, && t = 1,\ldots,T, \\
& -a_t \leq n_t - \mathbf{f}_t'\boldsymbol{\eta} \leq a_t, && t = 1,\ldots,T, \\
& \chi_k \in \{0,1\}, && k = 1,\ldots,K_{\text{factor}}, \\
& -\mathsf{b}\chi_k \leq \eta_k \leq \mathsf{b}\chi_k, && k = 1,\ldots,K_{\text{factor}}, \\
& \textstyle\sum_{k=1}^{K_{\text{factor}}}\chi_k \leq k_{\max}.
\end{aligned}
\tag{15}
$$

Let $\hat{\boldsymbol{\eta}}_{\mathsf{q},k_{\max}}$ and $\hat{\boldsymbol{\chi}}_{\mathsf{q},k_{\max}}$ denote the values of $\boldsymbol{\eta}$ and $\boldsymbol{\chi}$ that solve Problem 1. The constraint $\sum_{k=1}^{K_{\text{factor}}}\hat{\chi}_k \leq k_{\max}$ enforces that a maximum of $k_{\max}$ elements of $\hat{\boldsymbol{\eta}}_{\mathsf{q},k_{\max}}$ and $\hat{\boldsymbol{\chi}}_{\mathsf{q},k_{\max}}$ are nonzero. Thus, $\mathbf{f}'\hat{\boldsymbol{\eta}}_{\mathsf{q},k_{\max}}$ depends on a maximum of $k_{\max}$ factors. We term these factors the *best subset selection*. Equivalently, the best subset selection consists of the factors for which the corresponding $\hat{\chi}_k$ are nonzero. The nonzero elements of $\hat{\boldsymbol{\eta}}_{\mathsf{q},k_{\max}}$ are estimates of the SDF loadings on the selected factors.

We close this subsection by discussing computation. When $\mathsf{q} = 1$, Problem 1 is a mixed-integer linear program (MILP), whereas when $\mathsf{q} = 2$, Problem 1 is a mixed-integer quadratic program (MIQP). Fast, robust algorithms for solving both these types of optimization problems now exist in standard scientific computing environments. In the simulations and empirical application, we implemented our methods in `Julia` using the optimization package `JuMP` to interface with the `Gurobi` optimizer.

2.4 Consistency of the Selected Factors and their Loadings

We now present our main theoretical result, which establishes consistency of the selected factors and estimated loadings obtained by solving Problem 1.

To set the stage, recall that $\hat{\boldsymbol{\eta}}_{\mathsf{q},k_{\max}}$ and $\hat{\boldsymbol{\chi}}_{\mathsf{q},k_{\max}}$ are the values of $\boldsymbol{\eta}$ and $\boldsymbol{\chi}$ that solve Problem 1. Let $\boldsymbol{\eta}_{\mathsf{q},k_{\max}}$ denote the vector of SDF loadings that minimizes the distance measure $Q^{(\mathsf{q})}$ in (12) subject to the sparsity constraint (13). The nonzero elements of $\boldsymbol{\eta}_{\mathsf{q},k_{\max}}$ represent the SDF loadings for the best subset of factors. Let $\boldsymbol{\chi}_{\mathsf{q},k_{\max}}$ denote the binary vector whose entries indicate whether or not the corresponding factor is included in the best subset.

Theorem 1 *Suppose that Assumptions 2 and 3 in Appendix B hold. Then for any $\epsilon > 0$,*

$$\Pr\left(Q^{(\mathsf{q})}(\hat{\boldsymbol{\eta}}_{\mathsf{q},k_{\max}}) > Q^{(\mathsf{q})}(\boldsymbol{\eta}_{\mathsf{q},k_{\max}}) + \epsilon\right) \to 0 \tag{16}$$

as the sample size T becomes large. If, in addition, $\boldsymbol{\eta}_{\mathsf{q},k_{\max}}$ uniquely minimizes $Q^{(\mathsf{q})}$ subject to the sparsity constraint (13), then

$$\Pr(\hat{\boldsymbol{\chi}}_{\mathsf{q},k_{\max}} = \boldsymbol{\chi}_{\mathsf{q},k_{\max}}) \to 1 \tag{17}$$

and

$$\Pr\left(\left\|\hat{\boldsymbol{\eta}}_{\mathsf{q},k_{\max}} - \boldsymbol{\eta}_{\mathsf{q},k_{\max}}\right\| > \epsilon\right) \to 0 \tag{18}$$

as the sample size T becomes large.

Theorem 1 asserts that (i) the best subset selection recovers the best subset of factors with probability approaching one and (ii) the corresponding SDF loading estimates are consistent for the loadings on the model formed from the best subset of factors. This formally justifies the use of our methodology as a tool for both factor selection and estimation of SDF loadings. The proof of Theorem 1 is presented in Appendix B.

2.5 Ruling Out Unreasonably High Rewards-for-Risk

Cochrane and Saá-Requejo (2000) argue that SDFs with unrealistically high rewards-for-risk, such as very high Sharpe ratios, are unreasonable. To address this, we now show how our methodology can eliminate excessively high rewards-for-risk.

For tractability, we bound the gain-loss ratio rather than the Sharpe ratio. Bernardo and Ledoit (2000) show that unreasonably high gain-loss ratios can be avoided by placing a bound, which we denote b^{glr}, on the ratio of the highest to lowest values taken by the SDF across all states. As we show in more detail in Appendix A.3, the constraint

$$\frac{1}{\sqrt{\mathsf{b}^{\text{glr}}}} \leq n_t \leq \sqrt{\mathsf{b}^{\text{glr}}}, \qquad t = 1, \ldots, T, \tag{19}$$

can be added as a regularization to rule out unreasonably high rewards-for-risk. The value of $\mathsf{b}^{\text{glr}} \geq 1$ controls the maximum allowable Sharpe ratio, with higher b^{glr} implying less regularization and allowing higher Sharpe ratios. For further details on the mapping between b^{glr} and Sharpe ratios, see Appendix A.3.

To simultaneously select the best subset of factors and estimate the factors' SDF loadings, allowing for regularized rewards-for-risk, solve the following problem.

Problem 2 (Best Subset Selection with Regularized Rewards-for-Risk)

$$\begin{aligned}
\min_{\boldsymbol{\eta}, \boldsymbol{\chi}, \{n_t, a_t\}_{t=1}^T} \; & \frac{1}{T}\sum_{t=1}^{T} a_t^{\mathsf{q}} \\
\text{subject to} \quad & \mathbf{1} - \bar{\mathbf{c}}_T \leq \tfrac{1}{T}\textstyle\sum_{t=1}^{T} n_t \mathbf{R}_t \leq \mathbf{1} + \bar{\mathbf{c}}_T, \\
& -a_t \leq n_t - \mathbf{f}_t'\boldsymbol{\eta} \leq a_t, && t = 1, \ldots, T, \\
& \chi_k \in \{0, 1\}, && k = 1, \ldots, K_{\text{factor}}, \\
& -\mathsf{b}\chi_k \leq \eta_k \leq \mathsf{b}\chi_k, && k = 1, \ldots, K_{\text{factor}}, \\
& \textstyle\sum_{k=1}^{K_{\text{factor}}} \chi_k \leq k_{\max}, \\
& \tfrac{1}{\sqrt{\mathsf{b}^{\text{glr}}}} \leq n_t \leq \sqrt{\mathsf{b}^{\text{glr}}}, && t = 1, \ldots, T, \\
& \tfrac{1}{\sqrt{\mathsf{b}^{\text{glr}}}} \leq \mathbf{f}_t'\boldsymbol{\eta} \leq \sqrt{\mathsf{b}^{\text{glr}}}, && t = 1, \ldots, T, \\
& \tfrac{1}{T}\textstyle\sum_{t=1}^{T} \mathbf{f}_t'\boldsymbol{\eta} = 1.
\end{aligned} \tag{20}$$

Let $\hat{\boldsymbol{\eta}}_{\mathsf{q}, k_{\max}}$ denote the value of $\boldsymbol{\eta}$ that solves Problem 2. By adding the constraint

$$\frac{1}{\sqrt{\mathsf{b}^{\text{glr}}}} \leq \mathbf{f}_t'\boldsymbol{\eta} \leq \sqrt{\mathsf{b}^{\text{glr}}} \tag{21}$$

to Problem 2, we ensure that $y_{t+1}(\hat{\boldsymbol{\eta}}_{\mathsf{q}, k_{\max}}) = \mathbf{f}_{t+1}'\hat{\boldsymbol{\eta}}_{\mathsf{q}, k_{\max}}$ avoids unreasonably high

rewards-for-risk and is strictly positive. Additionally, the constraint

$$\frac{1}{T}\sum_{t=1}^{T}\mathbf{f}_t'\boldsymbol{\eta} = 1 \tag{22}$$

forces the estimated model-implied Radon–Nikodym derivative $y_{t+1}(\hat{\boldsymbol{\eta}}_{\mathsf{q},k_{\max}})$ to have unit sample mean. By stationarity of $\mathbf{f}_t$, this ensures that $y_{t+1}(\hat{\boldsymbol{\eta}}_{\mathsf{q},k_{\max}})$ has approximately unit expectation.

For Problem 2 to have a solution, $\mathfrak{b}^{\text{glr}}$ must be chosen large enough that there exist $(n_t)_{t=1}^T$ that satisfy $\mathbf{1} - \bar{\mathbf{c}}_T \leq \frac{1}{T}\sum_{t=1}^T n_t\mathbf{R}_t \leq \mathbf{1} + \bar{\mathbf{c}}_T$ and the condition $\frac{1}{\sqrt{\mathfrak{b}^{\text{glr}}}} \leq n_t \leq \sqrt{\mathfrak{b}^{\text{glr}}}$ for $t = 1, \ldots, T$. The smallest value of $\mathfrak{b}^{\text{glr}}$ that is compatible with these conditions is found by solving

$$\begin{aligned} \min_{\mathfrak{b},\{n_t\}_{t=1}^T} \quad & \mathfrak{b} \qquad \text{subject to} \quad \mathbf{1} - \bar{\mathbf{c}}_T \leq \tfrac{1}{T}\textstyle\sum_{t=1}^T n_t\mathbf{R}_t \leq \mathbf{1} + \bar{\mathbf{c}}_T, \\ & \qquad\qquad\qquad\qquad \tfrac{1}{\sqrt{\mathfrak{b}}} \leq n_t \leq \sqrt{\mathfrak{b}}, \quad t = 1, \ldots, T. \end{aligned} \tag{23}$$

Let $\underline{\mathfrak{b}}$ denote the minimizing value of $\mathfrak{b}$. Problem 2 has a solution if $\mathfrak{b}^{\text{glr}} \geq \underline{\mathfrak{b}}$.

2.6 Simulations Reveal Good Factor Selection Properties

In this section, we present a simulation exercise to evaluate the effectiveness of our method in identifying which factors are relevant and estimating their SDF loadings.

Our baseline simulation is conducted with $K_{\text{factor}} = 9$ (including the constant) and $K_{\text{asset}} = 48$ (resembling 48 industry portfolios). We set the bid-ask spread to zero for all assets. The first asset is a risk-free bond with a gross return $R_{f,t} = 1$. We model the gross return $R_{i,t}$ of asset $i = 2, \ldots, K_{\text{asset}}$ as

$$R_{i,t} \;=\; 1 + \mu_i\,\Delta + \sum_{k=2}^{K_{\text{factor}}} \beta_{i,k}\,f_{k,t} + \sigma_i^R\,\sqrt{\Delta}\,\epsilon_{i,t}^R, \tag{24}$$

with a time step of $\Delta = 1/12$ representing monthly intervals. The first factor is a constant. Specifically, $f_{1,t} = 1$ for all t. Additionally, all other factors have zero mean. We model the remaining factors $f_{k,t}$ for $k = 2, \ldots, K_{\text{factor}}$ as

$$f_{k,t} = \sigma_k^f\sqrt{\Delta}\,\epsilon_{k,t}^f. \tag{25}$$

Finally, we consider a Radon–Nikodym derivative given by

$$n_t \;=\; \eta_1\,f_{1,t} \;+\; \sum_{k=2}^{K_{\text{factor}}} \eta_k\,f_{k,t} \;+\; \sigma^{\perp}\,\sqrt{\Delta}\,\epsilon_t^{\perp}. \tag{26}$$

The shocks $\epsilon_{i,t}^R$, $\epsilon_{k,t}^f$ and $\epsilon_t^\perp$ are all $N(0,1)$ random variables. They are uncorrelated over time, but $\epsilon_{k,t}^f$ and $\epsilon_{\ell,t}^f$ have correlation $\rho_{k,\ell}^f$ for $k \neq \ell$. All other shocks are independent. The shocks $\epsilon_t^\perp$ are orthogonal to all factors and returns and are included to make the estimation more challenging.

Our aim is to determine whether our approach can reliably identify the relevant factors and estimate the associated SDF loadings. The parameters in equations (24)–(26) are chosen as follows. We set $\eta_8 = 0$ and $\eta_9 = 0$, implying that factors 8 and 9 are irrelevant. Additionally, we normalize $\eta_1 = 1$. The values of the remaining η_k, as well as σ_i^R, $\beta_{i,k}$, σ_k^f, $\sigma^\perp$, and $\rho_{k,\ell}^f$ vary *across* simulations, as follows:

1. The values of SDF loadings η_k, for $k \in \{2,3,4,5,6,7\}$ are drawn from a uniform distribution supported from -1.2 to -0.3.
2. Additionally, σ_i^R is uniformly distributed between 30% and 45%.
3. The $\beta_{i,k}$ coefficients in equation (24) are drawn from a uniform distribution supported between 0.05 and 0.3 for each $i = 2, \ldots, K_{\text{asset}}$ and $k = 2, \ldots, K_{\text{factor}}$.
4. Each correlation $\rho_{k,\ell}^f$ is uniformly distributed between 0.05 and 0.45. The parameter σ_k^f is uniformly distributed between 25% and 35%.
5. Finally, the parameter $\sigma^\perp$ is uniformly distributed between 8% and 12%.

The remaining parameters μ_i for $i = 2, \ldots, K_{\text{asset}}$ (see equation (24)), which represent expected returns, are obtained by solving

$$\mathbb{E}[(\eta_1 + \sum_{k=2}^{K_{\text{factor}}} \eta_k \, f_{k,t} + \sigma^\perp \sqrt{\Delta} \epsilon_t^\perp)(1 + \mu_i \, \Delta + \sum_{k=2}^{K_{\text{factor}}} \beta_{i,k} \, f_{k,t} + \sigma_i^R \sqrt{\Delta} \, \epsilon_{i,t}^R)] \;=\; 1, \qquad (27)$$

so that $\mathbb{E}[n_{t+1}\mathbf{R}_{t+1}] = \mathbf{1}$. By using equation (27) to determine μ_i, we enforce that, for each simulation, n is an admissible Radon–Nikodym derivative. Across simulations the minimum value of n_t is strictly positive.

Our exercise is based on $5,000$ simulations, each of which has $T = 360$ (monthly) time steps. We perform factor selection and estimation of SDF loadings by solving Problem 2 as an MILP (with $\mathsf{q} = 1$) and an MIQP (with $\mathsf{q} = 2$).

We compare results with a dense model that uses all $K_{\text{factor}} = 9$ factors with regularized rewards-for-risk when $\mathsf{q} = 1$ and $\mathsf{q} = 2$ (for details, see Problem 3 in Section 3.4). The case $\mathsf{q} = 2$ corresponds to estimation based on minimizing the Hansen and Jagannathan (1997) distance with positivity and regularized rewards-for-risk imposed.

One can map gain-loss ratios to Sharpe ratios, following Bernardo and Ledoit (2000). For example, a gain-loss ratio of 9 corresponds to a maximum allowable annualized Sharpe ratio of 4 (see the analysis in Appendix A.3). We also drop this gain-loss ratio regularization and compare results with estimation based on minimizing the Hansen and Jagannathan (1997) distance with positivity only.

Table I (Panels A and B) presents the mean absolute errors (MAEs) and MAE/(mean $|\eta_k|$), for each $k = 1, \ldots, K_{\text{factor}}$, between the estimated and true $\boldsymbol{\eta}$ across simulations and across the five approaches. In Panel C of Table I, we report the percentage of simulations in which each factor is correctly identified. This exercise is pertinent because the best subset selection method has the potential to eliminate irrelevant factors.

The best subset selection method (Panel A) proves advantageous for correctly identifying factors as relevant or irrelevant. Factors 8 and 9 are correctly identified as irrelevant in more than 99.92% of simulations while factors 2 through 7 are correctly identified as relevant in over 99.96% of simulations. Moreover, factor 1 — the constant factor — is correctly identified in all cases.

The MAEs for the irrelevant factors are essentially zero. This is true whether we use $\mathsf{q} = 1$ or $\mathsf{q} = 2$. For the remaining seven factors, the MAEs with best subset selection are smaller than those without. Estimations based on regularized rewards-for-risk produces virtually identical MAE to estimation without regularized rewards-for-risk. This suggests that the regularization of rewards-for-risk via the gain-loss ratio does not materially bias the SDF loading estimates in small samples.

In summary, our proposed methodology achieves excellent factor selection properties, successfully detecting irrelevant factors with more than 99.92% accuracy.

2.7 Optimal Factor Selection

Theorem 1 shows that our methodology consistently selects the best subset of factors. We now describe a bagged implementation of our method that is *optimal* for factor selection.

Our goal is to select the best $k_{\max}$-factor model from the set of all such models we could form from the universe of K_{factor} factors. Model selection methods can be sensitive to data perturbations. We mitigate this sensitivity by identifying the best subset of factors under small perturbations of the data and then choosing the subset that is selected for the majority of perturbations. This approach is based on the concept of bagging (Breiman (1996)). The resulting set of factors selected is inherently more stable, since

it minimizes sensitivity to perturbations of the data. Christensen et al. (2023) develop a general theory for problems such as these and show that model selection decisions performed in this way are optimal, including in scenarios where identification fails.

To implement this approach, we recommend the following bootstrap strategy, which we use in our empirical work. First, we partition the data into a number of consecutive blocks of length $B \geq 1$, similar to the block bootstrap method. Next, we use the multiplier or Bayesian (Rubin (1981)) bootstrap: We generate random weights drawn from an exponential distribution with mean 1 for each block and replace the time-series averages in Problems 1 and 2 by block-weighted averages, scaling the weights so they sum to one. We then solve Problems 1 and 2 for different draws of random weights. The set of factors that appears most frequently across the draws is chosen as the best subset selection.

2.8 Advantages of our Methodology

We close by highlighting several advantages of our method when compared to the existing literature. First, it is based on economically motivated criteria and it takes into account the impact of bid-ask spreads and short-sale constraints, which are important in practice. Our method excludes unrealistically high rewards-for-risk and guarantees a positive SDF, providing a more economically grounded basis for factor selection. Furthermore, unlike other methods that use LASSO-type penalizations on SDF loadings to achieve sparsity, our approach is not affected by the relative scaling of individual factors and avoids a regularization bias. Overall, our method offers a robust approach to culling the factor zoo.

Our method is backed by statistical theory with consistency and optimality guarantees for factor selection. Because our method is based on minimizing measures of mispricing, these guarantees do not require correct specification of the sparse factor model, nor do they rely on unrealistically strong assumptions on the data-generating process. If the true model is a factor model with at most $k_{\max}$ factors, then our method will correctly select the true subset of relevant factors and consistently estimate the true SDF loadings whether implemented with $\mathsf{q} = 1$ or $\mathsf{q} = 2$.

As with other distance-based estimation strategies, if the $k_{\max}$-factor model is misspecified, our method is consistent for the "pseudo-true" parameters that minimize the distance measure (12) subject to the sparsity constraint and other constraints. In this case, the pseudo-true parameters are those that minimize pricing errors. This is not the

case for model selection based on Bayesian methods, which will instead favor models that minimize Kullback–Leibler divergence between the assumed parametric specification and the true distribution of returns and/or factors.

Our approach also has several distinct computational advantages. It does not require searching through all possible $k_{\max}$-factor models, which is computationally infeasible, but rather relies on a mixed-integer optimization. Relevant factors and their loadings are estimated in a single step. It also avoids numerical instabilities associated with having to compute the inverse covariance matrix of a large number of asset returns and/or factors. Furthermore, our approach extends to conditional asset pricing models with time-varying SDF loadings (see Appendix A.2).

Thus far we have focused on sparse models based on the best subset of the factors. We show in Section 3.4 that our method extends to estimating *dense* models that incorporate *all* available factors. Factor loadings can be estimated by solving a linear or quadratic program, allowing for bid-ask spreads, short-sale constraints, regularized rewards-for-risk, and ensuring positivity of the SDF. Our empirical results in Section 3.4 also contribute to the ongoing debate about sparsity versus nonsparsity of the SDF.

3 Empirical Evidence

In this section, we provide empirical evidence on the best subset of factors. We estimate the best subset of factors and the factors' SDF loadings, ensuring that the resulting SDF is always positive and eliminating SDFs with unreasonably high rewards-for-risk. We further document properties of the resulting SDFs, including the identity of the best subset factors.

Our analysis reveals four central findings:

1. Models estimated using our method demonstrate positive out-of-sample alpha compared to standard benchmarks.

2. The models we selected generate realistic out-of-sample Sharpe ratios.

3. Models selected using our method, which involve a small number of factors, have out-of-sample performance that is superior to dense models involving all factors.

4. There is a conflict between selecting the best subset of factors jointly or sequentially, with the former consistently producing higher out-of-sample alphas.

3.1 Data and Implementation

To assess the effectiveness of our best subset selection approach, we conduct empirical tests using 203 stock portfolios as test assets. These portfolios cover 17 different industries, include 150 portfolios categorized by size and six other characteristics (book-to-market, momentum, accruals, beta, variance, and net issuance), 35 portfolios sorted by operating profitability and investment, and the returns on one-month risk-free Treasury bonds. Our study spans 51 years, from January 1972 to December 2022, comprising 612 monthly observations. Our goal is to determine how well our best subset selection methodology can price this diverse range of test portfolios.

We perform our analysis using $K_{\text{factor}} = 64$ factors, including a constant representing the risk-free rate, as listed in Table II. This set of factors comprises (i) five factors from the study of Fama and French (2015); (ii) three factors from Hou, Xue, and Zhang (2015, 2020); (iii) five factors from Daniel, Mota, Rottke, and Santos (2020, Table 8); (iv) three factors from AQR, including the $\text{HML}^{\text{devil}}$ factor used in the analysis of Barillas and Shanken (2018); and (v) 47 factors considered by Jensen, Kelly, and Pedersen (2023). Our design excludes newer factors that have absolute cross-correlation higher than 0.8 with existing factors, as these repackage existing sources of risk.

We select the best subset of factors and estimate their SDF loadings using 26 years of data, from January 1972 to December 1997, then evaluate performance, as measured by alpha relative to standard benchmarks, over the 25-year period of January 1998 to December 2022. We maintain the assumption that factor loadings are constant between the in- and out-of-sample periods, as in the empirical studies of Giglio, Liao, and Xiu (2020) and Lettau and Pelger (2020), among others.

We present results from two implementations of our methodology:

1. We select factors using the bagged implementation of our method described in Section 2.7, solving Problem 2 with $\mathsf{q} = 1$ (MILP) using 200 bootstraps with block length $B = 1$. The best subset selection is the collection of factors that appear together most often across bootstrap draws. SDF factor loadings are estimated in-sample and then fixed at those estimates for the out-of-sample period.

2. As above, but we now set $\mathsf{q} = 2$ (MIQP) to select factors and estimate their SDF loadings.

In both implementations we set $\mathsf{b}^{\text{glr}} = 9$, which is equivalent to capping the maximum annualized Sharpe ratio at four (see Appendix A.3). We also set the value of $\bar{\mathbf{c}}_T$ at 0.0028.

This value corresponds to a bid-ask spread of 0.56% of the mid-price. Our approach is consistent with the median bid-ask spread observed in studies that consider a variety of stocks traded on the NYSE, AMEX, and NASDAQ (e.g., Corwin and Schultz (2012, Table III)).

The approach to identifying the best subset selection of factors allows for flexibility in the choice of $k_{\max}$, that is, the maximum number of factors selected. We are motivated by empirical research on asset pricing that has identified numerous stock characteristics that are useful in predicting variations in the cross-section of expected returns. In order to simplify the complexity of these characteristics and summarize their impact, researchers have developed factor models that incorporate a small number of key characteristics. As new predictors continue to emerge, these models have been updated and expanded in response to new evidence. For example, workhorse factor models have been extended to include up to six factors. Our method brings discipline to this line of research, by determining which combination — in addition to the constant factor representing the risk-free rate, which is always selected — of seven, six, or five risk factors minimizes pricing errors in the cross-section.

3.2 Best Subset Selection Yields Superior Alpha

To set the stage for our analysis, we perform an out-of-sample exercise that demonstrates superior out-of-sample alpha of models selected using our method relative to standard benchmark models. In doing so, we provide evidence that the factors selected using our methodology effectively describe the return cross-section both in- and out-of-sample. As such, our findings show that our approach offers an effective solution to the factor zoo.

As explained by Cochrane (2005, Chapter 5) and Kozak, Nagel, and Santosh (2020, page 275), there is a one-to-one equivalence between the loadings on a linear SDF and the weights of the mean-variance efficient (MVE) portfolio. We calculate the excess returns of the MVE portfolio associated with the SDF implied by the best subset selection as

$$r^{\text{mve}}_{t+1} = \sum_{k=2}^{K_{\text{factor}}} \hat{\eta}_k \, f_{k,t+1}, \tag{28}$$

with the understanding that $\hat{\eta}_k = 0$ for any factor k that is not selected by our method. We then compare the out-of-sample excess returns of the MVE portfolio to those from three benchmarks: (i) the five-factor model of Fama and French (2015), (ii) the six-factor model of Barillas and Shanken (2018), and (iii) the five-factor model of Fama and French (2015) augmented to six factors by the addition of the momentum factor. We denote

these benchmarks FF5, BS6, and FF5+MOM, respectively. To ensure a comparison on an equal footing, we construct MVE portfolios for each of these benchmarks by estimating the loadings in the in-sample period and holding them fixed in the out-of-sample period. Following Kozak, Nagel, and Santosh (2020, page 290), we scale each MVE portfolio to have the same volatility as excess returns on the CRSP market factor.

Table III presents the alphas expressed in annualized percentage units calculated from the time-series regression of the excess returns of the SDF-implied MVE portfolio compared to those of the benchmarks for the out-of-sample period. These are estimated as the intercept in an OLS regression of the (scaled) returns of the MVE portfolio on the (scaled) returns on the benchmark portfolio for the out-of-sample period.

The alphas reported in Table III validate our method, showing that best subset selection SDF delivers positive out-of-sample alphas relative to all benchmarks, regardless of whether the $\mathsf{q} = 1$ or $\mathsf{q} = 2$ estimator is used and whether $k_{\max} = 6$, 7, or 8. Alphas relative to the FF5 model exceed 4% for five-, six-, or seven-factor models selected with our method. Alphas relative to other benchmarks are less dramatic but remain positive.

Below the alphas in Table III, we also report in parentheses the fraction of bootstrap draws for which the alphas are negative. We compute these by bootstrapping the out-of-sample returns data but holding the selection of factors and estimated factor loadings fixed. We may interpret these as p-values of individual hypothesis tests of a null of a non-positive alpha for each model against each benchmark, conditioning on the chosen set of factors and loadings.[2] These results indicate evidence of a positive alpha (conditional on the chosen model) for our six- and preferred seven-factor models against the FF5 and BS6 benchmarks, and weakly positive alphas relative to the FF5+MOM benchmark.

Taken all together, the results reported in Table III provide evidence to support our view that best subset selection models effectively capture the cross-section of expected stock returns both in- and out-of-sample. Their outperformance relative to all three benchmarks attests to the robustness of our approach, which produces five-, six-, and seven-factor models that yield positive alphas.

[2]Because the SDF loading estimator $\hat{\boldsymbol{\eta}}_{\mathsf{q},k_{\max}}$ is the solution to a linear or quadratic program subject to inequality constraints, $\hat{\boldsymbol{\eta}}_{\mathsf{q},k_{\max}}$ may not be asymptotically normally distributed and standard bootstrap confidence intervals may not have correct coverage (Shapiro, 1991; Dümbgen, 1993). It is difficult to compute non-conservative p-values that take into account uncertainty in the selected factors and estimated factor loadings. Therefore, we condition on the selected factors and estimated loadings.

3.3 Composition of the Best Subset

We now turn to analyzing the content of the factor models chosen using our method. Table IV shows that the five- and six-factor models selected using our method consist of the same factors with $\mathsf{q} = 1$ and $\mathsf{q} = 2$. Both of these models outperform all three benchmarks, with higher out-of-sample alphas for the six-factor model (see Table III). Additionally, the estimated SDF loadings are nearly identical, with differences only up to the last decimal point.

The seven-factor model selected with $\mathsf{q} = 2$ generates even higher out-of-sample alphas against all three benchmarks than the five- and six-factor models selected using our method. The MVE portfolio associated with this model achieves an annualized out-of-sample Sharpe ratio of 0.76, which exceeds those of the six- and five-factor models (see Table IV).

The composition of risk factors for this model also seems preferred on economic grounds to the seven-factor model with $\mathsf{q} = 1$. Table IV lists the factors of our preferred seven-factor model with $\mathsf{q} = 2$ as (1) SMB, (2) CMA, both from Fama and French (2015), (3) $\text{Q}^{\text{expected growth}}$ from the work of Hou, Xue, and Zhang (2015, 2020), (4) MKT^{*} proposed by Daniel, Mota, Rottke, and Santos (2020), (5) SMB^{*} the hedged version of Daniel, Mota, Rottke, and Santos (2020), (6) $\text{MOM}^{t-6:t-1}$ the momentum factor of Jegadeesh and Titman (1993), and (7) $\text{MKT}^{\text{equity}}$ of Banz (1981). These factors are considered the best combination of seven factors to explain the return cross-section according to our method.

We form intervals by estimating the factor loadings across bootstrap samples and taking the 2.5 and 97.5 percentiles of the loading estimates across draws. None of these intervals contain zero. Although this is not a formal hypothesis test of significance of the coefficients for the reasons explained in footnote 2, it does suggest that the coefficients of the selected factors differ meaningfully from zero. We then repeat the process, taking the minimum and maximum estimate across draws. There are only three factors, indicated by the absence of a bullet in Table IV, for which these intervals bracket zero.

According to Fama and French (2015), an effective way to explain expected stock returns is to consider the sensitivity to returns on five specific portfolios. These portfolios are the overall market (MKT), differences between small and large stocks (SMB), value and growth stocks (HML), robust and weak profitability (RMW), and low and high investment behavior (CMA).[3] However, our analysis suggests that the stock return cross-

[3]The size and value factors are widely acknowledged as key components in finance, accounting,

section is better explained by a different combination of five factors, which share only a few in common with those of Fama and French (2015).

One robust finding is the importance of a momentum factor. The momentum factor $\text{MOM}^{t-6:t-1}$ is selected by our five-, six- and preferred seven-factor models. This provides a partial explanation for why our models generate highest out-of-sample alphas against the FF5 benchmark, which does not feature a momentum factor, but smaller alphas against the BS6 and FF5+MOM benchmarks, both of which do.

3.4 Evidence on Sparsity

The previous evidence leaves unanswered two questions: (i) Do dense models featuring all factors perform better than sparse models estimated using our best subset selection method? and (ii) When does the complexity of factor models have a negative effect on out-of-sample alpha?

To address these questions, we now introduce an approach for estimating dense models that include all factors, while accommodating bid-ask spreads, short-sale constraints, regularized rewards for risk, and a strictly positive SDF. We formulate this estimation problem as a linear or quadratic program and provide statistical guarantees for it. We compare the performance — as measured by out-of-sample alphas — of models estimated using our best subset selection method with that of dense models that use all K_{factor} factors.

The estimation of SDF loadings for a dense model with regularized rewards-for-risk is accomplished by solving the following problem:

and mutual fund research. However, our analysis reveals that the HML factor is not frequently chosen in best subset analyses. This suggests that the impact of the HML factor is likely overshadowed by other factors, regardless of the factor structure. Our chosen models also demonstrate less reliance on factors like the Olson O-score, Altman Z-score, and Amihud measure, which are not perceived as highly reliable and such weak factors are eliminated in our selection process. Many other known factors appear dominated by our best subset and cannot be justified on empirical grounds.

Problem 3 (Dense Factor Models with Regularized Rewards-for-Risk)

$$
\begin{aligned}
\min_{\boldsymbol{\eta},\{n_t,a_t\}_{t=1}^T} \frac{1}{T}\sum_{t=1}^T a_t^{\mathsf{q}} & \\
\text{subject to} \quad & \mathbf{1}-\bar{\mathbf{c}}_T \leq \tfrac{1}{T}\textstyle\sum_{t=1}^T n_t\mathbf{R}_t \leq \mathbf{1}+\bar{\mathbf{c}}_T, & \\
& -a_t \leq n_t - \mathbf{f}_t'\boldsymbol{\eta} \leq a_t, & t=1,\ldots,T, \\
& \tfrac{1}{\sqrt{\mathsf{b}^{\text{glr}}}} \leq n_t \leq \sqrt{\mathsf{b}^{\text{glr}}}, & t=1,\ldots,T, \\
& \tfrac{1}{\sqrt{\mathsf{b}^{\text{glr}}}} \leq \mathbf{f}_t'\boldsymbol{\eta} \leq \sqrt{\mathsf{b}^{\text{glr}}}, & t=1,\ldots,T, \\
& \tfrac{1}{T}\textstyle\sum_{t=1}^T \mathbf{f}_t'\boldsymbol{\eta} = 1. &
\end{aligned}
\tag{29}
$$

In comparison with Problem 2, we drop the sparsity constraint and omit the indicator variables χ_k. Theorem 2 in Appendix A.1 provides statistical guarantees for SDF loadings estimated using this method.

The approach and set of constraints in Problem 3 differs from that of Kozak, Nagel, and Santosh (2020), who estimate dense models by minimizing HJ distance plus a ridge penalty (proportional to the sum of squared SDF loadings), which acts as a regularization. We instead impose an economic regularization by constraining the gain-loss ratio and accounting for bid-ask spreads. An advantage of regularizing via the gain-loss ratio is that the SDF estimated by solving Problem 3 is guaranteed to be positive, the SDF based on HJ distance with a ridge regularization cannot be guaranteed to be positive.

A further advantage of estimating SDF loadings by solving Problem 3 is that it avoids inverting a large sample covariance matrix of asset returns and/or factors. As a result, numerical instabilities associated with having a large number of assets and/or factors are avoided. This aspect gives our approach an advantage in high-dimensional settings over HJ distance and other methods designed for environments with a small to medium number of test assets and/or factors.[4]

We estimate the dense model using the same data as previously, namely 64 factors and 203 assets over 26 years of monthly observations, from January 1972 to December 1997. We then compute alphas relative to the FF5, BS6, and FF5+MOM benchmarks over the out-of-sample period of January 1998 to December 2022. Results are reported in Table V.

[4]Carrasco and Nokho (2024) show that for the HJ distance there is a formal mathematical equivalence between relaxing the equation $\mathbb{E}[n_{t+1}\mathbf{R}_{t+1}] = \mathbf{1}$ to an inequality and adding a ridge regularization to the inverse covariance matrix of returns. Thus, accommodating bid-ask spreads, while economically motivated, may also help to further avoid numerical instabilities associated with having a large number of assets and/or factors.

It is evident from Table V that the five-, six- and the preferred seven-factor models chosen using our best subset selection approach yield superior out-of-sample alpha to the dense model. Indeed, the dense model produces out-of-sample alphas of 1.7%, 1.3%, and −0.3% against the FF5, BS6, and FF5+MOM benchmarks, respectively. The dense model with 64 factors also produces a smaller out-of-sample Sharpe ratio than our preferred seven-factor model. We repeat the exercise with a subset of $K_{\text{factor}} = 50$ factors, listed in Table II. This has the effect of increasing slightly the out-of-sample alphas and Sharpe ratio relative to the 64-factor model, although the results still fall substantially short of our sparse five-, six- and seven-factor models.

In summary, we find no evidence that having a dense model that includes a large number of factors improves out-of-sample performance as measured by out-of-sample alphas relative to standard benchmarks or out-of-sample Sharpe ratios.

3.5 Sequential Best Subset Selection

The factors chosen through the best subset selection method show a pattern of robustness across $\mathsf{q} = 2$ and $\mathsf{q} = 1$, as seen in Tables III and IV. However, there are instances where a factor is present in the best five- or six-factor model but is excluded from the seven-factor model. This raises the question of whether it is prudent to select the best factors *sequentially*. In other words, we ask: If a factor is selected in the J-factor model, can we ensure that it will also be selected in the $(J + 1)$-factor model? A second question we ask is whether this approach would yield better empirical results out-of-sample.

This sequential feature is incorporated in our methodology by iterating Problem 2, increasing the value of $k_{\max}$ in each iteration. As we progress, we modify the constraint $\chi_k \in \{0, 1\}$ in display (20) of Problem 2 to be $\chi_k = 1$ for all factors that have already been selected in previous iterations. Thus, at each iteration, the optimization involves only a single nonzero integer rather than $k_{\max}$ nonzero integers.

Results for sequential selection are presented in Table VI. In this approach, the constant factor (representing the risk-free rate) is always chosen and the nonconstant factors are selected in a specific order, as shown in Table VI. There are some similarities, with the results in Table III, such as the presence of market, size, and momentum factors. However, comparing the alphas reported in Tables III and VI, we see that factor models selected sequentially generate substantially lower out-of-sample alphas than models selected by solving Problem 2 directly.

Table VI shows that the out-of-sample alphas across all benchmarks are generally

lower with larger models chosen sequentially. A potential explanation is that errors made early on with sequential selection become baked-in, leading to sub-optimal choices of factors at subsequent iterations. Overall, the best subset selection method is more effective than selecting factors sequentially.

4 Conclusion

This paper offers a practical solution to the challenge of factor selection for linear asset pricing models. Model selection methods typically require estimating and evaluating a large number of models individually or in combination, which is infeasible when the set of factors is large. Our approach resolves this issue. We simultaneously address factor selection and estimation of SDF loadings by solving a single mixed-integer linear or quadratic program. Fast, robust algorithms for solving these problems are widespread in scientific computing environments and will only continue to improve with time.

In addition to its computational appeal, our approach is grounded in economics. It is based on minimizing economically motivated criteria that represent in-sample pricing errors. Moreover, it naturally accommodates bid-ask spreads and short-sale constraints, enforces strict positivity of the SDF, and allows for regularized rewards-for-risk within the mixed-integer linear and quadratic program formulations.

Our methodology is also supported by theoretical guarantees for consistent factor selection and consistent estimation of SDF loadings, requiring only weak assumptions — in particular no parametric assumptions — on the data-generating process. Further, bagged implementations of our method are optimal for factor selection. Overall, our approach provides an efficient and theoretically sound solution for factor selection.

As we demonstrate, our methodology selects five-, six- and seven-factor models that generate superior out-of-sample alphas relative to standard benchmarks. Combined with the fact that our criteria minimize in-sample pricing errors, these results suggest that our method effectively captures the cross-section of returns both in- and out-of-sample. A key finding regarding the factor composition is the importance of a momentum factor, which is included in all our five-, six- and preferred seven-factor models.

We also provide a framework for examining the issue of sparsity versus nonsparsity of the SDF by introducing an approach for estimating dense models that include all factors, while accommodating bid-ask spreads, short-sale constraints, regularized rewards for risk, and a strictly positive SDF. We formulate this estimation problem as a linear or quadratic program. Our empirical analysis finds that dense models estimated in this

way produce alphas that are substantially smaller than those produced by five-, six- and seven-factor models estimated using our method, further demonstrating the effectiveness of our best subset selection approach.

Table I: **Mean Absolute Errors on SDF Loadings across Simulations**

Method		**Factors ($k_{\max} = 7$ and $K_{\text{factor}} = 9$)**								
									(Irrelevant)	
		1	2	3	4	5	6	7	8	9
	Panel A: **Best Subset Selection**									
1. MIQP (Problem 2, q=2)	MAE	0.000	0.030	0.031	0.033	0.034	0.035	0.037	0.000	0.000
	MAE/(mean $\lvert\eta_k\rvert$)	0.000	3.9	4.1	4.4	4.6	4.7	5.0		
2. MILP (Problem 2, q=1)	MAE	0.000	0.034	0.036	0.038	0.039	0.041	0.042	0.000	0.000
	MAE/(mean $\lvert\eta_k\rvert$)	0.000	4.5	4.7	5.1	5.3	5.5	5.6		
	Panel B: **Benchmarking to No Subset Selection (Dense Model)**									
3. QP (Problem 3, q=2)	MAE	0.000	0.030	0.032	0.034	0.035	0.036	0.038	0.039	0.040
	MAE/(mean $\lvert\eta_k\rvert$)	0.000	4.0	4.2	4.5	4.7	4.9	5.1		
4. LP (Problem 3, q=1)	MAE	0.000	0.035	0.036	0.039	0.040	0.042	0.043	0.045	0.046
	MAE/(mean $\lvert\eta_k\rvert$)	0.000	4.6	4.8	5.2	5.4	5.7	5.7		
5. HJ distance	MAE	0.000	0.030	0.031	0.033	0.034	0.036	0.037	0.038	0.039
	MAE/(mean $\lvert\eta_k\rvert$)	0.000	4.0	4.2	4.4	4.6	4.8	5.0		
	Panel C: **Percentage of simulations in which the factor is correctly identified as relevant or irrelevant**									
MIQP (Problem 2, q=2)		100	100	100	100	99.98	99.98	100	100	99.96
MILP (Problem 2, q=1)		100	100	100	100	99.98	99.96	99.98	100	99.92

Note: Across all simulations, we set $k_{\max} = 7$ and $\mathsf{b}^{\text{glr}} = 9$. We report mean absolute errors (MAEs) and MAE/(mean $|\eta_k|$) (in %) for estimating each of the nine SDF loadings η_k across 5,000 simulations. In our simulation design, factors 8 and 9 are irrelevant, with $\eta_k = 0$ for $k = 8, 9$. See equations (24)–(26). We compare five estimation methodologies, as follows:

1. The best subset selection method via MIQP ($\mathsf{q} = 2$) in Problem 2;

2. The best subset selection method via MILP ($\mathsf{q} = 1$) in Problem 2;

3. QP ($\mathsf{q} = 2$) counterparts without best subset selection in Problem 3;

4. LP ($\mathsf{q} = 1$) counterparts without best subset selection in Problem 3; and

5. Estimation by minimizing the Hansen and Jagannathan (1997) distance measure with positivity of the SDF enforced (but no regularization of rewards-for-risk).

For the best subset selection method in Problem 2, we also report the percentage of simulations in which each factor is correctly identified as being relevant or irrelevant.

Table II: **List of 64 Factors used in the Best Subset Selection Methodology**

Factor Name and Description

Panel A: Fama and French
$R_{f,t}$: Gross risk free return
MKT: Excess market returns
SMB: Small minus Big
HML: High minus Low
RMW: Robust Minus Weak
CMA: Conservative Minus Aggressive

Panel B: Hou, Xue, and Zhang
Q^{roe}: Return on equity
$Q^{investment}$: Investment
$Q^{expected\ growth}$: Expected growth

Panel C: Daniel, Mota, Rottke, and Santos
MKT* (characteristic-efficient)
SMB* (characteristic-efficient)
HML* (characteristic-efficient)
RMW* (characteristic-efficient)
CMA* (characteristic-efficient)

Panel D: AQR factors
HML^{devil}: HML Devil
QMJ: Quality minus junk
BAB: Betting-against-beta

Panel E: Jensen, Kelly, and Pedersen
$Accrual^{operating}$: Operating accruals†
$Investment^{corporate}$: Abnormal corporate investment†
$Growth^{book\ debt}$: Growth in book debt (3 years)†
NOA: Net operating assets
$Growth^{asset}$: Asset growth
Hiring: Hiring rate†
Liquidity: Pastor-Stambaugh (LIQV, 10-1 portfolio)
$Growth^{inventory}$: Inventory growth†
$Mispricing^{management}$: Mispricing factor (management)
CNOA: Change in net operating assets
PPI: Change PPE and Inventory
Reversal: Long-term reversal
$Growth^{sales}$: Sales growth (1 year)
$Leverage^{book}$: Book leverage
R&D/Sales: R&D-to-sales
Altman: Altman Z-score†
$Beta^{market}$: Market beta
$Beta^{downside}$: Downside beta
VOL^{idio}: Idiosyncratic volatility (from 3-factor Fama-French)
$MOM^{t-3:t-1}$: Price momentum t-3 to t-1
$MOM^{t-6:t-1}$: Price momentum t-6 to t-1
$MOM^{t-12:t-1}$: Price momentum t-12 to t-1 (Jegadeesh and Titman (1993))
$MOM^{t-12:t-7}$: Price momentum t-12 to t-7
$Revenue^{surprise}$: Standardized revenue surprise†
$CV^{dollar\ trading\ volume}$: Coefficient of variation for dollar trading volume†
ROE: Return on equity
OLSON: Ohlson O-score†
Cash Flows to Assets: Operating cash flow to assets
Profits to Lagged Book: Cash-based operating profits-to-lagged book assets†
Gross Profits to Assets: Gross profits-to-assets
$Mispricing^{performance}$: Mispricing factor (performance)
$Leverage^{operating}$: Operating leverage
$MKT^{correlation}$: Market correlation
Coskewness: Coskewness
Amihud: Amihud measure†
$Volume^{dollar}$: Dollar trading volume†
MKT^{equity}: Market equity (Banz (1981))
R&D-to-Market: R&D-to-market
$Skewness^{idio}$: Idiosyncratic skewness (from 3-factor Fama-French)
Short-term Reversal: Short-term reversal
RSV: Highest 5 days of return scaled by volatility
Skewness: Total skewness
Stock Issues: Net stock issues†
Debt-to-Market: Debt-to-market ratio
Payout Yield: Net payout yield†
Cash flow-to-price: Free cash flow-to-price ratio
Earnings-to-Price: Earnings-to-price ratio

Note: This table displays the various factors used in our empirical analysis and their corresponding descriptions. These factors are derived from five main sources: (i) Fama and French (2015), (ii) Hou, Xue, and Zhang (2015, 2020), (iii) Daniel, Mota, Rottke, and Santos (2020), (iv) AQR (as referenced in Barillas and Shanken (2018)), and (v) factors provided by Jensen, Kelly, and Pedersen (2023). Newer factors with an absolute correlation coefficient higher than 0.8 are omitted. Documentation on factors provided by Jensen, Kelly, and Pedersen (2023) is on the website of Theis Jensen. We also create a set of $K_{\text{factor}} = 50$ factors from $K_{\text{factor}} = 64$ factors, with the 14 omitted ones marked by †.

Table III: **Best Subset Selection Alpha**

	Panel A: MIQP (Problem 2, q=2)			Panel B: MILP (Problem 2, q=1)		
Factors	7	6	5	7	6	5
Benchmark	Out-of-sample Alpha (%)			Out-of-sample Alpha (%)		
FF5	5.6 (0.05)	4.6 (0.05)	4.0 (0.11)	4.0 (0.12)	4.6 (0.05)	4.1 (0.11)
BS6	5.4 (0.04)	4.5 (0.07)	3.2 (0.14)	3.1 (0.14)	4.5 (0.07)	3.3 (0.14)
FF5+MOM	3.3 (0.10)	2.5 (0.12)	1.5 (0.23)	1.3 (0.29)	2.5 (0.12)	1.6 (0.23)

Note: This table presents the alphas, expressed in annualized percentage units, calculated from time-series regressions of the excess returns of the SDF-implied MVE portfolio on those of the three benchmarks:

1. FF5: The five-factor model of Fama and French (2015);

2. BS6: The six-factor model of Barillas and Shanken (2018);

3. FF5+MOM: The six-factor model resulting from adding the momentum factor to the Fama and French (2015) model;

over the out-of-sample period. We select factors and estimate loadings as described in Section 3.1, using data from January 1972 to December 1997. The composition of each factor model and the respective SDF loadings are shown in Table IV. The best subset selection and the loadings are held fixed over the out-of-sample period of January 1998 to December 2022. As in Kozak et al. (2020), we map estimated SDF loadings onto the excess returns of the MVE portfolio and leverage the MVE portfolio excess returns to have the same volatility as that of the market. We regularize rewards-for-risk by setting $\mathsf{b}^{\text{glr}} = 9$, which corresponds to a maximum allowable Sharpe ratio of four. In parentheses, we report the fraction of bootstrap draws in which the out-of-sample alpha is positive.

Table IV: **Estimates of SDF Loadings and Out-of-Sample Sharpe Ratios**

Panel A: MIQP (Problem 2, q=2)

SDF Loadings for the Seven-Factor Model

constant	SMB	CMA	$Q^{\text{expected growth}}$	MKT*	SMB*	$MOM^{t-6:t-1}$	MKT^{equity}	Sharpe Ratio
1.2•	-16.4•	-10.4•	-5.2•	-9.0•	-7.9	-5.4•	16.8•	0.76

SDF Loadings for the Six-Factor Model

constant	SMB	CMA	MKT*	QMJ	$MOM^{t-6:t-1}$	MKT^{equity}	
1.2•	-28.8•	-12.7•	-7.4•	-6.3•	-4.2•	24.6•	0.71

SDF Loadings for the Five-Factor Model

constant	SMB	$Q^{\text{investment}}$	MKT*	$MOM^{t-6:t-1}$	MKT^{equity}	
1.1•	-21.0•	-10.5•	-6.8•	-6.4•	18.8•	0.59

Panel B: MILP (Problem 2, q=1)

SDF Loadings for the Seven-Factor Model

constant	SMB	$Q^{\text{investment}}$	MKT*	$MOM^{t-12:t-1}$	$MOM^{t-12:t-7}$	MKT^{equity}	RSV	
1.2•	-21.8•	-11.9	-7.7•	-11.6•	9.0•	19.6•	-3.1	0.59

SDF Loadings for the Six-Factor Model

constant	SMB	CMA	MKT*	QMJ	$MOM^{t-6:t-1}$	MKT^{equity}	
1.2•	-28.8•	-12.7•	-7.4•	-6.3•	-4.2•	24.6•	0.71

SDF Loadings for the Five-Factor Model

constant	SMB	$Q^{\text{investment}}$	MKT*	$MOM^{t-6:t-1}$	MKT^{equity}	
1.1•	-21.3•	-10.2•	-6.8•	-6.4•	19.0•	0.59

Note: This table presents estimates of factor loadings computed as described in Section 3.1 by solving Problem 2 with $\mathsf{q} = 2$ (upper panel) and $\mathsf{q} = 1$ (lower panel) using data from January 1972 to December 1997. The final column shows the resulting out-of-sample annualized Sharpe ratio of the mean-variance efficient portfolio. A bullet indicates that the factor loading estimate remains of the same sign (and nonzero) across *all* bootstrap draws.

Table V: **Dense Models: Out-of-Sample Alphas**

Model	**Sparse**	**Dense**	**Dense**
Factors	7	64	50
Benchmark	Out-of-Sample Alpha (%)		
FF5	5.6	1.7	2.0
BS6	5.4	1.3	2.0
FF5+MOM	3.3	-0.3	0.2
	Out-of-Sample Sharpe Ratio		
	0.76	0.46	0.63

Note: This table presents the alphas, expressed in annualized percentage units, calculated from time-series regressions of the excess returns of the SDF-implied MVE portfolio on those of the three benchmarks:

1. FF5: The five-factor model of Fama and French (2015);

2. BS6: The six-factor model of Barillas and Shanken (2018);

3. FF5+MOM: The six-factor model resulting from adding the momentum factor to the Fama and French (2015) model;

over the out-of-sample period. We solve Problem 3 with $\mathsf{q} = 2$ to estimate the SDF loadings, using data from January 1972 to December 1997. The SDF loadings are held fixed over the out-of-sample period of January 1998 to December 2022. As in Kozak et al. (2020), we map estimated SDF loadings onto the excess returns of the MVE portfolio and leverage the MVE portfolio excess returns to have the same volatility as that of the market. We present one set of results using all 64 factors listed in Table II, and another set using a subset of 50 factors, with the 14 omitted ones marked by † in Table II. We also report out-of-sample annualized Sharpe ratios for the MVE portfolios. For each model, we regularize rewards-for-risk by setting $\mathsf{b}^{\text{glr}} = 9$, which corresponds to a maximum allowable Sharpe ratio of four.

Table VI: **Sequential Best Subset Selection Alpha**

	Panel 1A: MIQP (Problem 2, q=2)			Panel 1B: MILP (Problem 2, q=1)		
Factors	7	6	5	7	6	5
Benchmark	Out-of-sample Alpha (%)			Out-of-sample Alpha (%)		
FF5	-1.4	2.6	3.5	-2.0	2.2	-0.1
BS6	-1.4	2.4	3.1	-2.0	1.7	-0.6
FF5+MOM	-2.8	0.8	1.4	-3.8	-0.3	-2.6

Factors	Panel 2A: MIQP (Problem 2, q=2)
7	SMB*, BAB, MKT*, HML$^{\text{devil}}$, PPE, MOM$^{t-6:t-1}$, R&D/Sales,
6	SMB*, BAB, MKT*, HML$^{\text{devil}}$, PPE, MOM$^{t-6:t-1}$
5	SMB*, BAB, MKT*, HML$^{\text{devil}}$, PPE
	Panel 2B: MILP (Problem 2, q=1)
7	BAB, MKT*, PPE, R&D/Sales, MOM$^{t-12:t-1}$, MOM$^{t-12:t-7}$, SMB*
6	BAB, MKT*, PPE, R&D/Sales, MOM$^{t-12:t-1}$, MOM$^{t-12:t-7}$
5	BAB, MKT*, PPE, R&D/Sales, MOM$^{t-12:t-1}$

Note: This table presents the alphas, expressed in annualized percentage units, calculated from time-series regressions of the excess returns of the SDF-implied MVE portfolio on those of the three benchmarks:

1. FF5: The five-factor model of Fama and French (2015);

2. BS6: The six-factor model of Barillas and Shanken (2018);

3. FF5+MOM: The six-factor model resulting from adding the momentum factor to the Fama and French (2015) model;

over the out-of-sample period. We select factors and estimate loadings sequentially, as described in Section 3.5, using data from January 1972 to December 1997. The composition of each factor model is shown in the lower panel. The factors and loadings are held fixed over the out-of-sample period of January 1998 to December 2022. As in Kozak et al. (2020), we map estimated SDF loadings onto the excess returns of the MVE portfolio and leverage the MVE portfolio excess returns to have the same volatility as that of the market. We regularize rewards-for-risk by setting $\mathsf{b}^{\text{glr}} = 9$, which corresponds to a maximum allowable Sharpe ratio of four.

References

Asness, C. and A. Frazzini (2013). The devil in HML's details. *Journal of Portfolio Management 39*, 49–68.

Banz, R. (1981). The relationship between return and market value of common stocks. *Journal of Financial Economics 9*, 3–18.

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

Bernardo, A. and O. Ledoit (2000). Gain, loss and asset pricing. *Journal of Political Economy 108*, 144–172.

Bertsimas, D., A. King, and R. Mazumder (2016). Best subset selection via a modern optimization lens. *Annals of Statistics 44*, 813–852.

Bertsimas, D. and B. Stellato (2022). Online mixed-integer optimization in milliseconds. *INFORMS Journal on Computing 34*(4), 2229–2248.

Breiman, L. (1996). Bagging predictors. *Machine Learning 24*, 123–140.

Bryzgalova, S., J. Huang, and C. Julliard (2023). Bayesian solutions for the factor zoo: We just ran two quadrillion models. *Journal of Finance 78*, 487–557.

Bybee, L., B. Kelly, and Y. Su (2023). Narrative asset pricing: Interpretable systematic risk factors from news text. *Review of Financial Studies 36*, 4759–4787.

Carrasco, M. and C. Nokho (2024). Hansen-Jagannathan distance with many assets. Working paper, Université de Montréal.

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

Christensen, T., H. Moon, and F. Schorfheide (2023). Optimal decision rules when payoffs are partially identified. Working paper, Yale University.

Cochrane, J. (2005). *Asset Pricing*. Princeton, NJ: Princeton University Press.

Cochrane, J. (2011). Discount rates. *Journal of Finance 66*, 1047–1108.

Cochrane, J. and J. Saá-Requejo (2000). Beyond arbitrage: Good-Deal asset price bounds in incomplete markets. *Journal of Political Economy 108*, 79–119.

Corwin, S. and P. Schultz (2012). A simple way to estimate bid-ask spreads from daily high and low prices. *Journal of Finance 67*(2), 719–759.

Daniel, K., L. Mota, S. Rottke, and T. Santos (2020). The cross-section of risk and returns. *Review of Financial Studies 33*, 1927–1979.

Didisheim, A., S. Ke, B. Kelly, and S. Malamud (2023). Complexity in factor pricing models. Working paper, NBER # 31689.

Dümbgen, L. (1993). On nondifferentiable functions and the bootstrap. *Probability Theory and Related Fields 95*, 125–140.

Fama, E. and K. French (2015). A five-factor asset pricing model. *Journal of Financial Economics 116*, 1–22.

Feng, G., S. Giglio, and D. Xiu (2020). Taming the factor zoo. *Journal of Finance 75*, 1327–1370.

Freyberger, J., A. Neuhierl, and M. Weber (2020). Dissecting characteristics nonparametrically. *Review of Financial Studies 33*, 2326–3277.

Giglio, S., Y. Liao, and D. Xiu (2020). Thousands of alpha tests. *Review of Financial Studies 30*, 1382–1423.

Giglio, S., D. Xiu, and D. Zhang (2024). Test assets and weak factors. *Journal of Finance (forthcoming)*, 1–21.

Gospodinov, N., R. Kan, and C. Robotti (2016). On the properties of the constrained Hansen-Jagannathan distance. *Journal of Empirical Finance 36*, 121–150.

Gospodinov, N., R. Kan, and C. Robotti (2019). Too good to be true? Fallacies in evaluating risk factor models. *Journal of Financial Economics 132*(1), 451–471.

Hansen, L. and R. Jagannathan (1997). Assessing specification errors in stochastic discount factor models. *Journal of Finance 52*(2), 557–590.

Hansen, L. and S. Richard (1987). The role of conditioning information in deducing testable restrictions implied by dynamic asset pricing models. *Econometrica 55*(3), 587–613.

Harvey, C. and Y. Liu (2019). A census of the factor zoo. Working paper, Duke University and Purdue University.

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

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

Hou, K., C. Xue, and L. Zhang (2020). Replicating anomalies. *Review of Financial Studies 33*, 2019–2133.

Jegadeesh, N. and S. Titman (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *Journal of Finance 48*(1), 65–91.

Jensen, T., B. Kelly, and L. Pedersen (2023). Is there a replication crisis in finance? *Journal of Finance 78*(5), 2465–2518.

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

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

Preite, M., R. Uppal, P. Zaffaroni, and I. Zviadadze (2024). What is missing in asset-pricing factor models? Working paper, EDHEC.

Rubin, D. (1981). The Bayesian bootstrap. *Annals of Statistics 9*(1), 130–134.

Shapiro, A. (1991). Asymptotic analysis of stochastic programs. *Annals of Operations Research 30*, 169–186.

Tsybakov, A. (2009). *Introduction to Nonparametric Estimation*. Springer.

Walker, S. (2013). Bayesian inference with misspecified models. *Journal of Statistical Planning and Inference 143*(10), 1621–1633.

Online Appendix: Culling the Factor Zoo

Appendix A Extensions of the Methodology

A.1 Estimation of Dense Factor Models

It is oftentimes of interest to estimate "dense" factor models that use all K_{factor} factors. Here the goal is to estimate the vector of factor loadings $\boldsymbol{\eta}_{\mathsf{q},K_{\text{factor}}}$ that minimizes the distance measure $Q^{(\mathsf{q})}(\boldsymbol{\eta})$ in display (12). This is achieved by solving Problem 4 below.

Our approach is analogous to Problem 1. We drop the sparsity constraint and omit the indicator variables χ_k. This leads us to the following problem, which is a linear (if $\mathsf{q}=1$) or quadratic (if $\mathsf{q}=2$) program:

Problem 4 (Estimation for Dense Factor Models)

$$\min_{\boldsymbol{\eta},\{n_t,a_t\}_{t=1}^T} \frac{1}{T}\sum_{t=1}^T a_t^{\mathsf{q}} \quad \text{subject to} \quad \begin{aligned} \mathbf{1}-\bar{\mathbf{c}}_T \leq \frac{1}{T}\sum_{t=1}^T n_t\mathbf{R}_t \leq \mathbf{1}+\bar{\mathbf{c}}_T, & \\ n_t \geq 0, & \quad t=1,\ldots,T, \\ -a_t \leq n_t - \mathbf{f}_t'\boldsymbol{\eta} \leq a_t, & \quad t=1,\ldots,T, \\ -\mathsf{b} \leq \eta_k \leq \mathsf{b}, & \quad k=1,\ldots,K_{\text{factor}}. \end{aligned}$$

For technical reasons, we maintain the constraints $-\mathsf{b} \leq \eta_k \leq \mathsf{b}$ for each of the elements of η, with b arbitrarily large but finite. In practice, b can be set sufficiently large that these constraints are never actually binding. The estimator $\hat{\boldsymbol{\eta}}_{\mathsf{q},K_{\text{factor}}}$ of $\boldsymbol{\eta}_{\mathsf{q},K_{\text{factor}}}$ is the value of $\boldsymbol{\eta}$ that solves Problem 4. Problem 3 is the counterpart dense factor model problem accounting for regularized rewards-for-risk.

The next theorem says that the Linear/Quadratic Programming estimator $\hat{\boldsymbol{\eta}}_{\mathsf{q},K_{\text{factor}}}$ in Problem 4 is consistent for $\boldsymbol{\eta}_{\mathsf{q},K_{\text{factor}}}$ as the sample size T becomes large.

Theorem 2 *Suppose that Assumptions 2 and 3 in Appendix B hold. Then for any $\epsilon > 0$,*

$$\Pr\left(Q^{(\mathsf{q})}(\hat{\boldsymbol{\eta}}_{\mathsf{q},K_{\text{factor}}}) > Q^{(\mathsf{q})}(\boldsymbol{\eta}_{\mathsf{q},K_{\text{factor}}}) + \epsilon\right) \to 0$$

as the sample size T becomes large, where $Q^{(\mathsf{q})}(\boldsymbol{\eta})$ denotes the distance measure in display (12). If, in addition, $\boldsymbol{\eta}_{\mathsf{q},K_{\text{factor}}}$ uniquely minimizes $Q^{(\mathsf{q})}$, then

$$\Pr\left(\left\|\hat{\boldsymbol{\eta}}_{\mathsf{q},K_{\text{factor}}} - \boldsymbol{\eta}_{\mathsf{q},K_{\text{factor}}}\right\| > \epsilon\right) \to 0$$

as the sample size T becomes large.

The proof of Theorem 2 is presented in Appendix B.

A.2 Time-Varying SDF Loadings

Our methodology extends to conditional asset pricing models with time-varying SDF loadings (e.g., Cochrane (1996) and Lettau and Ludvigson (2001)). Here we posit Radon–Nikodym derivatives of the form

$$y_{t+1}(\boldsymbol{\eta}_t) = \mathbf{f}_{t+1}'\boldsymbol{\eta}_t, \qquad \boldsymbol{\eta}_t = \mathbf{B}\mathbf{w}_t,$$

where $\mathbf{w}_t$ is a $K_{\text{cv}} \times 1$ vector of conditioning variables that are known at date t and $\mathbf{B}$ is a $K_{\text{factor}} \times K_{\text{cv}}$ matrix of parameters. The first element of $\mathbf{w}_t$ is a constant and the remaining elements are conditioning variables. Equivalently,

$$y_{t+1}(\boldsymbol{\eta}) = (\mathbf{f}_{t+1} \otimes \mathbf{w}_t)'\boldsymbol{\eta},$$

where $\boldsymbol{\eta} = \text{vec}(\mathbf{B}')$ and $\otimes$ denotes the Kronecker product. That is, $\boldsymbol{\eta} = (\mathbf{b}_1', \ldots, \mathbf{b}_{K_{\text{factor}}}')'$ where $\mathbf{b}_k'$ is the kth row of $\mathbf{B}$. Hence, we have

$$y_{t+1}(\boldsymbol{\eta}) = \mathbf{F}_{t+1}'\boldsymbol{\eta},$$

where $F_{t+1} = \mathbf{f}_{t+1} \otimes \mathbf{w}_t$.

Using the terminology of Lettau and Ludvigson (2001), we refer to $\mathbf{f}_t$ as fundamental factors and $\mathbf{F}_t$ as variable factors. Suppose the data are $(\mathbf{f}_t, \mathbf{R}_t, \mathbf{c}_t, \mathbf{w}_{t-1})_{t=1}^T$. The goal is to select the best $k_{\max}$ fundamental factors from the set of K_{factor} fundamental factors.

By similar steps to the derivation of Problem 1, (fundamental) factor selection and estimation of time-varying factor loadings is performed in a single step by solving the following MILP/MIQP:

Problem 5 (Methodology for Time-Varying SDF Loadings)

$$
\begin{aligned}
\min_{\boldsymbol{\eta},\{n_t,a_t\}_{t=1}^{T},\{\chi_k\}_{k=1}^{K_{\text{factor}}}} \quad & \frac{1}{T}\sum_{t=1}^{T} a_t^{\mathsf{q}} \\
\text{subject to} \quad & \mathbf{1}-\bar{\mathbf{c}}_T \leq \tfrac{1}{T}\textstyle\sum_{t=1}^{T} n_t \mathbf{R}_t \leq \mathbf{1}+\bar{\mathbf{c}}_T, \\
& n_t \geq 0, && t=1,\ldots,T, \\
& -a_t \leq n_t - \mathbf{F}_t'\boldsymbol{\eta} \leq a_t, && t=1,\ldots,T, \\
& \chi_k \in \{0,1\}, && k=1,\ldots,K_{\text{factor}}, \\
& -\mathsf{b}\chi_k \mathbf{1} \leq \mathbf{b}_k \leq \mathsf{b}\chi_k \mathbf{1}, && k=1,\ldots,K_{\text{factor}}, \\
& \textstyle\sum_{k=1}^{K_{\text{factor}}} \chi_k \leq k_{\max}.
\end{aligned}
$$

Relative to Problem 1, the only change is replacing $\mathbf{f}_t'\boldsymbol{\eta}$ with $\mathbf{F}_t'\boldsymbol{\eta}$ and replacing the constraint $-\mathsf{b}\chi_k \leq \eta_k \leq \mathsf{b}\chi_k$ with the constraint $-\mathsf{b}\chi_k\mathbf{1} \leq \mathbf{b}_k \leq \mathsf{b}\chi_k\mathbf{1}$ for a conformable vector of ones. This constraint becomes $\mathbf{b}_k \in [-\mathsf{b},\mathsf{b}]^{K_{\text{cv}}}$ when $\chi_k = 1$ or $\mathbf{b}_k = \mathbf{0}$ when $\chi_k = 0$. In the latter case, it has the effect of discarding the kth fundamental factor. The constraint $\sum_{k=1}^{K_{\text{factor}}} \chi_k \leq k_{\max}$ ensures that a model with at most $k_{\max}$ fundamental factors is selected.

Let $\hat{\boldsymbol{\eta}}_{\mathsf{q},k_{\max}}$ denote the value of $\boldsymbol{\eta}$ that solves Problem 5. This will be a vector of length $K_{\text{factor}} \times K_{\text{cv}}$. Partitioning $\hat{\boldsymbol{\eta}}_{\mathsf{q},k_{\max}}$ as $\hat{\boldsymbol{\eta}}_{\mathsf{q},k_{\max}} = (\hat{\mathbf{b}}_{\mathsf{q},1}', \ldots, \hat{\mathbf{b}}_{\mathsf{q},K_{\text{factor}}}')'$, we have

$$
\hat{\mathbf{B}}_{\mathsf{q},k_{\max}} = \begin{bmatrix} \hat{\mathbf{b}}_{\mathsf{q},1}' \\ \vdots \\ \hat{\mathbf{b}}_{\mathsf{q},K_{\text{factor}}}' \end{bmatrix}.
$$

By construction, at most $k_{\max}$ rows of $\hat{\mathbf{B}}_{\mathsf{q},k_{\max}}$ will be nonzero. These are indicated by $\chi_k = 1$ and correspond to the best subset selection of fundamental factors. Estimated loadings on the selected factors at date t are given by the nonzero elements of $\hat{\mathbf{B}}_{\mathsf{q},k_{\max}}\mathbf{w}_t$.

This methodology can be extended to select the best subset of conditioning variables $\mathbf{w}_t$. To do so, we introduce an additional set of K_{cv} binary variables χ_j, $j = 1,\ldots,K_{\text{cv}}$, where $\chi_j = 1$ indicates that conditioning variable j is selected and $\chi_j = 0$ indicates that the variable is discarded. Let $\mathbf{b}_{\cdot j}$ denote the jth column of $\mathbf{B}$. Then the best subset selection of $k_{\max}$ fundamental factors, the best subset selection of $j_{\max}$ conditioning variables, and the estimation of factor loadings can be performed in one step by solving

Problem 5 with the following additional constraints:

$$\begin{aligned}
&\chi_j \in \{0,1\}, && j = 1,\ldots,K_{\text{cv}},\\
&-\mathsf{b}\chi_j \mathbf{1} \le \mathbf{b}_{\cdot j} \le \mathsf{b}\chi_j \mathbf{1}, && j = 1,\ldots,K_{\text{cv}},\\
&\sum_{j=1}^{K_{\text{cv}}} \chi_j \le j_{\max},
\end{aligned}$$

and optimizing over $\boldsymbol{\eta}$, $\{n_t, a_t\}_{t=1}^{T}$, $\{\chi_k\}_{k=1}^{K_{\text{factor}}}$, and $\{\chi_j\}_{j=1}^{K_{\text{cv}}}$.

A.3 Discussion of Regularized Rewards-for-Risk

The gain-loss ratio, as developed by Bernardo and Ledoit (2000), is the expected gain (over and above the risk-free rate) of a security or portfolio divided by the expected loss.

Consider any risky asset k priced by an admissible SDF. We have (dropping time subscripts) $\mathbb{E}[nZ_k] = p_k$, for some p_k that satisfies $p_{b,k} \le p_k \le p_{a,k}$. Rearranging gives $\mathbb{E}[n(Z_k - p_k)] = 0$. Using the identity $a = \max(a,0) - \max(-a,0)$ for $a \in \mathbb{R}$, we obtain

$$\mathbb{E}[n \max(Z_k - p_k, 0)] = \mathbb{E}[n \max(p_k - Z_k, 0)]. \tag{A1}$$

Suppose that the state space is discrete and we index states by j. We denote the value of the Radon–Nikodym derivative in state j by n_j. Then it follows from equation (A1) that $\mathbb{E}[\max(Z_k - p_k, 0)] \times \inf_j n_j \le \mathbb{E}[\max(p_k - Z_k, 0)] \times \sup_j n_j$.

When $\inf_j n_j > 0$, we have

$$\frac{\mathbb{E}[\max(Z_k - p_{a,k}, 0)]}{\mathbb{E}[\max(p_{b,k} - Z_k, 0)]} \le \frac{\mathbb{E}[\max(Z_k - p_k, 0)]}{\mathbb{E}[\max(p_k - Z_k, 0)]} \le \frac{\sup_j n_j}{\inf_j n_j}. \tag{A2}$$

The left-most term in equation (A2) is the gain-loss ratio with bid-ask spreads. This is bounded by the the gain-loss ratio at the price p_k implied by n, which, in turn, is bounded by the ratio of the maximum to minimum values of the Radon–Nikodym derivative.

In view of (A2), one could rule out unreasonably high gain-loss ratios by enforcing $\sup_j n_j \le \mathsf{b}^{\text{glr}} \times \inf_j n_j$, where b^{glr} is a bound on the maximum gain-loss ratio, for some chosen b^{glr} satisfying $\mathsf{b}^{\text{glr}} \ge 1$. It is much easier from a computational standpoint to instead enforce the sufficient condition

$$\frac{1}{\sqrt{\mathsf{b}^{\text{glr}}}} \le n_j \le \sqrt{\mathsf{b}^{\text{glr}}} \quad \text{for all } j. \tag{A3}$$

Additionally, (A3) automatically enforces $n_j > 0$ in all states j. This leads to the regularized rewards-for-risk constraint

$$\frac{1}{\sqrt{\mathsf{b}^{\text{glr}}}} \leq n_t \leq \sqrt{\mathsf{b}^{\text{glr}}}, \qquad t = 1, \ldots, T,$$

in display (19).

How does $\mathsf{b}^{\text{glr}} = 9$ cap Sharpe ratios at 4? First note that for a random variable X with mean μ and $\Pr(m \leq X \leq M) = 1$, we have

$$\text{Var}[X] = (M - \mu)(\mu - m) - \mathbb{E}[(M - X)(X - m)] \leq (M - \mu)(\mu - m).$$

Hence, the bounds $\frac{1}{\sqrt{\mathsf{b}}} \leq n_t \leq \sqrt{\mathsf{b}}$ imply that

$$\text{Var}[n_t] \;\leq\; (\sqrt{\mathsf{b}} - 1)\,(1 - \frac{1}{\sqrt{\mathsf{b}}}). \tag{A4}$$

Equation (A4) motivates us to solve the following minimum second-moment problem (compare with Hansen and Jagannathan (1991)) enforcing positivity and allowing for bid-ask spreads:

$$\min_{\{n_t\}_{t=1}^T} n_t^2 \qquad \text{subject to} \qquad \begin{aligned} \mathbf{1} - \bar{\mathbf{c}}_T \leq \tfrac{1}{T} \textstyle\sum_{t=1}^T n_t \mathbf{R}_t \leq \mathbf{1} + \bar{\mathbf{c}}_T, & \\ n_t \geq 0, \qquad\qquad\qquad\qquad & t = 1, \ldots, T. \end{aligned}$$

Using our data set of 203 test asset returns over January 1972 to December 2022, the minimizing n_t has an annualized standard deviation of 2.938. This annualized standard deviation may seem high, but it is ex-post: The portfolio achieving this Sharpe ratio was unknowable in 1972.

We use the Hansen and Jagannathan (1991) bound to link the maximum annualized Sharpe ratio to the volatility of the Radon–Nikodym derivative, denoted σ_{RN}. Given σ_{RN}, we use inequality (A4) to compute a corresponding value of b^{glr}. This has the advantage that Sharpe ratios are more familiar and easier to interpret than gain-loss ratios. We compute the value of b^{glr} that solves

$$(\sqrt{\mathsf{b}^{\text{glr}}} - 1)\,(1 - \frac{1}{\sqrt{\mathsf{b}^{\text{glr}}}}) \;=\; (\sigma_{\text{RN}})^2 \;\times\; (1/12),$$

where the factor 1/12 is to convert to a monthly equivalent. Values of b^{glr} and corresponding maximum annualized Sharpe ratios are presented in Table VII.

Table VII: **Relation between $\mathfrak{b}^{\text{glr}}$ and Sharpe Ratios**

$\mathfrak{b}^{\text{glr}}$	6.982	9.0	14.605	22.956
Maximum annualized Sharpe ratio	3.5	4	5	6

Using all 51 years of data from January 1972 to December 2022 and using 203 test assets, the value of $\underline{\mathfrak{b}}$ defined in equation (23) is computed to be $\underline{\mathfrak{b}} = 2.6175$. All the possible choices of $\mathfrak{b}^{\text{glr}}$ in Table VII exceed $\underline{\mathfrak{b}}$, as required.

Appendix B Technical Results and Proofs

B.1 Interpretation of the Distance Measure when $\mathsf{q} = 1$

Here we formally derive an interpretation of the distance measure $\mathbb{D}_t^{(1)}(y_{t+1})$ in terms of Arrow–Debreu security prices, as stated in display (7). We formalize this in Lemma 1.

We first introduce some notation and a regularity condition. To simplify the presentation, we present this for the case in which all assets except the risk-free asset have a bid-ask spread. The argument extends to the case where there are spreads for some assets but not others at the cost of more complicated notation. We also drop time subscripts in all that follows. Thus, $\mathbb{E}$ should be interpreted as the conditional expectation $\mathbb{E}_t$, $\mathbb{P}$ as $\mathbb{P}_t$, n and y as n_{t+1} and y_{t+1}, and so forth. For any random variable W, we let essinf(W) denote the greatest lower bound B for which $\mathbb{P}(W \geq B) = 1$. Let $\check{\mathbf{Z}}$, $\check{\mathbf{p}}_a$, and $\check{\mathbf{p}}_b$ denote the vectors $\mathbf{Z}$, $\mathbf{p}_a$, and $\mathbf{p}_b$ with their first element (corresponding to the risk-free asset) removed. Let $A + B = \{a + b : a \in A, b \in B\}$ for two sets A, B. Define

$$\mathcal{E} = \{(\mathbb{E}[n] - 1, \mathbb{E}[n\check{\mathbf{Z}}]' - \check{\mathbf{p}}_a', \check{\mathbf{p}}_b' - \mathbb{E}[n\check{\mathbf{Z}}]')' : n \in L_+^2\} + \{0\} \times (\mathbb{R}_+)^{2(K_{\text{asset}}-1)}. \tag{A5}$$

Assumption 1 *The set $\mathcal{E}$ contains the origin in its interior.*

In words, Assumption 1 asserts that there exist admissible Radon–Nikodym derivatives at which any inequalities in (3) hold strictly. It is possible to provide more primitive sufficient conditions; see Appendix A of Hansen and Jagannathan (1997) for the case of equalities.

Lemma 1 *Suppose that Assumption 1 holds for the Arrow–Debreu environment described in Section 2.2. Suppose furthermore that $\mathbb{E}[y] = 1$ and $\mathbb{P}(y \geq 0) = 1$. Then the representation (7) holds.*

As a stepping stone to the proof of Lemma 1, we first state and prove a preliminary result showing that the infinite-dimensional minimization problem defining $\mathbb{D}^{(1)}$ in equality (4) has an equivalent finite-dimensional representation.

Lemma 2 *Suppose that Assumption 1 holds. Then*

$$\mathbb{D}^{(1)}(y) = \max_{\substack{\boldsymbol{\xi}_+,\boldsymbol{\xi}_- \in \mathbb{R}_+^{K_{\text{asset}}} : (\boldsymbol{\xi}_+)'\boldsymbol{\xi}_- = 0, \\ \operatorname{essinf}((\boldsymbol{\xi}_+ - \boldsymbol{\xi}_-)'\mathbf{Z}) \geq -1}} \frac{1}{2} \left(\mathbb{E}\left[\min\left\{ (\boldsymbol{\xi}_+ - \boldsymbol{\xi}_-)'\mathbf{Z}, 1 \right\} y \right] - \boldsymbol{\xi}_+'\mathbf{p}_a + \boldsymbol{\xi}_-'\mathbf{p}_b \right). \quad \text{(A6)}$$

Moreover, the set of maximizing $(\boldsymbol{\xi}_+, \boldsymbol{\xi}_-)$ *is nonempty, convex, and compact.*

In equation (A6), $\boldsymbol{\xi}_+$ and $\boldsymbol{\xi}_-$ are vectors of Lagrange multipliers on the bid and ask price inequalities. Note that these are complementary: At most, one of the bid and ask price inequalities is binding. Hence, at most, one of $\xi_{+,j}$ and $\xi_{-,j}$ is nonzero.

Proof of Lemma 2. We apply duality theory as exposited in Bonnans and Shapiro (2000, Chapter 2.5.3). In their notation, we let $X = X^* = L^2$ and $Y = Y^* = (\mathbb{R})^{2K_{\text{asset}}-1}$. The objective is $f(x) = \frac{1}{2}\mathbb{E}[|x - y|] + \infty \times \mathbb{I}[\operatorname{essinf}(x) < 0]$ with the understanding that $0 \times +\infty = 0$. Evidently, $f : X \to \mathbb{R} \cup \{+\infty\}$ is convex and lower semicontinuous. The constraint is of the form $G(x) \in K$, with $G(x) = \frac{1}{2}(\mathbb{E}[x] - 1, \mathbb{E}[x\check{\mathbf{Z}}]' - \check{\mathbf{p}}_a', \check{\mathbf{p}}_b' - \mathbb{E}[x\check{\mathbf{Z}}]')'$ and $K = \{0\} \times (\mathbb{R}_-)^{2(K_{\text{asset}}-1)}$.

The Lagrangian for problem (4) is

$$\mathcal{L}(n, \xi, \check{\boldsymbol{\xi}}_+, \check{\boldsymbol{\xi}}_-) = \frac{1}{2}\left(\mathbb{E}\left[|n - y| + n(\xi + (\check{\boldsymbol{\xi}}_+ - \check{\boldsymbol{\xi}}_-)'\check{\mathbf{Z}}) \right] - \xi - \check{\boldsymbol{\xi}}_+'\check{\mathbf{p}}_a + \check{\boldsymbol{\xi}}_-'\check{\mathbf{p}}_b \right). \quad \text{(A7)}$$

The Lagrangian dual is

$$\sup_{\xi, \check{\boldsymbol{\xi}}_+, \check{\boldsymbol{\xi}}_-} \mathcal{L}^\star(\xi, \check{\boldsymbol{\xi}}_+, \check{\boldsymbol{\xi}}_-), \quad \text{(A8)}$$

where

$$\begin{aligned} \mathcal{L}^\star(\xi, \check{\boldsymbol{\xi}}_+, \check{\boldsymbol{\xi}}_-) &= \inf_{n \in L_+^2} \frac{1}{2}\left(\mathbb{E}\left[|n - y| + n(\xi + (\check{\boldsymbol{\xi}}_+ - \check{\boldsymbol{\xi}}_-)'\check{\mathbf{Z}}) \right] - \xi - \check{\boldsymbol{\xi}}_+'\check{\mathbf{p}}_a + \check{\boldsymbol{\xi}}_-'\check{\mathbf{p}}_b \right) \\ &= \frac{1}{2}\left(\mathbb{E}\left[\inf_{u \geq 0} |u - y| + u(\xi + (\check{\boldsymbol{\xi}}_+ - \check{\boldsymbol{\xi}}_-)'\check{\mathbf{Z}}) \right] - \xi - \check{\boldsymbol{\xi}}_+'\check{\mathbf{p}}_a + \check{\boldsymbol{\xi}}_-'\check{\mathbf{p}}_b \right). \end{aligned} \quad \text{(A9)}$$

Shifting the minimization inside the expectation is justified by Theorem 14.60 in Rockafellar and Wets (2009). The inner minimization is performed pointwise and is of the form

$$\inf_{u \geq 0} |u - a| + \mathtt{t}\, u\,. \quad \text{(A10)}$$

The optimal value is $-\infty$ if $\mathtt{t} < -1$ and $\min\{\mathtt{t}, 1\} \times a$ if $\mathtt{t} \geq -1$. Evidently, it is without loss of generality to restrict the domain of the outer optimization to $(\xi, \check{\boldsymbol{\xi}}_+, \check{\boldsymbol{\xi}}_-)$, for which $\mathbb{P}((\xi + (\check{\boldsymbol{\xi}}_+ - \check{\boldsymbol{\xi}}_-)'\check{\mathbf{Z}}) \geq -1) = 1$.

Equivalence of the primal and dual values and nonemptyness, convexity, and compactness of the set of dual solutions now follows under Assumption 1 by Theorem 2.165 of Bonnans and Shapiro (2000). Note that the multipliers $\check{\boldsymbol{\xi}}_+$ and $\check{\boldsymbol{\xi}}_-$ are complementary (i.e., $(\check{\boldsymbol{\xi}}_+)'\check{\boldsymbol{\xi}}_- = 0$) because, for each asset, at most one of the bid- and ask-price inequalities is binding. Finally, note that we can rewrite $\xi = \xi_+ - \xi_-$ where $\xi_+ = \max(\xi, 0)$ and $\xi_- = -\min(\xi, 0)$. Evidently, $\xi_+ \times \xi_- = 0$. The result follows by setting $\boldsymbol{\xi}_+ = (R_f\xi_+, \check{\boldsymbol{\xi}}_+')'$ and $\boldsymbol{\xi}_- = (R_f\xi_-, \check{\boldsymbol{\xi}}_-')'$ and noting that the first elements of $\mathbf{Z}$, $\mathbf{p}_a$, and $\mathbf{p}_b$ are R_f^{-1}. ■
With Lemma 2 in hand, we are now in a position to prove Lemma 1.

Proof of Lemma 1. Define $A_1 = (\cup_{j=2}^{K_{\text{asset}}} A_j)^c$ so that $A_1, \ldots, A_{K_{\text{asset}}}$ form a partition of $\mathcal{S}$. Let $\tilde{\xi}_1 = \xi_{+,1} - \xi_{-,1}$ and $\tilde{\xi}_j = \tilde{\xi}_1 + \xi_{+,j} - \xi_{-,j}$ for $j = 2, \ldots, K_{\text{asset}}$. Then

$$\begin{aligned}\mathbb{E}\left[\min\left\{(\boldsymbol{\xi}_+ - \boldsymbol{\xi}_-)'\mathbf{Z}), 1\right\} y\right] &= \mathbb{E}\left[R_f^{-1} \min\left\{\sum_{j=1}^{K_{\text{asset}}} \tilde{\xi}_j \mathbb{I}[s \in A_j], R_f\right\} y\right] \\ &= \sum_{j=1}^{K_{\text{asset}}} \min\left\{\tilde{\xi}_j, R_f\right\} \tilde{p}_j,\end{aligned}$$

where $\tilde{p}_j = R_f^{-1}\mathbb{E}\left[y\,\mathbb{I}[s \in A_j]\right]$ is the price implied by y of the Arrow–Debreu security paying off in state A_j. Note that $\tilde{p}_1 = R_f^{-1} - \sum_{j=2}^{K_{\text{asset}}} \tilde{p}_j$ because $A_1, \ldots, A_{K_{\text{asset}}}$ form a partition and y has unit expectation. Hence,

$$\begin{aligned}&\mathbb{E}\left[\min\left\{(\boldsymbol{\xi}_+ - \boldsymbol{\xi}_-)'\mathbf{Z}), 1\right\} y\right] \\ &\qquad = \min\left\{\tilde{\xi}_1, R_f\right\} R_f^{-1} + \sum_{j=2}^{K_{\text{asset}}} \left(\min\left\{\tilde{\xi}_j, R_f\right\} - \min\left\{\tilde{\xi}_1, R_f\right\}\right) \tilde{p}_j.\end{aligned}$$

Moreover, as $p_{a,1} = p_{b,1} = R_f^{-1}$, the dual objective becomes (dropping the factor $\frac{1}{2}$)

$$\begin{aligned}\min\left\{0, R_f - \tilde{\xi}_1\right\} R_f^{-1} + \sum_{j=2}^{K_{\text{asset}}} \Bigg(\Bigg(\min\left\{\xi_{+,j} - \xi_{-,j}, R_f - \tilde{\xi}_1\right\} - \min\left\{0, R_f - \tilde{\xi}_1\right\}\Bigg)\tilde{p}_j \\ - \xi_{+,j}\, p_{a,j} + \xi_{-,j}\, p_{b,j}\Bigg). \quad \text{(A11)}\end{aligned}$$

We wish to maximize the objective function (A11) with respect to $(\boldsymbol{\xi}_+, \boldsymbol{\xi}_-) \in (\mathbb{R}_+)^{2K_{\text{asset}}}$,

subject to the constraints in (A6).

Let $I_a \subset \{2, \ldots, K_{\text{asset}}\}$ denote the indices for which $\tilde{p}_j > p_{a,j}$. Thus, I_a is empty if $\tilde{p}_j \leq p_{a,j}$ for all $j \in \{2, \ldots, K_{\text{asset}}\}$. Similarly, let $I_b \subset \{2, \ldots, K_{\text{asset}}\}$ denote the indices for which $\tilde{p}_j < p_{b,j}$. Evidently, it is optimal to choose $\xi_{j,+} = 0$ unless $j \in I_a$ and $\xi_{j,-} = 0$ unless $j \in I_b$. Moreover, inspection of the objective shows that for $j \in I_a$, it is optimal to set $\xi_{+,j} = \max\{R_f - \tilde{\xi}_1, 0\}$, and for $j \in I_b$ it is optimal to set $\xi_{-,j}$ as large as possible, subject to the constraints in (A6). These choices satisfy the complementarity constraint $(\boldsymbol{\xi}_+)'\boldsymbol{\xi}_- = 0$.

In order that $\operatorname{essinf}((\boldsymbol{\xi}_+ - \boldsymbol{\xi}_-)'\mathbf{Z}) \geq -1$ hold, it is necessary and sufficient that $\tilde{\xi}_j R_f^{-1} \geq -1$ for $j = 1, \ldots, K_{\text{asset}}$. Taking $j = 1$ this means we need $\tilde{\xi}_1 \geq -R_f$. Additionally, for $j \in I_a$, we need $\tilde{\xi}_j \equiv \tilde{\xi}_1 + \max(R_f - \tilde{\xi}_1, 0) \geq -R_f$, which is always satisfied. Moreover, for $j \in I_b$, we need $\tilde{\xi}_j \equiv \tilde{\xi}_1 - \xi_{j,-} \geq -R_f$. Hence, $\xi_{j,-} = \tilde{\xi}_1 + R_f$ for $j \in I_b$.

For these choices, the objective (A11) becomes

$$\min(0, R_f - \tilde{\xi}_1)R_f^{-1} + \max(0, R_f - \tilde{\xi}_1)\sum_{j \in I_a}(\tilde{p}_j - p_{a,j}) + (\tilde{\xi}_1 + R_f)\sum_{j \in I_b}(p_{b,j} - \tilde{p}_j) - \sum_{j \in I_b}\min(0, R_f - \tilde{\xi}_1)\tilde{p}_j,$$

with the understanding that a sum over an empty index is zero. It remains to optimize this quantity with respect to $\tilde{\xi}_1 \geq -R_f$. Now, as $\sum_{j \in I_b} \tilde{p}_j < R_f^{-1}$, it is never optimal to set $\tilde{\xi}_1 > R_f$, as we could do strictly better by choosing $\tilde{\xi}_1 = R_f$. Hence, $\tilde{\xi}_1 \in [-R_f, R_f]$. For such values, the objective (A11) becomes

$$(R_f - \tilde{\xi}_1)\sum_{j \in I_a}(\tilde{p}_j - p_{a,j}) + (\tilde{\xi}_1 + R_f)\sum_{j \in I_b}(p_{b,j} - \tilde{p}_j),$$

which we want to maximize for $\tilde{\xi}_1 \in [-R_f, R_f]$. As the objective is linear in $\tilde{\xi}_1$, the optimal value will be at an endpoint of the interval. Setting $\tilde{\xi}_1 = -R_f$ yields the value $2R_f \sum_{j \in I_a}(\tilde{p}_j - p_{a,j})$, whereas setting $\tilde{\xi}_1 = R_f$ yields the value $2R_f \sum_{j \in I_b}(p_{b,j} - \tilde{p}_j)$. ■

B.2 Statistical Guarantees

For brevity we only prove the results for the case $\mathsf{q} = 1$, as it is more novel. Proofs for the case $\mathsf{q} = 2$ follow similarly. As in Appendix B.1, we present proofs for the case in which all assets except the risk-free asset have a positive bid-ask spread to simplify notation. We first prove consistency of the "dense" estimator $\hat{\boldsymbol{\eta}}_{1,K_{\text{factor}}}$ from Problem 4. We then establish consistency of the Best Subset Selection and the estimator $\hat{\boldsymbol{\eta}}_{1,k_{\max}}$ from Problem 1.

We begin with two assumptions on the data-generating process. The first is a mild stationarity and integrability condition.

Assumption 2 *The state s_t is stationary and ergodic and the random vectors $(\mathbf{f}_t, \mathbf{R}_t, \mathbf{c}_t)$ are measurable functions of s_t with finite second moment.*

Note that requiring $(\mathbf{f}_t, \mathbf{R}_t, \mathbf{c}_t)$ to be functions of s_t is without loss of generality, since we can always augment the state with $(\mathbf{f}_t, \mathbf{R}_t, \mathbf{c}_t)$.

The second regularity condition is a constraint qualification condition on the set of admissible SDFs. This is an unconditional counterpart of Assumption 1 reframed for returns. We maintain the notation from Appendix B.1, but now take expectations under the stationary distribution. Let $\check{\mathbf{R}}$ and $\check{\mathbf{c}}$ denote the vectors $\mathbf{R}$ and $\mathbf{c}$ with their first element (corresponding to the risk-free asset) removed. Analogously to (A5), define

$$\mathcal{E} = \{(\mathbb{E}[n] - 1, \mathbb{E}[n\check{\mathbf{R}}]' - (\mathbf{1} + \mathbb{E}[\check{\mathbf{c}}])', (\mathbf{1} - \mathbb{E}[\check{\mathbf{c}}])' - \mathbb{E}[n\check{\mathbf{R}}]')' : \ n \in L^2_+\} + \{0\} \times (\mathbb{R}_+)^{2(K_{\text{asset}}-1)}.$$

Assumption 3 *The set $\mathcal{E}$ contains the origin in its interior.*

Assumption 3 ensures that the finite-sample constraint $\mathbf{1} - \bar{\mathbf{c}}_T \leq \frac{1}{T}\sum_{t=1}^{T} n_t \mathbf{R}_t \leq \mathbf{1} + \bar{\mathbf{c}}_T$ is feasible with probability approaching one, precluding the existence of what Gospodinov et al. (2016) refer to as an in-sample arbitrage portfolio.

Notation: In what follows, we abbreviate "with probability approaching one" to wpa1. We also drop the $^{(1)}$ superscript on the distance measure $Q^{(1)}(\boldsymbol{\eta})$ in display (12) and

simply write $Q(\boldsymbol{\eta})$. The estimators $\hat{\boldsymbol{\eta}}_{1,K_{\text{factor}}}$ and $\hat{\boldsymbol{\eta}}_{1,k_{\max}}$ minimize its sample counterpart

$$\begin{aligned} Q_T(\boldsymbol{\eta}) = \min_{\{n_t\}_{t=1}^T} \frac{1}{T}\sum_{t=1}^T |n_t - \mathbf{f}_t'\boldsymbol{\eta}| \\ \text{subject to} \qquad \mathbf{1} - \bar{\mathbf{c}}_T \le \tfrac{1}{T}\textstyle\sum_{t=1}^T n_t \mathbf{R}_t \le \mathbf{1} + \bar{\mathbf{c}}_T, \\ n_t \ge 0, \qquad\qquad t = 1, \dots, T, \end{aligned}$$

without and with the sparsity constraint (13), respectively.

Proof of Theorem 2. At least one admissible SDF exists by Assumption 3, and so $Q(\boldsymbol{\eta}) < \infty$ for all $\boldsymbol{\eta}$. We establish further properties in a sequence of claims.

Claim 1: Q is convex.

Proof of Claim 1: Take $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2$. Let $\epsilon > 0$. Choose $n_1, n_2 \in L_+^2$ with $1 - \mathbb{E}[\mathbf{c}] \le \mathbb{E}[n_i \mathbf{R}] \le \mathbf{1} + \mathbb{E}[\mathbf{c}]$ so that $\mathbb{E}[|n_i - \mathbf{f}'\boldsymbol{\eta}_i|] \le Q(\boldsymbol{\eta}_i) \le \mathbb{E}[|n_i - \mathbf{f}'\boldsymbol{\eta}_i|] + \epsilon$ for $i = 1, 2$. Then, for $\tau \in (0,1)$, because $n := \tau n_1 + (1-\tau) n_2$ also satisfies $1 - \mathbb{E}[\mathbf{c}] \le \mathbb{E}[n\mathbf{R}] \le \mathbf{1} + \mathbb{E}[\mathbf{c}]$, we have

$$\begin{aligned} Q(\tau\boldsymbol{\eta}_1 + (1-\tau)\boldsymbol{\eta}_2) &\le \mathbb{E}[|\tau(n_1 - \mathbf{f}'\boldsymbol{\eta}_1) + (1-\tau)(n_2 - \mathbf{f}'\boldsymbol{\eta}_2)|] \\ &\le \tau\mathbb{E}[|n_1 - \mathbf{f}'\boldsymbol{\eta}_1|] + (1-\tau)\mathbb{E}[|n_2 - \mathbf{f}'\boldsymbol{\eta}_2|] \le \tau Q(\boldsymbol{\eta}_1) + (1-\tau) Q(\boldsymbol{\eta}_2) + \epsilon\,. \end{aligned}$$

Convexity follows because ϵ is arbitrary.

For the next claim we first introduce some notation. Let

$$q(\boldsymbol{\eta}, \boldsymbol{\xi}_+, \boldsymbol{\xi}_-) = \frac{1}{2}\left(\mathbb{E}\left[\min\left\{(\boldsymbol{\xi}_+ - \boldsymbol{\xi}_-)'\mathbf{R}, 1\right\}\mathbf{f}'\boldsymbol{\eta}\right] - \boldsymbol{\xi}_+'(\mathbf{1} + \mathbb{E}[\mathbf{c}]) + \boldsymbol{\xi}_-'(\mathbf{1} - \mathbb{E}[\mathbf{c}])\right),$$

where $\boldsymbol{\xi}_+, \boldsymbol{\xi}_- \in \mathbb{R}^{K_{\text{asset}}}$.

Claim 2: In what follows, let essinf denote essential infimum with respect to the stationary probability measure. We have the equivalent representation

$$Q(\boldsymbol{\eta}) = \max_{\substack{(\boldsymbol{\xi}_+, \boldsymbol{\xi}_-) \in (\mathbb{R}_+)^{2K_{\text{asset}}} : (\boldsymbol{\xi}_+)'\boldsymbol{\xi}_- = 0, \\ \text{essinf}((\boldsymbol{\xi}_+ - \boldsymbol{\xi}_-)'\mathbf{Z}) \ge -1}} q(\boldsymbol{\eta}, \boldsymbol{\xi}_+, \boldsymbol{\xi}_-),$$

where, for any $\boldsymbol{\eta}$, the set of maximizing $(\boldsymbol{\xi}_+, \boldsymbol{\xi}_-)$ is compact and nonempty.

Proof of Claim 2: The proof uses identical arguments to those of Lemma 2, replacing y with $\mathbf{f}'\boldsymbol{\eta}$, $\mathbf{Z}$ with $\mathbf{R}$, $\mathbf{p}_a$ with $1 + \mathbb{E}[\mathbf{c}]$, and $\mathbf{p}_b$ with $1 - \mathbb{E}[\mathbf{c}]$.

Claim 3: Wpa1, the equivalent representation

$$Q_T(\boldsymbol{\eta}) = \max_{\substack{(\boldsymbol{\xi}_+,\boldsymbol{\xi}_-)\in(\mathbb{R}_+)^{2K_{\text{asset}}}:(\boldsymbol{\xi}_+)'\boldsymbol{\xi}_-=0,\\ \min_{t\le T}((\boldsymbol{\xi}_+-\boldsymbol{\xi}_-)'\mathbf{Z}_t)\ge -1}} q_T(\boldsymbol{\eta},\boldsymbol{\xi}_+,\boldsymbol{\xi}_-)$$

holds uniformly in $\boldsymbol{\eta}$, where

$$q_T(\boldsymbol{\eta},\boldsymbol{\xi}_+,\boldsymbol{\xi}_-) = \frac{1}{2}\left(\frac{1}{T}\sum_{t=1}^{T}\min\left\{(\boldsymbol{\xi}_+-\boldsymbol{\xi}_-)'\mathbf{R}_t,1\right\}\mathbf{f}_t'\boldsymbol{\eta} - \boldsymbol{\xi}_+'(\mathbf{1}+\bar{\mathbf{c}}_T) + \boldsymbol{\xi}_-'(\mathbf{1}-\bar{\mathbf{c}}_T)\right).$$

Moreover, for each $\boldsymbol{\eta}$, the set of maximizing $(\boldsymbol{\xi}_+,\boldsymbol{\xi}_-)$ is nonempty and compact.

Proof of Claim 3: By Assumption 3, we can choose finitely many $n^{(1)},\ldots,n^{(M)}\in L_+^2$ and vectors $\mathbf{v}^{(1)},\ldots,\mathbf{v}^{(M)}\in\mathbb{R}^{2(K_{\text{asset}}-1)}$ such that the origin is in the interior of the convex hull of

$$\Big\{\left(\mathbb{E}[n^{(m)}-1],\mathbb{E}[n^{(m)}\check{\mathbf{R}}]' - (\mathbf{1}+\mathbb{E}[\check{\mathbf{c}}])', (\mathbf{1}-\mathbb{E}[\check{\mathbf{c}}])' - \mathbb{E}[n^{(m)}\check{\mathbf{R}}]'\right)' + (0,\mathbf{v}^{(m)\prime})' : \\ m=1,\ldots,M\Big\}.$$

It follows by Assumption 2 and the ergodic theorem that wpa1 the origin is in the interior of the convex hull of

$$\Big\{\left(\tfrac{1}{T}\textstyle\sum_{t=1}^{T} n_t^{(m)}-1, \tfrac{1}{T}\sum_{t=1}^{T} n_t^{(m)}\check{\mathbf{R}}_t' - (\mathbf{1}+\bar{\check{\mathbf{c}}}_T)', (\mathbf{1}-\bar{\check{\mathbf{c}}}_T)' - \tfrac{1}{T}\textstyle\sum_{t=1}^{T} n_t^{(m)}\check{\mathbf{R}}_t'\right)'\Big\} \\ + (0,\mathbf{v}^{(m)\prime})' :\ m=1,\ldots,M\Big\},$$

where $\bar{\check{\mathbf{c}}}_T = \frac{1}{T}\sum_{t=1}^{T}\check{\mathbf{c}}_t$. The result follows analogously to Claim 2, replacing the stationary probability measure with the empirical measure.

Claim 4: For any fixed $\boldsymbol{\eta}$, we have $|Q_T(\boldsymbol{\eta}) - Q(\boldsymbol{\eta})| \to_p 0$.

Proof of Claim 4: First, note by Claim 2 that $Q(\boldsymbol{\eta}) = q(\boldsymbol{\eta},\boldsymbol{\xi}_+,\boldsymbol{\xi}_-)$ for some vectors $(\boldsymbol{\xi}_+,\boldsymbol{\xi}_-)\in(\mathbb{R}_+)^{2K_{\text{asset}}}$ for which $(\boldsymbol{\xi}_+)'\boldsymbol{\xi}_- = 0$ and $\operatorname{ess\,inf}((\boldsymbol{\xi}_+-\boldsymbol{\xi}_-)'\mathbf{R}) \ge -1$. Then, by the ergodic theorem, we have that $q_T(\boldsymbol{\eta},\boldsymbol{\xi}_+,\boldsymbol{\xi}_-)\to_p q(\boldsymbol{\eta},\boldsymbol{\xi}_+,\boldsymbol{\xi}_-) = Q(\boldsymbol{\eta})$. Also note that $\min_{t\le T}(\boldsymbol{\xi}_+-\boldsymbol{\xi}_-)'\mathbf{R}_t \ge \operatorname{ess\,inf}((\boldsymbol{\xi}_+-\boldsymbol{\xi}_-)'\mathbf{R})$ with probability one. It then follows by Claim 3 that wpa1,

$$Q_T(\boldsymbol{\eta}) \ge q_T(\boldsymbol{\eta},\boldsymbol{\xi}_+,\boldsymbol{\xi}_-),$$

from which we conclude that $Q_T(\boldsymbol{\eta}) \ge Q(\boldsymbol{\eta}) + o_p(1)$.

To prove the lower bound, fix $\epsilon > 0$ and choose $n^{(0)} \in L^2_+$ with $\mathbf{1} - \mathbb{E}[\mathbf{c}] \leq \mathbb{E}[n^{(0)}\mathbf{R}] \leq 1 + \mathbb{E}[\mathbf{c}]$ such that $\mathbb{E}[|n^{(0)} - \boldsymbol{\eta}'\mathbf{f}|] \leq Q(\boldsymbol{\eta}) + \epsilon$. Let $n^{(1)}, \ldots, n^{(M)}$ and $\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(M)}$ be as in the proof of Claim 3. But note that $\frac{1}{T}\sum_{t=1}^T n_t^{(0)}\mathbf{R}_t \to_p \mathbb{E}[n^{(0)}\mathbf{R}]$ and $\bar{\mathbf{c}}_T \to_p \mathbb{E}[\mathbf{c}]$ by the ergodic theorem. This means that wpa1 there exists a convex combination $\hat{n} = \sum_{m=0}^M \hat{\tau}_m n^{(m)}$ of $n^{(0)}, \ldots, n^{(M)}$ that satisfies $1 - \bar{\mathbf{c}}_T \leq \frac{1}{T}\sum_{t=1}^T \hat{n}_t\mathbf{R}_t \leq \mathbf{1} + \bar{\mathbf{c}}_T$ and where the weight $\hat{\tau}_0 \to_p 1$. Then since $\hat{n}$ is feasible for the sample problem, we have

$$\begin{aligned} Q_T(\boldsymbol{\eta}) \leq \frac{1}{T}\sum_{t=1}^T |\hat{n}_t - \boldsymbol{\eta}'\mathbf{f}_t|] &\leq \frac{1}{T}\sum_{t=1}^T |n_t^{(0)} - \boldsymbol{\eta}'\mathbf{f}_t| + \frac{1}{T}\sum_{t=1}^T |\hat{n}_t - n_t^{(0)}| \\ &\leq \frac{1}{T}\sum_{t=1}^T |n_t^{(0)} - \boldsymbol{\eta}'\mathbf{f}_t| + (1 - \hat{\tau}_0)\sum_{m=1}^M \left(\frac{1}{T}\sum_{t=1}^T |n_t^{(m)} - n_t^{(0)}|\right) \\ &\to_p \mathbb{E}[|n^{(0)} - \boldsymbol{\eta}'\mathbf{f}|] \leq Q(\boldsymbol{\eta}) + \epsilon\,. \end{aligned}$$

Hence, $Q_T(\boldsymbol{\eta}) \leq Q(\boldsymbol{\eta}) + o_p(1)$.

Claim 5: Fix any bounded subset $\mathcal{B}$ of $\mathbb{R}^{K_{\text{factor}}}$. Then, for all $\epsilon > 0$, we have

$$\Pr\left(\sup_{\eta \in \mathcal{B}}(Q(\boldsymbol{\eta}) - Q_T(\boldsymbol{\eta})) > 2\epsilon\right) \to 0\,.$$

Proof of Claim 5: The argument follows Section 6 of Pollard (1991). Note from the definition of $Q_T(\boldsymbol{\eta})$ that, whenever there exists $\{n_t\}_{t=1}^T$ with $1 - \bar{\mathbf{c}}_T \leq \frac{1}{T}\sum_{t=1}^T n_t\mathbf{R}_t \leq \mathbf{1} + \bar{\mathbf{c}}_T$ (which is true wpa1 in view of the proof of Claim 3), we have that for any $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2 \in \mathbb{R}^{K_{\text{factor}}}$,

$$|Q_T(\boldsymbol{\eta}_1) - Q_T(\boldsymbol{\eta}_2)| \leq \left(\frac{1}{T}\sum_{t=1}^T \|\mathbf{f}_t\|\right)\|\boldsymbol{\eta}_1 - \boldsymbol{\eta}_2\|,$$

where $\frac{1}{T}\sum_{t=1}^T \|\mathbf{f}_t\| \leq C$ wpa1 by Assumption 2. Cover $\mathcal{B}$ by a finite grid of points G separated by at most $\delta = \epsilon/C$, so that the cubes formed by joining points in G have their edges all parallel to the coordinate axes. Take any $\boldsymbol{\eta} \in \mathcal{B}$ and write it as $\boldsymbol{\eta} = \sum_i \tau_i\boldsymbol{\eta}_i$, where the $\boldsymbol{\eta}_i \in G$ are vertices of the δ-cube containing $\boldsymbol{\eta}$. Then by convexity (Claim 1),

whenever $\frac{1}{T}\sum_{t=1}^{T}\|\mathbf{f}_t\| \leq C$, we have

$$\begin{aligned} Q(\boldsymbol{\eta}) \leq \sum_i \tau_i Q(\boldsymbol{\eta}_i) &= \sum_i \tau_i Q_T(\boldsymbol{\eta}_i) + \tau_i(Q(\boldsymbol{\eta}_i) - Q_T(\boldsymbol{\eta}_i)) \\ &\leq Q_T(\boldsymbol{\eta}) + \max_i |Q_T(\boldsymbol{\eta}_i) - Q_T(\boldsymbol{\eta})| + \max_i (Q(\boldsymbol{\eta}_i) - Q_T(\boldsymbol{\eta}_i)) \\ &\leq Q_T(\boldsymbol{\eta}) + \epsilon + \max_{\tilde{\boldsymbol{\eta}} \in G}(Q(\tilde{\boldsymbol{\eta}}) - Q_T(\tilde{\boldsymbol{\eta}})). \end{aligned}$$

But $\max_{\tilde{\boldsymbol{\eta}}\in G}(Q(\tilde{\boldsymbol{\eta}}) - Q_T(\tilde{\boldsymbol{\eta}})) \leq \epsilon$ wpa1 by Claim 4, so we have that

$$\sup_{\boldsymbol{\eta}\in\mathcal{B}}(Q(\boldsymbol{\eta}) - Q_T(\boldsymbol{\eta})) \leq 2\epsilon \quad \text{wpa1},$$

which completes the proof of Claim 5.

With these claims established, we now turn to proving the first result. Note by construction that the parameter space $[-B, B]^{K_{\text{factor}}}$ for $\boldsymbol{\eta}$ is compact. Hence, by Claims 4 and 5 and definition of $\hat{\boldsymbol{\eta}}_{1,K_{\text{factor}}}$ that for any $\epsilon > 0$, the inequalities

$$Q(\boldsymbol{\eta}_{1,K_{\text{factor}}}) + \epsilon \geq Q_T(\boldsymbol{\eta}_{1,K_{\text{factor}}}) \geq Q_T(\hat{\boldsymbol{\eta}}_{1,K_{\text{factor}}}) \geq Q(\hat{\boldsymbol{\eta}}_{1,K_{\text{factor}}}) - \epsilon$$

hold wpa1.

For the second result, note that as Q is finite-valued and convex on $\mathbb{R}^{K_{\text{factor}}}$ by Claim 1, it is continuous (by Rockafellar (1970, Corollary 10.1.1)) on $[-B, B]^{K_{\text{factor}}}$. As Q is uniquely minimized at $\boldsymbol{\eta}_{1,K_{\text{factor}}}$, we have $\inf_{\boldsymbol{\eta}:\|\boldsymbol{\eta}-\boldsymbol{\eta}_{1,K_{\text{factor}}}\|>\delta} Q(\boldsymbol{\eta}) > Q(\boldsymbol{\eta}_{1,K_{\text{factor}}})$ for all $\delta > 0$. Hence, in view of the first claim, we have $Q(\hat{\boldsymbol{\eta}}_{1,K_{\text{factor}}}) \leq \inf_{\boldsymbol{\eta}:\|\boldsymbol{\eta}-\boldsymbol{\eta}_{1,K_{\text{factor}}}\|>\delta} Q(\boldsymbol{\eta})$ wpa1, and so $\|\hat{\boldsymbol{\eta}}_{1,K_{\text{factor}}} - \boldsymbol{\eta}_{1,K_{\text{factor}}}\| \leq \delta$ wpa1. ■

Proof of Theorem 1. Let $\mathcal{B}_{k_{\max}}$ denote the subset of $\mathbb{R}^{K_{\text{factor}}}$ where at most $k_{\max}$ elements of $\boldsymbol{\eta}$ are non-zero. Note that we may express $\mathcal{B}_{k_{\max}}$ as the union of sets $\mathcal{B}_{(\ell)}$, $\ell = 1, \ldots, L$ with $L = \sum_{k=1}^{k_{\max}} \binom{K_{\text{factor}}}{k}$ where each $\boldsymbol{\eta} \in \mathcal{B}_{(\ell)}$ has the same $K_{\text{factor}} - k$ elements fixed to zero and the remaining $k \leq k_{\max}$ elements are allowed to vary. We let $\ell^\star_{k_{\max}}$ denote the index of the set to which $\boldsymbol{\eta}_{1,k_{\max}}$ belongs, indicating the best subset of factors. Similarly, we let $\hat{\ell}^\star_{k_{\max}}$ denote the index of the set to which $\hat{\boldsymbol{\eta}}_{1,k_{\max}}$ belongs, indicating the best subset selection.

First, note that $\hat{\boldsymbol{\eta}}_{1,k_{\max}}$ can be expressed as

$$\hat{\boldsymbol{\eta}}_{1,k_{\max}} := \hat{\boldsymbol{\eta}}_{(\hat{\ell}^\star_{k_{\max}})}, \quad Q_T(\hat{\boldsymbol{\eta}}_{(\hat{\ell}^\star_{k_{\max}})}) \leq Q_T(\hat{\boldsymbol{\eta}}_{(\ell)}),\ \ell = 1, \ldots, L, \quad \hat{\boldsymbol{\eta}}_{(\ell)} = \arg\min_{\boldsymbol{\eta}\in\mathcal{B}_{(\ell)}} Q_T(\boldsymbol{\eta})\,.$$

Similarly,

$$\boldsymbol{\eta}_{1,k_{\max}} := \boldsymbol{\eta}_{(\ell^\star_{k_{\max}})}, \quad Q(\boldsymbol{\eta}_{(\ell^\star_{k_{\max}})}) \leq Q(\boldsymbol{\eta}_{(\ell)}),\ \ell = 1, \ldots, L, \quad \boldsymbol{\eta}_{(\ell)} = \arg\min_{\boldsymbol{\eta} \in \mathcal{B}_{(\ell)}} Q(\boldsymbol{\eta})\,.$$

Arguing as in the proof of Theorem 2, for any $\epsilon > 0$ we have $Q_T(\boldsymbol{\eta}_{(\ell)}) \leq Q(\boldsymbol{\eta}_{(\ell)}) + \epsilon$ wpa1 for all $\ell = 1, \ldots, L$ (by Claim 4) and $Q_T(\hat{\boldsymbol{\eta}}_{(\ell)}) \geq Q(\hat{\boldsymbol{\eta}}_{(\ell)}) - \epsilon$ for all $\ell = 1, \ldots, L$ wpa1 (by Claim 5). Hence, wpa1 we have

$$\begin{aligned} Q(\boldsymbol{\eta}_{1,k_{\max}}) \equiv Q(\boldsymbol{\eta}_{(\ell^\star)}) \leq Q(\hat{\boldsymbol{\eta}}_{1,k_{\max}}) \equiv Q(\hat{\boldsymbol{\eta}}_{(\hat{\ell}^\star)}) &\leq Q_T(\hat{\boldsymbol{\eta}}_{(\hat{\ell}^\star)}) + \epsilon \\ &\leq Q_T(\boldsymbol{\eta}_{(\ell^\star)}) + \epsilon \leq Q(\boldsymbol{\eta}_{(\ell^\star)}) + 2\epsilon \equiv Q(\hat{\boldsymbol{\eta}}_{1,k_{\max}}) + 2\epsilon, \end{aligned}$$

which proves (16).

We now establish result (17). To this end, note that establishing $\hat{\boldsymbol{\chi}}_{1,k_{\max}} = \boldsymbol{\chi}_{1,k_{\max}}$ is equivalent to establishing $\hat{\ell}^\star = \ell^\star$. As Q is uniquely minimized over $\mathcal{B}_{k_{\max}}$ at $\boldsymbol{\eta}_{1,k_{\max}}$, we have $Q(\boldsymbol{\eta}_{1,k_{\max}}) < \min_{\ell \neq \ell^\star} Q(\boldsymbol{\eta}_{(\ell)})$. Hence, by (16), we have $Q(\hat{\boldsymbol{\eta}}_{1,k_{\max}}) < \min_{\ell \neq \ell^\star} Q(\boldsymbol{\eta}_{(\ell)})$ wpa1, and, hence, that $\hat{\ell}^\star = \ell^\star$ wpa1.

Finally, consistency of $\hat{\boldsymbol{\eta}}_{k_{\max}}$ follows analogously to the proof of Theorem 2. ■

References

Bernardo, A. and O. Ledoit (2000). Gain, loss and asset pricing. *Journal of Political Economy 108*, 144–172.

Bonnans, J. and A. Shapiro (2000). *Perturbation Analysis of Optimization Problems*. Springer.

Cochrane, J. (1996). A cross-sectional test of an investment-based asset pricing model. *Journal of Political Economy 104*, 572–621.

Gospodinov, N., R. Kan, and C. Robotti (2016). On the properties of the constrained Hansen-Jagannathan distance. *Journal of Empirical Finance 36*, 121–150.

Hansen, L. and R. Jagannathan (1991). Implications of security market data for dynamic economies. *Journal of Political Economy 99*, 225–261.

Hansen, L. and R. Jagannathan (1997). Assessing specification errors in stochastic discount factor models. *Journal of Finance 52*(2), 557–590.

Lettau, M. and S. Ludvigson (2001). Resurrecting the (C)CAPM: A cross-sectional test when risk premia are time-varying. *Journal of Political Economy 109*, 1238–1287.

Pollard, D. (1991). Asymptotics for least absolute deviation regression estimators. *Econometric Theory* 7(2), 186–199.

Rockafellar, R. (1970). *Convex Analysis*. Princeton University Press.

Rockafellar, R. and R. Wets (2009). *Variational Analysis*. Springer.