Geometry of the SDF
The Factor Zoo as Structured Ambiguity

Andrew Perry*

DRAFT: April 26, 2026
DO NOT DISTRIBUTE

ABSTRACT

I develop an information-geometric framework for selecting priced risk factors from a high-dimensional candidate set. The stochastic discount factor is cast as a point on a statistical manifold of exponential-affine pricing kernels, and the manifold's dual flatness yields an exact Pythagorean decomposition of Kullback-Leibler divergence across nested SDFs. The decomposition drives a greedy forward-selection algorithm that is isomorphic to the portfolio problem of an investor with multiplier preferences. The investor's ranking of candidates by worst-case welfare loss coincides with the statistical ranking by KL divergence, and the Hansen-Jagannathan distance emerges as a leading-order approximation. A Wald test on estimated risk prices and an entropy bound calibrated to the investor's detection-error probability provide distinct stopping criteria. Applied to a rich set of asset returns, the algorithm selects a parsimonious SDF spanning roughly nine thematic categories. Debt issuance, quality, and profit growth form a stable core; accruals is absent from every specification. Selected factors derive a substantial share of their pricing content from non-Gaussian features of the return distribution.

JEL classification: C13, C52, C58, C61, D81, G11, G12

*Andrew Perry is with the Haas School of Business, U.C. Berkeley. Acknowledgments:

I. Introduction

The cross-section of expected stock returns remains unsettled. Cochrane (2011) characterized the state of empirical asset pricing as having a "zoo of [...] factors" and the zoo has only continued to expand. Harvey, Liu, and Zhu (2015) documents over 300 factors published in top journals. The proliferation is not merely troubling; it is epistemically disabling. Every test of a candidate factor is conducted against a benchmark, and the benchmark itself is contested. Statistical significance is therefore conditional on which fraction of the zoo one chooses to control for, which makes the evaluation problem circular. The zoo grows because there is no benchmark model. There is no procedure for factor selection that simultaneously respects economic theory and provides criteria for which factors meaningfully contribute to pricing. This paper closes that gap.

I develop an information-geometric framework that recasts factor selection as the traversal of a statistical manifold of admissible pricing kernels. The manifold $\mathbb{M}$ consists of exponential-affine stochastic discount factors (SDF), which form an exponential family whose dual flatness equips the space with a canonical divergence measure, the Kullback-Leibler (KL) divergence. I show that the traversal of $\mathbb{M}$ is isomorphic to the portfolio problem of an investor with multiplier preferences in the sense of Hansen and Sargent (2001), and that an exact Pythagorean decomposition of KL divergence across nested SDFs drives a greedy forward-selection algorithm. The algorithm is not a heuristic approximation. It is an exact manifold traversal at each step.

The mechanism rests on the dual flatness of $\mathbb{M}$. The exponential family of SDFs generates an e-flat manifold, and the moment-restricted space is m-flat under the mixture connection. This dual flatness yields an exact Pythagorean decomposition of KL divergence. This allows for a decomposition of residual pricing error to the true (unobservable) SDF. Therefore, minimizing the residual error after adding candidate factor $\mathbf{g}$ is mathematically equivalent to maximizing the divergence of the new factor relative to the baseline set of factors $\mathbb{S}$.

The economic isomorphism follows from the same exponential-tilt structure. A multiplier-preference investor's worst-case distortion h^* is itself an exponential tilt and lives on $\mathbb{M}$. The investor's ranking of candidate factors by worst-case welfare loss coincides with the purely statistical ranking by KL divergence. The two are not aligned by assumption. They coincide as a consequence of the e-flat geometry interacting with the exponential form of the worst-case distortion. The Hansen and Jagannathan (1997) (HJ) distance is recovered from this framework as a second order approximation. Unlike the HJ metric, KL divergence respects the higher order moments of the return distribution. The geometric criterion is therefore crash-sensitive in a way that mean-variance pricing is not, prioritizing factors whose pricing content concentrates in the left tail of the return distribution where marginal utility is high.

The framework yields two distinct stopping criteria. A Wald test on the estimated risk price $\hat{\theta}_{\mathbf{g}}$ screens for statistical detectability and an entropy bound $\bar{\eta}$, calibrated from the investor's detection-error probability ε as in Hansen and Sargent (2010), screens for economic materiality. Each guards a distinct failure mode. The Wald test prevents false inclusion driven by sampling noise; the entropy bound prevents kitchen-sink overfitting that admits factors whose genuine pricing signal falls below

the threshold the adversary's distortion can rationalize. Neither criterion can substitute for the other.

I apply the procedure to a candidate universe of 157 factors, comprising the 151 characteristic-based long-short factors of Jensen, Kelly, and Pedersen (2023) augmented with the six standard Fama-French factors, and price them against 73 principal components of 4,686 characteristic-sorted decile portfolios spanning 666 months from Chen and Zimmermann (2022). At the baseline calibration, the algorithm selects 14 factors spanning 9 thematic categories.

Across a sweep of parameter values a stable core of themes emerges. Debt issuance and quality each enter in 6 of 7 trials, profit growth in 5 of 7. Seasonality, value, the equity premium, and size form a middle tier represented in 3 to 4 trials. Accruals never enters the selected model, despite the candidate universe containing 6 accruals factors, a pattern consistent with substantial overlap between the accruals signal and the quality theme that does enter. Path dependence operates primarily within themes through rotation among near-substitutes. Between-theme orderings are comparatively stable. The algorithm therefore recovers a consistent ranking of the cross-sectional pricing structure rather than an arbitrary subset of the candidate universe.

The selected factors derive a substantial share of their pricing content from non-Gaussian features of the return distribution. The KL contribution of each factor decomposes additively into a Gaussian component, which captures the mean-variance pricing content available to a standard SDF, and a higher-order component, which captures skewness, kurtosis, and tail structure. This pattern is the structural counterpart of the robust investor's sensitivity to model misspecification. The adversarial tilt concentrates probability mass on tail states, and factors whose pricing information loads on those states are correspondingly informative.

The framework relates closely to several strands of prior work. Feng, Giglio, and Xiu (2020) develop a double-selection LASSO procedure that evaluates the marginal contribution of new factors against a high-dimensional control set. Their criterion is statistical, namely significance after controlling for redundancy, whereas the criterion here is economic, derived from the welfare loss of an investor under ambiguity, and it coincides with the statistical criterion. Lettau and Pelger (2020) modify the construction of factors through risk-premium PCA to incorporate cross-sectional pricing information. This paper takes the candidate factors as given and modifies the selection rule.

Ghosh, Julliard, and Taylor (2016) introduce KL divergence to asset pricing as a tool for nonparametric estimation of an unobservable SDF component in the consumption CAPM framework. Their KL is a fitting criterion for a single SDF; here KL is a ranking criterion across nested SDFs. ichi Amari (2020) supplies the dually flat manifold structure and the generalized Pythagorean theorem that make the decomposition exact. Kitamura and Stutzer (1997) develop the Minimum Cross-Entropy (MCE) estimator that I use to estimate the exponential-affine SDF. Prior implementations in asset pricing solve the inner problem for fixed risk prices or restrict the parameter space. This paper is, to my knowledge, the first to estimate the full nested exponential-affine SDF over both Lagrange multipliers and risk prices under the pricing restrictions.

The paper is organized as follows: Section II develops the geometric and economic theories,

Section III describes the forward selection algorithm and the architecture of the MCE optimizer, Section IV gives the result of the method applied to progressively richer universe of assets, and Section V concludes.

II. Theory

The central question of empirical asset pricing is which risks the market prices, and how severely. I approach this question geometrically. The search for the true stochastic discount factor is equivalent to a search through a space of probability measures, and that space has a structure defined by the distribution of asset returns. My goal in this section is to make that structure precise, exploit it, and show that it delivers both a tractable estimation algorithm and a complete economic theory of why investors hold the factors they do, and no more. The factor zoo is not tamed by an arbitrary cutoff; it is tamed by a criterion that a rational agent under uncertainty would independently arrive at.

The section proceeds in two parts. Section II.A develops the statistical geometry of the SDF manifold. I define the manifold of admissible pricing kernels, equip it with a divergence measure, derive the projection geometry that drives the selection algorithm, and extract a statistical stopping rule that determines when a candidate factor's risk price is detectable. This machinery delivers a complete algorithm and a significance filter, but it does not provide an economic criterion for when residual mispricing is tolerable. Section II.B supplies that criterion. I introduce a representative investor with multiplier preferences in the sense of Hansen and Sargent (2001, 2020), derive the robust SDF and the misspecification premium that governs pricing anomalies, and show that the investor's preference-based ranking of candidate factors by welfare loss is identical to the purely statistical ranking by KL divergence. The investor's welfare analysis produces an economic stopping rule that determines when the pricing kernel should cease expanding. The section closes by observing that the geometric algorithm and the investor's equilibrium selection describe the same phenomenon from two vantage points.

A. Statistical Geometry

The search for the true stochastic discount factor can be cast as a projection of the true data-generating process onto a statistical manifold of admissible pricing models. By embedding asset pricing into the framework of information geometry, the heuristic practice of hunting for factors is replaced by a principled traversal of a statistical manifold. This subsection formalizes the space of admissible SDFs, derives the projection geometry that drives the selection algorithm, and establishes a statistical stopping rule for the greedy procedure.

Let $(\Omega, \mathfrak{F}, \mathcal{P})$ be a probability space representing the physical economy, supporting a vector of asset excess returns $\mathbf{r}_t^e \in \mathbb{R}^N$ and a vector of candidate risk factors $\mathbf{f}_t \in \mathbb{R}^K$. Standard asset pricing theory posits the existence of a strictly positive stochastic discount factor m_t such that

$$\mathbf{E}_{\mathcal{P}}\left[m_t \mathbf{r}_t^e\right] = \mathbf{0}_N. \tag{1}$$

The family of admissible pricing kernels must satisfy several requirements simultaneously. The SDF must be strictly positive everywhere, both to enforce no-arbitrage and to ensure that it defines a valid probability measure via its Radon-Nikodym interpretation. It must be closed under projection, so that restricting the factor set produces a valid SDF within the same family. It must admit a canonical divergence, enabling a well-defined measure of distance between nested models. And it must nest cleanly, so that an augmented model embeds the restricted model as a special case. These requirements are not independent. The unique family of strictly positive distributions satisfying all four is the exponential family.

I restrict attention to the family of exponential-affine pricing kernels

$$m_t(\boldsymbol{\theta}) := \exp\left(-\boldsymbol{\theta}^\top \mathbf{f}_t - \psi(\boldsymbol{\theta})\right), \tag{2}$$

where $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^K$ is the vector of factor risk prices and $\psi(\boldsymbol{\theta})$ is the cumulant generating function

$$\psi(\boldsymbol{\theta}) := \ln\left(\mathbf{E}_{\mathcal{P}}\left[\exp\left(-\boldsymbol{\theta}^\top \mathbf{f}_t\right)\right]\right). \tag{3}$$

The normalization in (3) ensures that $\mathbf{E}_{\mathcal{P}}\left[m_t(\boldsymbol{\theta})\right] = 1$, so that $m_t(\boldsymbol{\theta})$ can be interpreted as the Radon-Nikodym derivative inducing a risk-neutral probability measure $\mathcal{Q}(\boldsymbol{\theta})$. The construction guarantees $m_t(\boldsymbol{\theta}) > 0$ for all valid parameters, placing every candidate model within the open simplex of probability distributions where information geometry is valid. Economically, this positivity enforces the no-arbitrage condition by construction, precluding the negative state prices that plague linear factor models during market stress.

Assumption 1 (Regularity): *The joint distribution of* $(\mathbf{f}, \mathbf{r}^e)$ *under the physical measure* $\mathcal{P}$ *has a well-defined moment generating function for* $\mathbf{f}$ *in a neighborhood of the origin, and the support of* $\mathbf{f}$ *under* $\mathcal{P}$ *is not contained in any proper affine subspace of* $\mathbb{R}^K$.

Under the finite-moment and full-rank conditions of Assumption 1, the set

$$\mathbb{M} = \left\{\mathcal{Q}(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^K\right\}$$

forms a K-dimensional statistical manifold, which I term the *SDF manifold*. In the language of information geometry, $\mathbb{M}$ constitutes an exponential family of distributions. This structure is not merely a statistical convenience; it endows the space of pricing models with a rigorous form of geometric flatness.

A statistical manifold is flat if there exists a coordinate system in which geodesics, the shortest paths between distributions, are straight lines. The manifold $\mathbb{M}$ is specifically e-flat (*exponentially flat*) because its geodesics are linear in the natural parameters $\boldsymbol{\theta}$ ichi Amari (2020). This linearity

in log-probability space is the property that enables orthogonal decomposition of pricing errors, analogous to projecting vectors in Euclidean space, but on a surface whose geometry is dictated by the distribution of asset returns.

To quantify the distance between the true pricing kernel and a candidate model, I employ the Kullback-Leibler (KL) divergence, also known as relative entropy. This measure generalizes Shannon entropy Shannon (1948) by measuring the expected informational loss incurred when the unobservable true risk-neutral measure $\mathcal{Q}^\dagger$ is approximated by a candidate model $\mathcal{Q}(\boldsymbol{\theta})$.[1]

Let $\mathcal{Q}^\dagger$ denote the true, unobservable risk-neutral measure that correctly prices all assets. The divergence from the truth to a candidate model $\mathcal{Q}(\boldsymbol{\theta}) \in \mathbb{M}$ is

$$D_{KL}(\mathcal{Q}^\dagger || \mathcal{Q}(\boldsymbol{\theta})) := \int_\Omega \ln \frac{d\mathcal{Q}^\dagger}{d\mathcal{Q}(\boldsymbol{\theta})} d\mathcal{Q}^\dagger = \mathbf{E}_{\mathcal{Q}^\dagger}\left[\ln\left(\frac{d\mathcal{Q}^\dagger}{d\mathcal{Q}(\boldsymbol{\theta})}\right)\right]. \tag{4}$$

Minimizing the KL divergence minimizes the expected informational loss incurred when approximating the true pricing distribution $\mathcal{Q}^\dagger$ with the model $\mathcal{Q}(\boldsymbol{\theta})$. Although statistically equivalent to maximizing the expected log-likelihood, this criterion serves a distinct economic purpose: it selects the pricing kernel that least distorts state probabilities relative to investors' true marginal utilities. The estimator naturally aligns risk premia with observed returns while penalizing unnecessary complexity in the factor representation.

The KL divergence is not a true metric in the topological sense, satisfying neither symmetry nor the triangle inequality. It is, however, the canonical divergence for exponential families. This status is not accidental. The variational preference axioms of Maccheroni, Marinacci, and Rustichini (2006) establish that KL divergence is the unique measure of model distance, up to affine transformations, consistent with coherent decision-making under ambiguity. The exponential-entropy structure is dictated by axioms, not assumed for tractability.

The geometry of the SDF manifold $\mathbb{M}$ is governed by a dual structure that is fundamental to the estimation strategy. The manifold $\mathbb{M}$ is e-flat under the exponential connection, while the space of mixture distributions is m-flat under the mixture connection. A central result in information geometry ichi Amari (2020) establishes that this dual flatness guarantees the projection of the true distribution onto the model manifold is unique and satisfies a generalized Pythagorean theorem.

The econometric goal is to identify the model $\mathcal{Q}^{\mathbb{S}}$ within a restricted factor span $\mathbb{S} \subseteq \mathbb{F}$ that best approximates the true pricing kernel $\mathcal{Q}^\dagger$. This optimization defines $\mathcal{Q}^{\mathbb{S}}$ as the m-projection of the true measure $\mathcal{Q}^\dagger$ onto the e-flat submanifold $\mathbb{M}^{\mathbb{S}}$.

Consider a nested factor structure where we augment a base set of factors $\mathbb{S}$ with a candidate factor $\mathbf{g}$, so that $\mathbb{M}^{\mathbb{S}} \subset \mathbb{M}^{(\mathbb{S},\mathbf{g})}$. Since both the restricted measure $\mathcal{Q}^{\mathbb{S}}$ and the augmented measure $\mathcal{Q}^{(\mathbb{S},\mathbf{g})}$ are absolutely continuous with respect to $\mathcal{P}$, the Radon-Nikodym derivative of the augmented measure with respect to the restricted measure exists and equals the ratio of their pricing kernels.

[1] Shannon entropy quantifies the absolute uncertainty of a single probability distribution. The KL divergence generalizes this to the relative uncertainty between two distributions, measuring the excess information required to describe the true distribution using an approximating model. The KL divergence is foundational across disciplines, from information theory and coding theory to machine learning and statistical physics.

Because $\mathbb{M}^{\mathbb{S}}$ is a submanifold of $\mathbb{M}^{(\mathbb{S},\mathbf{g})}$ and both are e-flat, the generalized Pythagorean theorem for dually flat manifolds ichi Amari (2020) applies. For the chain of distributions $\mathcal{Q}^{\dagger}$ (true), $\mathcal{Q}^{(\mathbb{S},\mathbf{g})}$ (augmented model), and $\mathcal{Q}^{\mathbb{S}}$ (restricted model), the divergence decomposes exactly as

$$D_{KL}(\mathcal{Q}^{\dagger}||\mathcal{Q}^{\mathbb{S}}) = D_{KL}(\mathcal{Q}^{\dagger}||\mathcal{Q}^{(\mathbb{S},\mathbf{g})}) + D_{KL}(\mathcal{Q}^{(\mathbb{S},\mathbf{g})}||\mathcal{Q}^{\mathbb{S}}). \tag{5}$$

This identity is the theoretical cornerstone of the approach. It admits a transparent geometric interpretation. The restricted model $\mathcal{Q}^{\mathbb{S}}$ is the projection of the true measure $\mathcal{Q}^{\dagger}$ onto the submanifold $\mathbb{M}^{\mathbb{S}}$. When the submanifold is expanded by one dimension through the addition of factor $\mathbf{g}$, the new projection $\mathcal{Q}^{(\mathbb{S},\mathbf{g})}$ lies closer to the truth. The three points, $\mathcal{Q}^{\dagger}$, $\mathcal{Q}^{(\mathbb{S},\mathbf{g})}$, and $\mathcal{Q}^{\mathbb{S}}$, form a right triangle on the manifold in the information-geometric sense, with the two legs orthogonal. The decomposition is exact due to the e-flat structure of $\mathbb{M}$.

The economic content is equally direct. The left-hand side is the total mispricing under the restricted model, fixed once $\mathbb{S}$ is chosen. The first term on the right is the residual mispricing after adding $\mathbf{g}$, the pricing error that remains unexplained. The second term is the information gain from $\mathbf{g}$, the mispricing that the new factor resolves. Since the total mispricing is fixed *a priori*, minimizing the residual mispricing is mathematically equivalent to maximizing the information gain

$$\underset{\mathbf{g}\in\mathbb{F}\setminus\mathbb{S}}{\arg\min}\, D_{KL}(\mathcal{Q}^{\dagger}\|\mathcal{Q}^{(\mathbb{S},\mathbf{g})}) \;\equiv\; \underset{\mathbf{g}\in\mathbb{F}\setminus\mathbb{S}}{\arg\max}\, D_{KL}(\mathcal{Q}^{(\mathbb{S},\mathbf{g})}\|\mathcal{Q}^{\mathbb{S}}). \tag{6}$$

This equivalence motivates greedy forward selection: at each step, add the factor that maximizes the divergence between the augmented and baseline models. The equivalence in (6) holds at each selection step for the current restricted model $\mathcal{Q}^{\mathbb{S}}$. Across steps, as $\mathbb{S}$ expands and the model is re-estimated, the total error $D_{KL}(\mathcal{Q}^{\dagger}||\mathcal{Q}^{\mathbb{S}})$ itself changes, so the decomposition does not telescope into a cumulative sum.

On the e-flat manifold, the KL divergence between two exponential-affine SDFs admits a representation as the Bregman divergence generated by the cumulant generating function $\psi(\boldsymbol{\theta})$

$$D_{KL}\left(\mathcal{Q}^{(\mathbb{S},\mathbf{g})}||\mathcal{Q}^{\mathbb{S}}\right) = \psi\left(\boldsymbol{\theta}^{\mathbb{S}}\right) - \psi\left(\boldsymbol{\theta}^{(\mathbb{S},\mathbf{g})}\right) + \nabla_{\boldsymbol{\theta}}\psi\left(\boldsymbol{\theta}^{(\mathbb{S},\mathbf{g})}\right)^{\top}\left(\boldsymbol{\theta}^{(\mathbb{S},\mathbf{g})} - \boldsymbol{\theta}^{\mathbb{S}}\right). \tag{7}$$

The Bregman form makes the pricing content of the divergence transparent. The first two terms, $\psi(\boldsymbol{\theta}^{\mathbb{S}}) - \psi(\boldsymbol{\theta}^{(\mathbb{S},\mathbf{g})})$, measure the change in the log-normalizing constant when the new factor is added to the SDF. Since $\psi(\boldsymbol{\theta})$ is precisely the adjustment that ensures the pricing kernel integrates to unity, this difference reflects how much the model must redistribute probability mass across states of the world to maintain no-arbitrage once the new risk dimension is priced. A large change indicates that the new factor substantially reshapes the state-price density; a small change indicates that the existing factors already account for most of the relevant variation in state prices. The third term is an inner product between the gradient $\nabla_{\boldsymbol{\theta}}\psi(\boldsymbol{\theta}^{(\mathbb{S},\mathbf{g})})$, which is the vector of expected factor payoffs under the augmented model's risk-neutral measure, and the parameter displacement $\boldsymbol{\theta}^{(\mathbb{S},\mathbf{g})} - \boldsymbol{\theta}^{\mathbb{S}}$. Together, these terms decompose the information gain into a normalization effect and a repricing

effect. A large divergence occurs when the inclusion of factor $\mathbf{g}$ forces a substantial revision of the state-price density, either by shifting the level of the normalizing constant or by altering the risk-neutral expectations of factor payoffs. If $\mathbf{g}$ is redundant with the factors already in $\mathbb{S}$, the augmented parameters will differ negligibly from $\boldsymbol{\theta}^{\mathbb{S}}$, the normalization will barely change, and the divergence will be near zero.

The gradient of the Bregman divergence links this representation directly to pricing errors. For a candidate factor $\mathbf{g}$ outside the current span $\mathbb{S}$, the derivative of the KL divergence with respect to its loading $\theta_{\mathbf{g}}$ is proportional to the factor's alpha under $\mathcal{P}$:

$$\frac{\partial}{\partial \theta_{\mathbf{g}}} D_{KL} \propto \mathbf{E}_{\mathcal{P}} \left[m_t(\boldsymbol{\theta}^{\mathbb{S}}) g_t \right] \neq 0. \tag{8}$$

A nonzero alpha is a displacement in the tangent space of the manifold at $\mathcal{Q}^{\mathbb{S}}$, and its magnitude is measured by the KL divergence in that direction. The Bregman representation therefore provides a direct link between the statistical criterion used for factor selection and the economic magnitude of the pricing error that the omitted factor generates. Maximizing the divergence is equivalent to searching the zoo for the source of risk that carries the largest unexplained risk premium in information units.

The local geometry of $\mathbb{M}$ provides the connection to the Hansen and Jagannathan (1997) (HJ) distance that is prevalent in empirical finance. For infinitesimally close models parameterized by $\boldsymbol{\theta}$ and $\boldsymbol{\theta} + d\boldsymbol{\theta}$, the KL divergence is locally quadratic and governed by the Fisher information metric

$$D_{KL}(\mathcal{Q}(\boldsymbol{\theta}) || \mathcal{Q}(\boldsymbol{\theta} + d\boldsymbol{\theta})) \approx \frac{1}{2} d\boldsymbol{\theta}^{\top} I(\boldsymbol{\theta}) d\boldsymbol{\theta}, \tag{9}$$

where $I(\boldsymbol{\theta}) = [i_{jk}(\boldsymbol{\theta})]$ is the Fisher information matrix, given by the Hessian of the cumulant generating function

$$i_{jk}(\boldsymbol{\theta}) = \frac{\partial^2 \psi}{\partial \theta_j \partial \theta_k}(\boldsymbol{\theta}) = \mathrm{Cov}_{\mathcal{Q}(\boldsymbol{\theta})}(f_{jt}, f_{kt}).$$

This metric acts as a volatility-adjusted ruler on the manifold. A displacement in parameter space represents a meaningful change in the pricing model only when it occurs in a direction of high factor variance under the risk-neutral measure. The intrinsic nature of the metric ensures that the selection criterion is invariant to the rescaling or rotation of the factor space.

The HJ distance D_{HJ} measures the minimum Euclidean distance between a candidate SDF and the set of admissible pricing kernels in the space of square-integrable random variables. It is a second-moment metric, fundamentally linked to mean-variance efficiency and linear beta pricing, and it inherits the geometry of L^2. Deviations from the admissible set are measured symmetrically: positive and negative pricing errors receive equal weight, and the strict positivity of the SDF plays no formal role in the distance calculation. This last point is consequential. The HJ projection frequently implies an SDF that is negative in extreme states, precisely the crash states where the true SDF should be largest, because the Euclidean geometry has no mechanism to penalize violations of positivity.

The information-geometric approach subsumes the HJ framework as a special case. A Taylor expansion of D_{KL} around the null hypothesis $\theta_{\mathbf{g}} = 0$ yields

$$D_{KL}(\mathcal{Q}^{(\mathbb{S},\mathbf{g})}||\mathcal{Q}^{\mathbb{S}}) = \frac{1}{2}D^2_{HJ}(\mathcal{Q}^{(\mathbb{S},\mathbf{g})}||\mathcal{Q}^{\mathbb{S}}) + \mathcal{O}_{\mathcal{P}}\left(\left\|\boldsymbol{\theta}^{(\mathbb{S},\mathbf{g})} - \boldsymbol{\theta}^{\mathbb{S}}\right\|_2^3\right). \tag{10}$$

For small deviations from the baseline model, or equivalently in the Gaussian limit where all cumulants of order three and above vanish, maximizing the KL divergence is identical to maximizing the reduction in HJ distance. The two criteria agree precisely where the linear, quadratic world is an adequate approximation to the true return distribution. The HJ distance is therefore not superseded by the geometric framework; it is recovered from it as the leading term of a curvature expansion on $\mathbb{M}$.

The KL divergence departs from HJ through the higher-order terms, and it does so in a directionally asymmetric way that is economically meaningful. The divergence penalizes models that assign insufficient probability mass to states where the true SDF is high, which are generically the crash states where marginal utility is elevated. A candidate factor that hedges downside tail risk generates a strictly larger KL divergence than a factor with symmetric volatility and an identical unconditional Sharpe ratio. The geometric selection criterion is therefore inherently crash-sensitive, prioritizing factors that capture the wedge between the physical and risk-neutral measures in the left tail even when standard mean-variance analysis would treat them as equivalent to their symmetric counterparts. These higher-order terms encode the skewness, kurtosis, and co-kurtosis of the return distribution, the precise features that drive the premium demanded for bearing asymmetric downside exposure and fat-tailed losses in states of high marginal utility.

The geometric machinery delivers a complete ranking algorithm: project, measure divergence, add the factor that moves furthest along the manifold. It does not, however, say when to stop. Two questions must be addressed. First, *can we detect a nonzero risk price at all?* Second, *is the resulting mispricing economically meaningful?* This subsection addresses the first question; Section II.B addresses the second.

The natural approach to determining whether adding $\mathbf{g}$ significantly changes the model is to compare the nested models via their KL divergences. However, as Vuong (1989) established, nested model comparison via KL divergence cannot distinguish the models when the restricted model lies on the boundary of the augmented model's parameter space, as it does here when $\theta_{\mathbf{g}} = 0$. The appropriate test statistic is instead the Wald statistic on the estimated risk price $\hat{\theta}_{\mathbf{g}}$.

Proposition 1 (Statistical Stopping Rule): *Under the exponential-affine SDF with factor* $\mathbf{g}^{\perp}$ *as the orthogonal component of* $\mathbf{g}$ *with respect to* $\mathbb{S}$*, the Wald statistic* $W_T := \hat{\theta}^2_{\mathbf{g}}/\hat{\sigma}^2_{\hat{\theta}_{\mathbf{g}}}$ *satisfies:*

1. *Under* $H_0 : \theta_{\mathbf{g}} = 0, \quad W_T \to \chi^2_1$.
2. *Under* $H_1 : \theta_{\mathbf{g}} \neq 0, \quad W_T \to \chi^2_1(\delta_{\mathbf{g}})$*, with noncentrality*

$$\delta_{\mathbf{g}} = \frac{T(\theta^*_{\mathbf{g}})^2}{\sigma^2_{\theta_{\mathbf{g}},\infty}}$$

where $\sigma^2_{\theta_g,\infty}$ denotes the long-run asymptotic variance of $\sqrt{T}\hat{\theta}_{\mathbf{g}}$.

3. *To second order in $\theta^*_{\mathbf{g}}$, the noncentrality parameter coincides with twice the KL divergence between the augmented and restricted models scaled by the long run variance,*

$$\delta_{\mathbf{g}} = \frac{T}{\Omega_{\mathbf{gg}}} \frac{2D_{KL}(\mathcal{Q}^{(\mathbb{S},\mathbf{g})}||\mathcal{Q}^{\mathbb{S}})}{\left(\nabla_{\theta_{\mathbf{g}}}\psi(\boldsymbol{\theta}^{(\mathbb{S},\mathbf{g})})\right)^2}.$$

Proposition 1 admits a dual interpretation. Statistically, the Wald test asks whether the estimated risk price $\hat{\theta}_{\mathbf{g}}$ is large relative to its standard error. Economically, the noncentrality $\delta_{\mathbf{g}}$ links the test's power directly to the KL divergence. A factor with a large KL divergence but a poorly estimated risk price, perhaps because of heavy tails or serial dependence in its returns, can produce a small $\delta_{\mathbf{g}}$ and fail the test despite its economic relevance.

The Wald test creates a feasible set of factors

$$\mathbb{J} := \left\{\mathbf{g} : W_T(\mathbf{g}) > \chi^2_1(1-\alpha)\right\}$$

for which a true pricing signal is detected at α significance level. Within this set, factors can be ranked by their KL divergence D_{KL}. What the test does not provide is an economic criterion for when a factor's omission creates tolerable mispricing. A factor might pass the Wald test with a highly significant but economically negligible risk price. This gap, the absence of an economic threshold for acceptable mispricing, motivates the investor's problem in the following subsection.

The geometric machinery of this subsection delivers a selection algorithm and a statistical significance filter. The Pythagorean decomposition provides an exact, step-by-step criterion for ranking candidate factors. The KL divergence supplies a canonical objective directly related to model implied pricing errors. The Taylor expansion around the HJ distance clarifies precisely where the geometric approach departs from conventional methods and what it adds: crash sensitivity, positivity enforcement, and higher-order risk premia. The Wald test provides a statistical filter. What remains missing is an economic criterion for when residual mispricing is tolerable. For that, I need an investor.

B. Investor's Portfolio Choice Problem

The purpose of this subsection is twofold. First, I provide economic foundations for the geometric procedure developed above, showing that a rational investor facing model uncertainty would follow precisely the ranking criterion that the manifold geometry dictates. Second, I derive an economic bound on tolerable mispricing that, combined with the statistical stopping rule of Proposition 1, determines the equilibrium composition of the pricing kernel. The economic model is not novel; I adopt the multiplier preferences of Hansen and Sargent (2001, 2020). What is new is the connection to the manifold geometry, the application to the factor zoo, and the resulting characterization of the stopping threshold.

Consider a representative investor who lives for one period, is endowed with initial wealth W_0,

and must allocate across N assets with excess returns $\mathbf{r}^e \in \mathbb{R}^N$. The economy is driven by K candidate risk factors $\mathbf{f} \in \mathbb{R}^K$ that jointly determine the distribution of returns. The investor holds a reference probability model $\mathcal{P}$ that they believe is approximately correct, but they acknowledge that $\mathcal{P}$ may be misspecified. To formalize this concern, the investor evaluates their portfolio under a family of distorted probability measures, guarding against the worst-case alternative. Following Hansen and Sargent (2001), the investor has multiplier preferences

$$V(\mathbf{w}) := \min_{h \geq 0,\, \mathbf{E}_{\mathcal{P}}[h]=1} \{\mathbf{E}_{\mathcal{P}}\left[hu(W_1)\right] + \vartheta \mathbf{E}_{\mathcal{P}}\left[h \ln(h)\right]\} \tag{11}$$

subject to the budget constraint $W_1 = W_0(R_f + \mathbf{w}^\top \mathbf{r}^e)$, where $u : \mathbb{R} \to \mathbb{R}$ is strictly increasing and concave, $\vartheta \in \mathbb{R}_{++}$ is the *welfare sensitivity to misspecification*, and R_f is the gross risk-free rate.

The distortion h is a Radon-Nikodym derivative satisfying $h \geq 0$ and $\mathbf{E}_{\mathcal{P}}[h] = 1$, so that the reweighted measure $\mathcal{Q} = h\mathcal{P}$ is a valid probability distribution. When $h(\omega) > 1$, the distortion assigns state ω more probability mass than the reference model suggests; when $h(\omega) < 1$, the state is downweighted. The distortion represents the "adversary's" hypothesis about how the world differs from the investor's baseline.

The first term in (11) evaluates expected utility under the distorted measure $hd\mathcal{P}$. The second term penalizes deviations from the reference model by the relative entropy $D_{KL}(h\mathcal{P}||\mathcal{P}) = \mathbf{E}_{\mathcal{P}}\left[h \ln(h)\right]$, which measures the statistical cost of the distortion. The parameter ϑ governs the tradeoff between severity and plausibility. A large ϑ makes large distortions prohibitively expensive, confining the adversary to a small entropy neighborhood of the reference; the investor is nearly confident in the baseline. A small ϑ imposes little cost on large distortions, allowing the adversary to consider distortions far from the reference; the investor is highly ambiguity-averse. In the limit $\vartheta \to \infty$, the penalty dominates, forcing $h \equiv 1$, and the investor reduces to a standard expected-utility maximizer.[2]

In the context of the factor zoo $\mathbb{F}$, the investor anchors to a baseline pricing kernel that loads on a subset $\mathbb{S} \subseteq \mathbb{F}$ of established factors, with the baseline SDF given by (2) and the induced risk-neutral measure $\mathcal{Q}^{\mathbb{S}}$. The investor has confidence in $\mathcal{Q}^{\mathbb{S}}$; it embodies the factors that have survived rigorous empirical testing. However, the investor acknowledges that the factor zoo contains $K - |\mathbb{S}|$ additional candidates, each representing a direction along which the baseline might be misspecified. This is structured ambiguity: the set of plausible alternative models is parameterized by the factor zoo, and the investor's ambiguity set is the SDF manifold $\mathbb{M}$.

The investor's full problem is

[2] Multiplier preferences admit an axiomatic foundation through the variational preferences of Maccheroni et al. (2006), which generalize the max-min expected utility of Gilboa and Schmeidler (1989). The entropy penalty $\vartheta D_{KL}(\mathcal{Q}||\mathcal{P})$ is the unique penalty, up to affine transformations, that satisfies translation invariance and is generated by a convex, lower-semicontinuous function on the space of probability measures. The KL divergence therefore emerges as the unique model distance consistent with the preference axioms.

$$\max_{\mathbf{w}\in\mathbb{R}^N} V(\mathbf{w}) \equiv \max_{\mathbf{w}\in\mathbb{R}^N} \min_{h\geq 0,\, \mathbf{E}_{\mathcal{P}}[h]=1} \left\{\mathbf{E}_{\mathcal{P}}\left[hu(W_1)\right] + \vartheta \mathbf{E}_{\mathcal{P}}\left[h\ln(h)\right]\right\}. \tag{12}$$

The inner problem in (12) is a constrained minimization of a functional that is linear in h plus a strictly convex entropy penalty. The solution is the standard Gibbs variational result.

Proposition 2 (Worst-Case Distortion): *For any fixed portfolio* $\mathbf{w}$*, the worst-case density distortion that solves the inner problem in* (12) *is*

$$h^*(\mathbf{w}) = \frac{\exp\left(-\vartheta^{-1}u(W_1)\right)}{\mathbf{E}_{\mathcal{P}}\left[\exp\left(-\vartheta^{-1}u(W_1)\right)\right]}. \tag{13}$$

The worst-case measure $\mathcal{Q}^*_{\mathbf{w}} := h^*(\mathbf{w})d\mathcal{P}$ *is an exponential tilt of the reference model in the direction of low utility.*

The worst-case model is not selected by the econometrician; it emerges endogenously from the investor's preferences. The exponential tilt in (13) upweights states where utility is low and downweights states where utility is high. States in which the investor's portfolio suffers severe losses, the crash states where marginal utility is elevated, receive the most probability mass under the distortion. The severity of the tilt is governed by ϑ: investors with smaller welfare sensitivity to misspecification produce more extreme distortions, concentrating the worst-case measure more sharply on adverse outcomes.

Each omitted factor represents a direction along which the adversary can attack the baseline model. If a candidate factor $\mathbf{g} \notin \mathbb{S}$ carries a nonzero risk price $\theta^*_{\mathbf{g}} \neq 0$, the baseline model $\mathcal{Q}^{\mathbb{S}}$ misprices assets exposed to $\mathbf{g}$, generating an anomaly $\alpha^{\mathbb{S}}_{\mathbf{g}}$ in the pricing errors. The adversary exploits precisely these unhedged directions: the worst-case distortion h^* tilts probability mass toward states where the omitted factor takes extreme values, concentrating on the crash states that the baseline ignores. Including factor $\mathbf{g}$ in the SDF removes it from the adversary's toolkit by pricing the risk that was previously left as an anomaly. An investor with small ϑ faces a powerful adversary who can create severe distortions along every unhedged direction, making even modest unexplained premia costly. This investor hedges aggressively, including more factors to shrink the adversary's action set. An investor with large ϑ faces a constrained adversary who can inflict only mild distortions; unexplained premia of moderate size are tolerable, and the investor includes fewer factors.

The distortion h^* is central to the stopping rule derived below. Its entropy, $\eta := \mathbf{E}_{\mathcal{P}}\left[h^*\ln(h^*)\right]$, bounds the amount of mispricing that the investor can rationalize as misspecification rather than as evidence of a genuinely omitted factor.

The worst-case distortion $h^*(\mathbf{w})$ is an exponential function of $u(W_1)$, which in turn depends on the factor realizations through the portfolio return $(\mathbf{w}^*)^\top\mathbf{r}^e$. When utility is affine in the factors, as it is to first order for any smooth utility function in a neighborhood of the optimum, the distortion $h^*(\mathbf{w})$ is an exponential-affine function of $\mathbf{f}$. An exponential-affine function of $\mathbf{f}$ is precisely a point on the SDF manifold $\mathbb{M}$. The worst-case model therefore lives on the same manifold that the

geometric algorithm traverses. This is the structural reason why the investor's problem and the manifold projection problem produce the same answer: the adversary's optimal distortion, derived from preference axioms, lands on the very surface that the information-geometric framework uses to organize pricing models.

Substituting the worst-case distortion back into the outer problem and taking first-order conditions yields the Euler equation

$$\mathbf{E}_{\mathcal{P}}\left[h^*(\mathbf{w}^*)u'(W_1)\mathbf{r}^e\right] = \mathbf{0}_N,$$

from which the equilibrium SDF is identified as

$$m^{robust} \propto \underbrace{h^*(\mathbf{w}^*)}_{\text{misspecification}} \times \underbrace{m^{ccapm}}_{\text{consumption risk}} \tag{14}$$

where the proportionality constant ensures that m^{robust} correctly prices the risk-free asset, and $m^{ccapm} \propto u'(W_1)$ is the baseline CCAPM SDF. The equilibrium SDF is the product of two components. The marginal-utility component prices risk exactly as in the standard consumption CAPM. The ambiguity distortion h^* tilts the pricing kernel toward worst-case states, adding a premium for risks that the reference model may be ignoring. The standard SDF is nested as the special case $\vartheta \to \infty$. The model-implied pricing restriction is $\mathbf{E}_{\mathcal{P}}\left[m^{robust}\mathbf{r}^e\right] = \mathbf{0}_N$.

The decomposition of the SDF in (14) generates a corresponding decomposition of expected excess returns that separates compensation for known risk from compensation for model uncertainty.

Lemma 1 (Risk Premia Decomposition): *Under the SDF* (14)*, the expected excess return on asset* i *decomposes as*

$$\mathbf{E}_{\mathcal{P}}\left[r_i^e\right] = -\left(\frac{\mathrm{Cov}_{\mathcal{P}}(m^{ccapm}, r_i^e)}{\mathbf{E}_{\mathcal{P}}\left[m^{ccapm}\right]} + \frac{\mathrm{Cov}_{\mathcal{P}}(h^*, m^{ccapm}r_i^e)}{\mathbf{E}_{\mathcal{P}}\left[m^{ccapm}\right]}\right). \tag{15}$$

The first term is the standard risk premium under the baseline model, compensation for covariance with consumption risk. The second term is the misspecification premium, the additional compensation that the investor demands for bearing exposure to the worst-case distortion h^*. The relationship between the misspecification premium and the baseline pricing error is made explicit by the identity

$$\mathbf{E}_{\mathcal{P}}\left[m^{ccapm}r_i^e\right] = -\,\mathrm{Cov}_{\mathcal{P}}(h^*, m^{ccapm}r_i^e).$$

The left-hand side is the pricing error of the baseline, the component of the expected return that consumption risk fails to explain. The right-hand side states that this anomaly is governed by the covariance between the worst-case distortion and the SDF-weighted return, exactly as the standard risk premium is governed by the covariance between the SDF and excess returns. The distortion h^* therefore functions as a second pricing kernel, operating on the directions the baseline misses.

The sign of the misspecification premium has a transparent interpretation. When h^* and

$m^{ccapm} r_i^e$ move in opposite directions, the asset delivers low SDF-weighted returns in the states that the distortion makes more probable. The alpha is positive, and investors demand compensation for bearing exposure to a risk the baseline ignores. When they comove, the asset provides insurance against model uncertainty and commands a lower return. Since h^* is an exponential tilt toward low-utility states, this covariance is largest for assets exposed to omitted factors that are active in the left tail of the return distribution. When no factors are omitted, $h^* \equiv 1$, the covariance vanishes, and the standard Euler equation is recovered.

The misspecification premium is the economic content of the nonzero tangent displacement identified in the geometry of Section II.A. The displacement in the manifold's tangent space at $\mathcal{Q}^{\mathbb{S}}$, measured by the gradient $\partial_{\theta_{\mathbf{g}}} D_{KL}$ in (8), is the investor's misspecification premium.

The misspecification premium establishes that each omitted factor generates a welfare cost for the investor. The natural question is how to rank candidate factors by the severity of this cost. The following result shows that the investor's preference-based ranking reduces to a purely statistical quantity.

Proposition 3 (Factor Ranking): *Under the multiplier preferences* (11) *and the exponential-affine SDF specification, the worst-case welfare loss from misspecification in the direction of factor* $\mathbf{g}$*, to leading order, satisfies*

$$|\Delta V_{\mathbf{g}}| \leq \sqrt{2\vartheta \mathbf{V}_{\mathcal{Q}^{\mathbb{S}}}(u(W_1^*))} \sqrt{D_{KL}(\mathcal{Q}^{(\mathbb{S},\mathbf{g})} || \mathcal{Q}^{\mathbb{S}})}. \tag{16}$$

The ranking of candidates by worst-case welfare loss is identical to the ranking by KL divergence:

$$|\Delta V_{\mathbf{g}}| > \left|\Delta V_{\mathbf{g}'}\right| \iff D_{KL}(\mathcal{Q}^{(\mathbb{S},\mathbf{g})} || \mathcal{Q}^{\mathbb{S}}) > D_{KL}(\mathcal{Q}^{(\mathbb{S},\mathbf{g}')} || \mathcal{Q}^{\mathbb{S}}).$$

Proposition 3 is the result that connects the economic model to the statistical geometry. The investor solves a portfolio problem, evaluates the worst-case welfare loss from each direction of misspecification, and finds that the KL divergence between the augmented and restricted models is the quantity governing how much indirect utility can deteriorate under the worst-case perturbation. The KL divergence appears as a consequence of this economic analysis, not as an assumed objective.

The content of the proposition is that the preference-based ranking of factors by welfare loss is identical to the purely statistical ranking by divergence. This is not guaranteed *a priori*; it relies on the specific interaction between the exponential-tilt structure of the worst-case distortion and the e-flat geometry of the SDF manifold. The prefactor $\sqrt{2\vartheta \mathbf{V}_{\mathcal{Q}^{\mathbb{S}}}(u(W_1^*))}$ is common to all candidates. An investor with smaller ϑ suffers a smaller absolute welfare loss from each omission, but the relative ranking of candidates is invariant to ϑ. All investors with multiplier preferences, regardless of their individual sensitivity to misspecification, agree on the ordering.

The ranking in Proposition 3 tells the investor which factor to add next, but it does not tell them when to stop. Without a stopping criterion, the investor would include the entire factor zoo. The key observation is that not every anomaly reflects a genuinely omitted factor. The investor knows their pricing model is wrong, and some of the apparent mispricing can be attributed to

misspecification rather than to a truly missing risk source.

The entropy of the worst-case distortion h^* provides the criterion. The investor holds the baseline $\mathcal{Q}^{\mathbb{S}}$ and considers extending to $\mathcal{Q}^{(\mathbb{S},\mathbf{g})}$. Both are members of the exponential family on $\mathbb{M}$. The marginal divergence $D_{KL}(\mathcal{Q}^{(\mathbb{S},\mathbf{g})}||\mathcal{Q}^{\mathbb{S}})$ measures how far the augmented model is from the baseline on $\mathbb{M}$. The investor's decision reduces to the following question: *Is this difference evidence of a genuine missing factor, or could it be an artifact of the adversary's distortion?*

If $D_{KL}(\mathcal{Q}^{(\mathbb{S},\mathbf{g})}||\mathcal{Q}^{\mathbb{S}}) \leq \bar{\eta}$, where $\bar{\eta}$ is the adversary's entropy budget, then there exists a distortion h within the adversary's feasible set that could generate the same divergence. The investor cannot rule out that the apparent improvement from adding $\mathbf{g}$ was manufactured by the adversary. Including $\mathbf{g}$ would attribute to systematic risk what could be misspecification, and there is no basis for preferring $\mathcal{Q}^{(\mathbb{S},\mathbf{g})}$ over $\mathcal{Q}^{\mathbb{S}}$.

If $D_{KL}(\mathcal{Q}^{(\mathbb{S},\mathbf{g})}||\mathcal{Q}^{\mathbb{S}}) > \bar{\eta}$, the divergence exceeds what any distortion within the entropy budget can produce. The investor must expand the model to account for the missing factor.

Corollary 1 (Economic Stopping Criterion): *The investor includes factor* $\mathbf{g}$ *only if*

$$D_{KL}(\mathcal{Q}^{(\mathbb{S},\mathbf{g})}||\mathcal{Q}^{\mathbb{S}}) > \bar{\eta}. \tag{17}$$

The entropy budget $\bar{\eta}$ bounds the total divergence the adversary can produce in any direction on $\mathbb{M}$, so it is an upper bound on the divergence the adversary could produce specifically in the direction of $\mathbf{g}$. The condition is therefore conservative.

The stopping criterion admits a connection to the Hansen-Jagannathan (HJ) bound and the GRS test statistic. Bounding h^* by the HJ inequality yields $SR_{h^*} \leq \sqrt{e^{2\eta}-1}$, where $\eta = \mathbf{E}_{\mathcal{P}}\left[h^* \ln(h^*)\right]$ is the entropy of the worst-case distortion. The set of pricing errors that can be rationalized by misspecification is

$$\mathbb{A}(\eta) := \left\{ \boldsymbol{\alpha} \in \mathbb{R}^N : \boldsymbol{\alpha}^\top \Sigma^{-1} \boldsymbol{\alpha} \leq e^{2\eta} - 1 \right\},$$

where Σ is the covariance matrix of the test assets.

An investor includes factor $\mathbf{g}$ if the pricing errors created by its omission lie outside $\mathbb{A}(\eta)$, that is, $\boldsymbol{\alpha}_{\mathbf{g}} \notin \mathbb{A}(\eta)$. Geometrically, the entropy bound $\bar{\eta}$ defines a ball on the manifold centered at $\mathcal{Q}^{\mathbb{S}}$. Factors whose marginal divergence falls within this ball are indistinguishable from adversarial distortion. Factors outside the ball represent genuine manifold curvature that must be captured by expanding the pricing kernel.

The entropy budget $\bar{\eta}$ is in principle endogenous to the selected factor set, since the entropy of h^* depends on which factors are already included in the SDF. I resolve this endogeneity by pre-calibrating $\bar{\eta}$ before starting the forward selection, analogous to fixing the significance level before conducting a sequence of hypothesis tests. The pre-calibration is accomplished through the detection error probability ε as in Hansen and Sargent (2010).

The detection error probability is defined as the average of Type I and Type II error when testing $\mathcal{Q}^{\mathbb{S}}$ against $\mathcal{Q}^{(\mathbb{S},\mathbf{g})}$

$$\varepsilon := \frac{1}{2}\left(\mathbf{P}_{\mathcal{Q}^{\mathbb{S}}}\left\{\text{Reject } \mathcal{Q}^{\mathbb{S}} \text{ for } \mathcal{Q}^{(\mathbb{S},\mathbf{g})}\right\} + \mathbf{P}_{\mathcal{Q}^{(\mathbb{S},\mathbf{g})}}\left\{\text{Fail to reject } \mathcal{Q}^{\mathbb{S}}\right\}\right).$$

A high ε (close to 1/2) means the two models are statistically indistinguishable; a low ε (close to 0) means the investor is confident in their ability to discriminate between the baseline and the alternative. Under a Gaussian approximation, the closed-form mapping from ε to $\bar{\eta}$ is

$$\bar{\eta} = \frac{2\left(\Phi^{-1}(1-\varepsilon)\right)^2}{T}, \tag{18}$$

where Φ^{-1} is the inverse standard normal CDF and T is the sample size. The investor's choice of ε and the sample size T are the only inputs to the stopping condition.[3]

The relationship between ε and factor inclusion runs counterintuitively and warrants careful discussion. Low ε means the investor can clearly attribute observed pricing errors to the adversary's distortion rather than to an omitted priced factor. This implies a large entropy budget $\bar{\eta}$. The investor tolerates more apparent mispricing before concluding that a genuine factor is missing, so fewer factors pass the inclusion threshold. High ε means the investor cannot reliably distinguish adversarial distortion from genuine factor omission. This implies a small entropy budget $\bar{\eta}$, so more factors must be included to account for residual pricing errors that cannot be confidently assigned to misspecification. The mapping is counterintuitive because one might expect an ambiguity-averse investor (high ε) to include fewer factors, having already hedged against model uncertainty. The resolution is that when the adversary's power is large and indistinguishable from a missing factor, the investor is compelled to add factors precisely because they cannot rule out that the mispricing is genuine.

The economic stopping criterion of Corollary 1 and the statistical stopping rule of Proposition 1 address distinct failure modes, and neither can perform the other's role.

The KL divergence is bounded below by zero, so sampling noise will produce a nonzero $\hat{D}_{KL}$ in any finite sample. Even when the economic criterion is satisfied, the investor must verify that the signal is detectable and not merely sampling noise. This is the role of the Wald test. The statistical criterion prevents false inclusion, the admission of factors that carry no true pricing signal. The economic criterion prevents overfitting, the inclusion of factors whose genuine signal is swamped by the amount of mispricing the adversary's entropy budget can rationalize.

Consider the two failure modes separately. A factor with $\theta^*_{\mathbf{g}} = 0$ will produce a nonzero estimated divergence in finite samples through sampling variation alone. If this spurious divergence exceeds $\bar{\eta}$, the economic criterion would erroneously admit the factor. The Wald test guards against this by requiring that the estimated risk price be large relative to its standard error. Conversely, a factor with a small but highly significant risk price might pass the Wald test easily while producing a divergence below $\bar{\eta}$. Including it would add a parameter for minimal pricing improvement, the classic kitchen-sink problem. The economic criterion guards against this by requiring that the

[3]This mapping has been validated by simulation to be an excellent approximation under non-Gaussian distributions (Student-t, Gaussian mixtures) at empirically relevant sample sizes.

divergence exceed the entropy budget.

Combining the two criteria yields the investor's complete inclusion decision.

Definition 1 (Equilibrium Stopping Rule): *The robust investor includes the candidate* $\mathbf{g}$ *with the highest* $\hat{D}_{KL}$ *if and only if both conditions hold:*

$$\hat{D}_{KL}(\mathcal{Q}^{(\mathbb{S},\mathbf{g})} \| \mathcal{Q}^{\mathbb{S}}) > \bar{\eta}, \tag{19}$$

$$W_T(\mathbf{g}) > c_\alpha, \tag{20}$$

where $\bar{\eta}$ *is calibrated via* (18) *and* c_α *is the* χ^2_1 *critical value at significance level* α*.*

The four-quadrant structure of the $(\xi_{\mathbf{g}}, \delta_{\mathbf{g}})$ space, where $\xi_{\mathbf{g}} := D_{KL}(\mathcal{Q}^{(\mathbb{S},\mathbf{g})} || \mathcal{Q}^{\mathbb{S}})$ is the population KL divergence and $\delta_{\mathbf{g}}$ is the Wald noncentrality, generates a taxonomy of pricing anomalies

	Large $\delta_{\mathbf{g}}$	Small $\delta_{\mathbf{g}}$
Large $\xi_{\mathbf{g}}$	Both pass → **Include**	Econ pass, stat fail → **Exclude**
Small $\xi_{\mathbf{g}}$	Econ fail, stat pass → **Exclude**	Both fail → **Exclude**

Table I: Taxonomy of the factor zoo. $\xi_{\mathbf{g}}$ is the population KL divergence and $\delta_{\mathbf{g}}$ is the Wald noncentrality.

The upper-left quadrant contains well-behaved factors with detectable, meaningful risk prices. A market-based factor may have a large and precisely estimated risk price because market returns are relatively well-behaved statistically. Both criteria are satisfied, and the factor enters the SDF.

The upper-right quadrant contains factors whose mispricing is genuine but whose data cannot yet support inclusion. An options-based tail-risk factor may carry substantial misspecification content, reflected in a large $\xi_{\mathbf{g}}$, but the extreme kurtosis of option returns inflates long-run variance $\Omega_{\mathbf{gg}}$ relative to information $I_{\mathbf{gg}}$, simultaneously raising the economic threshold and depressing the Wald noncentrality $\delta_{\mathbf{g}}$. These factors may migrate to the upper-left as sample sizes grow or estimation techniques improve.

The lower-left quadrant contains factors with small but tightly estimated risk prices, the kitchen-sink problem. A low-volatility factor might pass the Wald test with ease because its returns are smooth and $\Omega_{\mathbf{gg}}$ is small, but the divergence itself is negligible, falling below $\bar{\eta}$. Including it would add a parameter that buys almost no pricing improvement in the SDF.

The lower-right quadrant contains noise or data mined factors.

The geometric algorithm of Section II.A and the robust investor's equilibrium selection of Section II.B describe the same phenomenon from two vantage points. The geometry provides the ranking criterion and the computational machinery. The economic model provides the stopping threshold and the welfare interpretation. The two converge: the investor's preference-derived ranking by welfare loss coincides with the geometry's ranking by KL divergence, and both stopping conditions, statistical detectability and economic materiality, emerge from each framework inde-

pendently. As $T \to \infty$, the Wald noncentrality grows without bound for every factor with a nonzero risk price, $\bar{\eta} \to 0$ through the calibration in (18), and both criteria are eventually satisfied for every priced factor. The selected set converges to $\mathbb{S}^* := \{\mathbf{g} \in \mathbb{R}^K : \theta^*_{\mathbf{g}} \neq 0\}$, the population truth. The following Methodology section operationalizes this unified framework for empirical implementation.

III. Methodology

This section details the construction of the Kullback-Leibler Stochastic Discount Factor (KL-SDF) and the estimation techniques required to implement it. The methodology proceeds in two distinct stages. First, I describe a greedy forward-selection algorithm to identify the set of risk factors that maximally expands the span of priced risk, systematically enlarging the effective dimension of the stochastic discount factor. Second, I estimate the parameters of the resulting exponential-affine SDF using a Minimum Cross-Entropy (MCE) approach that is robust to the heavy tails and non-linearities inherent in financial data. This paper is, to my knowledge, the first to estimate the full nested exponential-affine SDF under pricing restrictions.

A. Forward Selection Algorithm

A primary challenge in empirical asset pricing is the identification of the true pricing kernel from a factor zoo populated by hundreds of potential risk candidates. Since the number of possible factor combinations grows exponentially with the size of the candidate pool, an exhaustive search is computationally infeasible. The factor zoo consists of approximately 151 factors, corresponding to 2^{151} model specifications, more than the number of nanoseconds elapsed since the Big Bang ($\approx 4 \times 10^{26}$), meaning no algorithm running since the formation of the universe could have evaluated even a negligible fraction of the candidate space. Instead, I construct the SDF using a stepwise forward-selection procedure rooted in information geometry.

The algorithm builds the pricing kernel dimension by dimension. It begins with a null model and sequentially adds the single factor that yields the largest marginal increase in the manifold's curvature. Formally, let $\mathbb{S}$ denote the set of currently selected factors and $\mathcal{Q}^{\mathbb{S}}$ denote the risk-neutral probability measure induced by those factors. In each iteration, the algorithm considers every remaining candidate factor $\mathbf{g}$ from the pool $\mathbb{F}$. It computes the KL divergence between the augmented model $\mathcal{Q}^{(\mathbb{S},\mathbf{g})}$, which includes the new factor, and the current restricted model $\mathcal{Q}^{\mathbb{S}}$. The candidate that maximizes this divergence is permanently added to the basis set $\mathbb{S}$. Econometrically, the procedure echoes forward stepwise regression, but with an intrinsic curvature-based criterion rather than the R^2 or F-statistic used in classical model selection. The analogy to Akaike (1974) is useful: both approaches use KL-oriented scoring, but the selection criterion here operates directly on the statistical manifold of exponential-affine SDFs rather than on a linear likelihood.

The theoretical justification for this greedy approach rests on the dualistic geometry of the exponential family. The manifold of admissible SDFs is e-flat, while the space of distributions satisfying the moment conditions is m-flat. This dual flatness admits a generalized Pythagorean

decomposition (5). At each step of the algorithm, the factor maximizing the KL divergence is the one whose orthogonal component under $\mathcal{Q}^{\mathbb{S}}$ creates the largest independent displacement in the tangent space of the SDF manifold. The economic intuition is direct. Consider the population analogue: the omission of a priced factor from the SDF induces a non-zero alpha, a systematic pricing error, on portfolios loaded on that factor. The KL divergence translates this pricing error into information units, so the factor with the largest divergence is the one whose omission would generate the largest alpha under the current model. By targeting this factor at each step, the algorithm expands the effective dimension of the manifold rather than traversing directions already spanned by $\mathbb{S}$.

The complete procedure is presented in Algorithm 1. I now walk through its structure, explaining the rationale for each step.

Algorithm 1 Greedy D_{KL} Forward Selection Algorithm

Require: Factor zoo $\mathbb{F}$, asset excess returns $R^e \in \mathbb{R}^{T\times N}$, entropy threshold $\bar{\eta}$

$\mathbb{S} \leftarrow \emptyset$

$\hat{\boldsymbol{\theta}}^{\mathbb{S}} \leftarrow \mathbf{0}_K$

while true do

  $\mathcal{Q}^{\mathbb{S}} \leftarrow m_t\left(\boldsymbol{\theta}^{\mathbb{S}}\right)\left(\sum_{s=1}^{T} m_s\left(\boldsymbol{\theta}^{\mathbb{S}}\right)\right)^{-1}$

  $\tilde{\mathbf{f}}_t^{\mathbb{S}} \leftarrow \mathrm{ZCA}_{\mathcal{Q}^{\mathbb{S}}}(\mathbf{f}_t^{\mathbb{S}})$

  for all $\mathbf{g}_t \in \mathbb{F} \setminus \mathbb{S}$ **do**

    $\tilde{\mathbf{g}}_t \leftarrow \mathrm{ZCA}_{\mathcal{P}}(\mathbf{g}_t)$

    $\tilde{\mathbf{g}}_t^{\perp} \leftarrow \tilde{\mathbf{g}}_t - \mathrm{Proj}^{\mathcal{Q}^{\mathbb{S}}}_{\mathrm{span}(\mathbb{S})}(\tilde{\mathbf{g}}_t)$

    $X_t \leftarrow [\tilde{\mathbf{f}}_t^{\mathbb{S}}, \tilde{\mathbf{g}}_t^{\perp}]$

    $\hat{\boldsymbol{\theta}}^{\mathbb{S},\mathbf{g}} \leftarrow \textsc{MCEOptimization}(X_t, R^e)$

    $D_{KL}^{\mathbf{g}} \leftarrow \psi(\hat{\boldsymbol{\theta}}^{\mathbb{S}}) - \psi(\hat{\boldsymbol{\theta}}^{(\mathbb{S},\mathbf{g})}) + \nabla_{\boldsymbol{\theta}}\psi(\hat{\boldsymbol{\theta}}^{(\mathbb{S},\mathbf{g})})(\hat{\boldsymbol{\theta}}^{(\mathbb{S},\mathbf{g})} - \hat{\boldsymbol{\theta}}^{\mathbb{S}})$

  $\mathbb{J} \leftarrow \left\{\mathbf{g} : W_T(\mathbf{g}_t) > \chi_1^2(1-\alpha)\right\}$

  $\mathbf{g}^* \leftarrow \arg\max_{\mathbf{g}\in\mathbb{J}} D_{KL}^{\mathbf{g}}$

  if $\mathbb{J} = \emptyset$ or $D_{KL}^{\mathbf{g}^*} \leq \bar{\eta}$ **then**

    break ▷ no candidate can explain remaining pricing errors

  $\mathbb{S} \leftarrow \mathbb{S} \cup \{\mathbf{g}^*\}$

  $\hat{\boldsymbol{\theta}}^{\mathbb{S}} \leftarrow \hat{\boldsymbol{\theta}}^{(\mathbb{S},\mathbf{g})}$

return $\mathbb{S}$

The algorithm begins with an empty factor set $\mathbb{S}$ and a zero parameter vector. Under the null model, the SDF is a constant $m_t = 1$ for all t, implying that the risk-neutral measure coincides

with the physical measure and no risk is priced. Each subsequent iteration perturbs this baseline by adding the single factor whose inclusion forces the largest revision of the state-price density.

At each iteration, the algorithm computes the implied probability weights $\pi_t(\hat{\boldsymbol{\theta}}^{\mathbb{S}})$ induced by the current SDF. These weights define the risk-neutral measure $\mathcal{Q}^{\mathbb{S}}$ under which all subsequent operations within the iteration are performed. As factors are added and $\hat{\boldsymbol{\theta}}^{\mathbb{S}}$ changes, the tilted measure evolves, so that orthogonalization, whitening, and divergence evaluation all reflect the geometry of the current model rather than the geometry of the raw data.

Before evaluating a candidate, I apply zero-phase component analysis (ZCA) whitening to the factor panel under the current tilted probabilities. ZCA whitening sphericalizes the factor innovations, converting elliptical contours in the optimization surface to circular ones, and improves the condition number of the Hessian for the MCE estimation that follows.

Lemma 2 (Affine Reparameterization Invariance): *The forward selection algorithm is invariant to affine reparameterization of the exponential family.*

Since whitening is an affine transformation of the factor space via $\hat{\Sigma}_{\mathcal{Q}^{\mathbb{S}}}^{-1/2}$, Lemma 2 guarantees that the divergence improvement $D_{KL}(\mathcal{Q}^{(\mathbb{S},\mathbf{g})}||\mathcal{Q}^{\mathbb{S}})$ and the ordering of candidates are unchanged. ZCA therefore serves purely as a numerical conditioning device, with no effect on the economic content of the selection criterion.

Candidate factors in the zoo are highly correlated. In the raw data, multicollinearity inflates the variance of parameter estimates and flattens the optimization landscape, making it difficult to isolate the marginal contribution of any single factor. Line 7 projects each whitened candidate $\tilde{\mathbf{g}}_t$ onto the orthogonal complement of span($\mathbb{S}$) under $\mathcal{Q}^{\mathbb{S}}$, retaining only the residual $\mathbf{g}_t^{\perp}$.

Lemma 3 (Orthogonal Projection Invariance): *Replacing a candidate factor* $\mathbf{g}$ *with its orthogonal residual* $\mathbf{g}^{\perp} := \mathbf{g} - \text{Proj}^{\mathcal{Q}^{\mathbb{S}}}_{\text{span}(\mathbb{S})}(\mathbf{g})$ *under the current risk-neutral measure* $\mathcal{Q}^{\mathbb{S}}$ *leaves the augmented SDF manifold and the KL divergence unchanged.*

Lemma 3 states that the information gain of a factor is determined solely by its component outside the existing span. The proof, provided in Appendix A.F, follows from the fact that the component of $\mathbf{g}$ within span($\mathbb{S}$) is already absorbed by the existing parameters $\hat{\boldsymbol{\theta}}^{\mathbb{S}}$, contributing zero marginal divergence. This Gram-Schmidt Gram (1883); Schmidt (1907) orthogonalization step delivers two operational benefits. First, it stabilizes the numerical conditioning of the design matrix $X_t = [\tilde{\mathbf{f}}_t^{\mathbb{S}}, \tilde{\mathbf{g}}_t^{\perp}]$, since the columns are now uncorrelated under $\mathcal{Q}^{\mathbb{S}}$. Second, it acts as a natural filter against redundant factors: if $\mathbf{g}^{\perp} \approx \mathbf{0}$, the candidate is nearly spanned by $\mathbb{S}$, and its KL contribution will be negligible regardless of its raw correlation with returns.

For each candidate, the algorithm constructs the augmented design matrix X_t, estimates the full parameter vector $\hat{\boldsymbol{\theta}}^{(\mathbb{S},\mathbf{g})}$ via MCE optimization, detailed in Section III.B, and computes the KL divergence of the augmented model relative to the current model using Equation (7).

Not every candidate with positive KL divergence warrants inclusion. Line 12 applies a Wald test to all remaining candidates, retaining in the feasible set $\mathbb{J}$ only those factors whose risk price is statistically distinguishable from zero at the chosen significance level. Line 13 selects from this

feasible set the factor with the highest KL divergence. This two-stage screening separates the statistical question, *Is the risk price significantly different from zero?*, from the economic question *Which surviving candidate contributes the most to the span of priced risk?*.

The algorithm terminates when either the feasible set is empty, meaning no remaining candidate has a statistically significant risk price, or the maximum KL divergence within the feasible set falls below the entropy bound $\bar{\eta}$.

The output of the algorithm is an ordered set of factors $\mathbb{S}$ and a corresponding parameter vector $\hat{\boldsymbol{\theta}}^{\mathbb{S}}$ that together define a parsimonious exponential-affine pricing kernel. The ordering is economically meaningful: each factor is selected because, conditional on those already in the model, it resolves the largest remaining mispricing in the cross-section. The stopping rule ensures that the final model includes every factor whose contribution to the span of priced risk exceeds the threshold implied by the investor's tolerance for model misspecification, and excludes those that do not.

B. Estimating the Exponential-Affine SDF

Once a candidate factor set $\mathbb{S}$ is selected, the second stage of the methodology requires a robust technique to estimate the risk prices. The natural candidate is the Generalized Method of Moments (GMM), which minimizes a quadratic form of the sample pricing errors $\mathbf{h}_T(\boldsymbol{\theta})^\top W \mathbf{h}_T(\boldsymbol{\theta})$. However, the exponential structure of the pricing kernel creates a specific pathology. The exponential function $m_t(\boldsymbol{\theta})$ allows the SDF to take values arbitrarily close to zero. In finite samples, a GMM optimizer can exploit this property by driving $\boldsymbol{\theta}$ to extreme magnitudes, attenuating $m_t(\boldsymbol{\theta})$ in crash states and artificially minimizing the quadratic objective. The result is a boundary solution where the estimated SDF concentrates probability mass on a subset of observations while ignoring the rest. Economically, such a solution implies that the marginal utility of wealth is negligible during market crashes, the precise opposite of standard equilibrium theory.

I instead frame estimation as an information projection problem. The goal is to find a risk-neutral probability measure $\mathcal{Q}$ that satisfies the no-arbitrage conditions while minimizing the informational distance to the physical measure $\mathcal{P}$. This is the Minimum Cross-Entropy (MCE) estimator.

Let $\mathbf{f}_t \in \mathbb{R}^K$ denote the vector of selected risk factors and $\mathbf{r}_t^e \in \mathbb{R}^N$ denote the vector of asset excess returns, with $N > K$. The objective is to estimate the parameter vector $\boldsymbol{\theta}$ for the exponential-affine pricing kernel (2) subject to the fundamental asset pricing restrictions $\mathbf{E}_{\mathcal{P}}\left[m_t \mathbf{r}_t^e\right] = \mathbf{0}_N$.

Fix a risk-price vector $\boldsymbol{\theta}$ and define the set of all probability distributions satisfying the pricing restrictions as

$$\mathbb{P}(\boldsymbol{\theta}) := \left\{\mathcal{P} : \mathbf{E}_{\mathcal{P}}\left[m_t(\boldsymbol{\theta})\mathbf{r}_t^e\right] = \mathbf{0}_N\right\}.$$

The primal problem minimizes the Kullback-Leibler divergence to the true distribution $\mathcal{P}^\dagger$

$$\min_{\mathcal{P}\in\mathbb{P}(\boldsymbol{\theta})} D_{KL}(\mathcal{P}^\dagger||\mathcal{P}) \equiv \min_{\mathcal{P}\in\mathbb{P}(\boldsymbol{\theta})} \int \ln \frac{\mathcal{P}^\dagger}{\mathcal{P}} d\mathcal{P}. \tag{21}$$

The formulation asks for the minimal reweighting of the pricing kernel required to fairly price the assets. If the factor model is correctly specified, there exists a $\boldsymbol{\theta}$ such that the pricing errors are zero under the empirical measure, and the divergence vanishes. If the model is misspecified, the estimator finds the closest valid pricing model in the information-geometric sense. A key advantage over GMM is that the MCE approach enforces non-negativity of the SDF by construction: exponential tilting cannot produce negative probability weights, so the implied pricing kernel remains strictly positive even under misspecification, eliminating the boundary solutions that arise when GMM drives parameters to extreme magnitudes.

Solving the primal problem directly is intractable because it requires optimization over the space of all probability distributions. Kitamura and Stutzer (1997) show that this constrained optimization can be transformed into an unconstrained dual problem via Lagrange multipliers. Introducing Lagrange multipliers $\boldsymbol{\gamma} \in \mathbb{R}^N$ for the moment constraints, the dual objective function is the cumulant generating function of the pricing errors

$$\phi_T(\boldsymbol{\gamma}; \boldsymbol{\theta}) := \ln \left(\frac{1}{T} \sum_{t=1}^{T} \exp \left(\boldsymbol{\gamma}^\top m_t(\boldsymbol{\theta}) \mathbf{r}_t^e \right) \right). \tag{22}$$

The estimation proceeds as a nested optimization

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^K} \max_{\boldsymbol{\gamma} \in \mathbb{R}^N} \phi_T(\boldsymbol{\gamma}; \boldsymbol{\theta}), \tag{23}$$

which can equivalently be written as a nested minimization problem via Fenchel duality by introducing the convex conjugate of the cumulant generating function of the pricing restrictions. In practice, I solve the nested min-min formulation because the inner problem is strictly convex with a unique global solution, and the envelope formula delivers analytic outer gradients. As shown by Kitamura and Stutzer (1997), the MCE estimator is asymptotically equivalent to two-step efficient GMM, inheriting its consistency and asymptotic normality while avoiding the finite-sample pathologies described above.

The estimation procedure separates the optimization into two stages that exploit the different curvature properties of the objective. Although the MCE framework has been developed theoretically by Kitamura and Stutzer (1997) and applied in simplified settings Ghosh et al. (2016), that implementation solves only the inner problem for fixed risk prices or impose additional restrictions that reduce the parameter space. The full nested optimization over both the Lagrange multipliers $\boldsymbol{\gamma}$ and the exponential-affine risk prices $\boldsymbol{\theta}$ has not, to my knowledge, been implemented before. Because the outer problem is non-convex and the exponential structure amplifies small parameter perturbations into large changes in the state-price density, the practical details of the optimization matter for reproducibility. I therefore describe the two loops in turn.

The estimation procedure separates the optimization into two distinct stages that exploit the different curvature properties of the objective.

The inner-$\boldsymbol{\gamma}$ loop, detailed in Algorithm 2, solves for the optimal Lagrange multipliers $\boldsymbol{\gamma}^*(\boldsymbol{\theta})$

holding the factor parameters $\boldsymbol{\theta}$ fixed. This sub-problem minimizes the dual objective $\phi_T(\boldsymbol{\gamma}, \boldsymbol{\theta})$, which is strictly convex in $\boldsymbol{\gamma}$, guaranteeing a unique global solution. The resulting vector $\boldsymbol{\gamma}^*$ has a clear economic interpretation: it represents the shadow prices of the moment constraints. A large element in $\boldsymbol{\gamma}^*$ indicates that the corresponding asset is particularly difficult to price with the current factors, necessitating a significant distortion of the probability weights to satisfy its Euler equation.

Algorithm 2 Inner $\boldsymbol{\gamma}$ Loop

Require: $\{\mathbf{r}_t^e \in \mathbb{R}^N\}_{t=1}^T$, $\{\mathbf{f}_t \in \mathbb{R}^K\}_{t=1}^T$, fixed $\boldsymbol{\theta}$, tolerance ε_γ, max iterations $L_{\max}$

Precompute $m_t(\boldsymbol{\theta}) \leftarrow \exp\left(-\boldsymbol{\theta}^\top \mathbf{f}_t - \psi(\boldsymbol{\theta})\right)$ for $t = 1, \ldots, T$

where $\psi(\boldsymbol{\theta}) = \ln \frac{1}{T} \sum_{t=1}^T \exp\left(-\boldsymbol{\theta}^\top \mathbf{f}_t\right)$

Set $\mathbf{h}_t \leftarrow m_t(\boldsymbol{\theta}) \mathbf{r}_t^e$ for $t = 1, \ldots, T$

Initialize $\boldsymbol{\gamma}^{(0)} \leftarrow \mathbf{0}$; $\ell \leftarrow 0$

while $\ell \leq L_{\max}$ **do**

  $s_t^{(\ell)} \leftarrow \boldsymbol{\gamma}^{(\ell)\top} \mathbf{h}_t$ for $t = 1, \ldots, T$

  $\tilde{s}_t \leftarrow s_t^{(\ell)} - \max_t s_t^{(\ell)}$ for $t = 1, \ldots, T$ ▷ Used for numerical stability in log-sum-exp.

  $w_t^{(\ell)} \leftarrow \dfrac{\exp(\tilde{s}_t)}{\sum_{s=1}^T \exp(\tilde{s}_s)}$ for $t = 1, \ldots, T$

  $\mathbf{h}^{(\ell)} \leftarrow \sum_{t=1}^T w_t^{(\ell)} \mathbf{h}_t$

  if $\|\mathbf{h}^{(\ell)}\|_\infty < \varepsilon_\gamma$ **then** ▷ $\nabla_\gamma \phi_T(\boldsymbol{\gamma}^{(\ell)}) = \mathbf{0}$

    break

  $d^{(\ell)} \leftarrow$ L-BFGS

  $\alpha^{(\ell)} \leftarrow$ Guarded Strong Wolfe

  $\boldsymbol{\gamma}^{(\ell+1)} \leftarrow \boldsymbol{\gamma}^{(\ell)} + \alpha^{(\ell)} d^{(\ell)}$

  $\ell \leftarrow \ell + 1$

return $\boldsymbol{\gamma}^\star \leftarrow \boldsymbol{\gamma}^{(\ell)}$, $w_t^\star \leftarrow w_t(\boldsymbol{\gamma}^\star; \boldsymbol{\theta})$ for $t = 1, \ldots, T$

The outer-$\boldsymbol{\theta}$ loop, detailed in Algorithm 3, minimizes the profiled objective function $J_T(\boldsymbol{\theta}) = \phi_T(\boldsymbol{\gamma}^*(\boldsymbol{\theta}), \boldsymbol{\theta})$ with respect to the factor loadings $\boldsymbol{\theta}$. This step identifies the SDF parameters that minimize the distributional distortion found in the inner loop. While the inner problem is convex, the outer problem is generally non-convex due to the nonlinearity of the exponential SDF. I exploit the Envelope Theorem to compute analytic gradients for the outer optimization: because $\boldsymbol{\gamma}^*$ minimizes the inner objective, the gradient of the profiled objective depends only on the partial derivative with respect to $\boldsymbol{\theta}$, avoiding the computational expense and instability of numerical differentiation.

The dual architecture offers several advantages over standard GMM in the exponential-affine setting. The objective function penalizes higher moments of the pricing errors, providing inherent

Algorithm 3 Outer $\boldsymbol{\theta}$ Loop

Require: $\{\mathbf{r}_t^e \in \mathbb{R}^N\}_{t=1}^T$, $\{\mathbf{f}_t \in \mathbb{R}^K\}_{t=1}^T$, tolerances $\varepsilon_\theta, \varepsilon_\gamma, \varepsilon_{\text{rel}}$, max iterations $L_{\max}$

Initialize $\boldsymbol{\theta}^{(0)}$ (e.g., $\mathbf{0}$/multi-start)
$\ell \leftarrow 0$
$J^{(-1)} \leftarrow \infty$
while $\ell < L_{\max}$ **do**
  $\psi^{(\ell)} \leftarrow \ln T^{-1} \sum_{s=1}^T e^{-\boldsymbol{\theta}^{(\ell)\top}\mathbf{f}_s}$
  $m_t \leftarrow e^{-\boldsymbol{\theta}^{(\ell)\top}\mathbf{f}_t - \psi^{(\ell)}}$ for $t = 1, \ldots, T$
  $g_t \leftarrow m_t \mathbf{r}_t^e$ for $t = 1, \ldots, T$
  $(\boldsymbol{\gamma}^*, \{w_t^*\}) \leftarrow$ InnerGammaLoop$(\{g_t\}, \varepsilon_\gamma)$
  $\nabla^{(\ell)} \leftarrow -\sum_{t=1}^T w_t^* m_t(\boldsymbol{\theta}^{(\ell)})(\boldsymbol{\gamma}^*)^\top \mathbf{r}_t^e \left(\mathbf{f}_t - T^{-1}\sum_{\tau=1}^T m_\tau(\boldsymbol{\theta}^{(\ell)})\mathbf{f}_\tau\right)$
  $J^{(\ell)} \leftarrow \ln T^{-1} \sum_{t=1}^T e^{\boldsymbol{\gamma}^{\star\top} g_t}$
  if $\|\nabla^{(\ell)}\|_\infty \le \varepsilon_\theta$ **and** $\left|J^{(\ell)} - J^{(\ell-1)}\right| \le \varepsilon_{\text{rel}}(1 + |J^{(\ell)}|)$ **then**
    break
  $d^{(\ell)} \leftarrow$ L-BFGS
  $\alpha^{(\ell)} \leftarrow$ Guarded Strong Wolfe
  $\boldsymbol{\theta}^{(\ell+1)} \leftarrow \boldsymbol{\theta}^{(\ell)} + \alpha^{(\ell)} d^{(\ell)}$
  $\ell \leftarrow \ell + 1$
return $\widehat{\boldsymbol{\theta}} \leftarrow \boldsymbol{\theta}^{(\ell)}$

robustness against the outliers that often derail quadratic GMM objectives. The log-sum-exp structure of ϕ_T acts as a smooth approximation to the maximum function, preventing the optimizer from fixating on single extreme observations or crash states. The separation of the linear moment constraints (handled in the inner loop) from the nonlinear parameter search (handled in the outer loop) ensures that the Euler equations are satisfied to high precision at every step of the estimation, in contrast to GMM, which often treats constraints as soft targets and may violate the fundamental pricing restrictions in finite samples. The full derivation of the dual objective and its gradient is provided in Appendix B.B.

The low signal-to-noise ratio characteristic of financial returns creates three interrelated computational challenges for the exponential-affine SDF. First, the objective surface is asymptotically flat over broad regions of the parameter space, depriving gradient-based methods of a reliable descent direction. Second, the exponential map amplifies small perturbations in $\boldsymbol{\theta}$ into large changes in $m_t(\boldsymbol{\theta})$, producing degenerate probability distributions in which a handful of observations absorb nearly all the implied probability mass. Third, these two features interact to produce numerical cliffs: abrupt changes in the objective from parameter perturbations that are negligible in Euclidean norm but large in the natural geometry of the exponential family.

To address these challenges, I developed a bespoke optimization architecture whose full details are provided in Appendix C. The key techniques are as follows. ZCA whitening of factors and test assets before estimation sphericalizes the Hessian, eliminating the ill-conditioning induced by correlated factors. Guard-safe line searching monitors the effective sample size and weight concentration at every candidate step, rejecting steps that produce pathological probability distributions. Backtracking and trust-region strategies handle the non-convexity of the outer problem, preventing stagnation in shallow local minima. Multi-start initialization from both structured starting points (linear risk premia estimates $\boldsymbol{\lambda} = \Sigma_F \boldsymbol{\theta}$) and random perturbations provides robustness against convergence to local optima.

A separate constraint arises from random matrix theory. The concentration ratio N/T governs the eigenvalue distribution of the sample covariance matrix of test asset returns via the Marchenko-Pastur law. When N/T is large, the smallest sample eigenvalues are contaminated by noise, falling below the Marchenko-Pastur edge and destabilizing both the ZCA whitening operator and the inner-loop Hessian. I impose $N/T < 0.1$ throughout the estimation to ensure that the sample eigenvalues remain well above the noise floor. The practical consequence is that the method requires long time series. This constraint limits the number of test assets that can be used at any given estimation step and motivates the specific test asset specifications employed in the empirical section.

The MCE estimator translates the economic requirement of no-arbitrage into a well-posed information-geometric optimization. By working in the dual space, the estimator avoids the pathologies that afflict quadratic objectives in exponential models and recovers a pricing kernel that is positive, internally consistent, and driven by the full distribution of asset returns rather than a handful of influential observations. The nested structure is more demanding computationally than standard GMM, but the payoff is an estimator whose finite-sample behavior reflects the theoretical properties that motivate the exponential-affine specification in the first place.

IV. Empirics

This section applies the forward selection algorithm to two progressively richer empirical settings. The first is an illustrative exercise on the 49 Fama-French industry portfolios with the six standard Fama-French factors, which makes the algorithm's mechanics transparent in a well-understood cross-section. The second applies the algorithm at full scope, using 157 candidate factors drawn from the Jensen et al. (2023) characteristic universe and the six Fama-French factors, priced against 73 principal components of 4,686 characteristic-sorted decile portfolios. The central finding is that the algorithm selects a parsimonious set of thematic clusters rather than individual factors. Profit growth, profitability, quality, and value account for the majority of the cumulative KL budget; momentum, investment, low leverage, and low risk are marginal entrants to the selected model. Accruals is never selected by the Algorithm across all parameter configurations examined. The remainder of the section documents this result, traces its origins to the path dependence of the greedy procedure, and assesses its stability across the space of tuning parameters.

A. Discussion of Parameters

The forward selection algorithm is governed by three tuning parameters, each controlling a distinct aspect of factor inclusion. Before presenting the main results, I outline each parameter and its economic interpretation. These three criteria form a hierarchy, not merely a set of interchangeable robustness checks. First, the Wald test at significance level α acts as a necessary condition: a candidate factor must show statistically significant incremental pricing content before entering the feasible set at a given iteration. Second, the entropy distortion budget $\bar{\eta}$ imposes an economic significance criterion, conditional on statistical significance. Third, the minimum effective sample size (ESS) threshold serves as a diagnostic guard against degenerate importance weights, rather than as a model-quality criterion.

A loose Wald significance level admits factors whose incremental KL contribution may be driven by sampling noise, inflating the apparent entropy budget absorbed by the selected set. A tight significance level may exclude genuinely priced factors in finite samples, particularly those whose risk price is economically meaningful but imprecisely estimated. The costs of these two errors are asymmetric. A false inclusion overstates robustness by attributing noise to priced risk; a false exclusion understates the entropy budget required to span the worst-case distortion. Because the multiplier-preference framework penalizes worst-case outcomes, the welfare consequences need not offset each other. I report the main results at $\alpha = 0.005$ and examine sensitivity to the alternatives in the robustness section.

Recall from Section II that $\bar{\eta}$ represents the maximum admissible KL divergence of the worst-case measure from the data-generating measure. Instead of calibrating $\bar{\eta}$ directly, I use the detection error probability ε for calibration. When ε is high, $\bar{\eta}$ is small: the worst-case measure is nearly statistically indistinguishable from the baseline, so the investor attributes less to model misspecification and perceives more factors as priced risk. Conversely, when ε is low, $\bar{\eta}$ is large, and more pricing error is attributed to the adversary. Throughout, I set $\varepsilon = 0.10$ as the baseline[4].

Finally, consider the ESS threshold. The exponential tilt that defines the MCE measure $\mathcal{Q}^{\mathbb{S}}$ reweights each observation by its compatibility with the pricing kernel. Factors that achieve large Sharpe ratios by exploiting only a few extreme returns concentrate probability mass on those observations. This produces a distribution that is tilted far from uniform. The ESS quantifies this concentration. It measures the effective number of observations that receive non-negligible weight under $\mathcal{Q}^{\mathbb{S}}$, and is inversely related to the squared Euclidean distance between the uniform distribution and the vector of model-implied probability weights Martino, Elvira, and Louzada (2017). When the tilt is mild and the pricing content is broadly supported, ESS remains close to T. In contrast, when a factor's contribution is driven by only a few influential data points, ESS falls sharply. This signals in-sample overfitting rather than genuine risk pricing. To guard against this, I impose a threshold of $ESS/T \geq 0.30$, requiring that at least thirty percent of the sample

[4] A detection error probability of $\varepsilon = 0.10$ means the Bayesian agent misclassifies the two measures 10% of the time. As ε rises toward 0.50, the two measures become indistinguishable. Equation (18) relates the entropy budget of the adversary to the error detection probability.

retain meaningful probability weight under the tilted measure.[5]

B. Illustrative Example

I apply the forward selection algorithm to a setting designed to make the procedure's mechanics transparent before turning to richer specifications. The test assets are the 49 Fama-French Industry value-weighted portfolios, observed monthly from July 1969 to August 2025. The candidate factor set consists of the six standard Fama-French factors (EQP[6], SMB, HML, RMW, CMA) augmented with momentum (MOM). This configuration serves as a useful baseline for several reasons. The 49 industry portfolios are among the most widely studied test asset cross-sections in empirical asset pricing, and their limited dispersion in expected returns is well documented. The candidate factors have clear economic interpretations grounded in decades of research, which permits direct evaluation of the algorithm's selections against prior knowledge. The combination of a low-dimensional factor zoo and a well-understood test asset universe also allows the results to be cross-checked against the large existing literature on HJ-distance bounds and mean-variance efficiency tests applied to these same portfolios. Table II reports the full selection path, displaying the top-ranked feasible and infeasible candidates at each iteration.

Column (1) reports the estimated risk price $\hat{\theta}_{\mathbf{g}}$ for each candidate factor, with heteroskedasticity-robust standard errors in parentheses. Column (2) reports the incremental KL divergence $\Delta D_{KL}^{\mathbf{g}}$, which measures the information gain from adding factor $\mathbf{g}$ to the current SDF. The standard error is reported below. Column (3) translates this information gain into an annualized Sharpe ratio. This represents the maximum reward-to-risk ratio an investor could extract from the factor's marginal contribution to the pricing kernel. Column (4) reports the entropy bound $\bar{\eta}$, calibrated from the detection error probability ε. A factor whose incremental KL divergence exceeds this threshold contributes pricing information that an ambiguity-averse investor cannot attribute to model misspecification. Column (5) reports the Wald statistic W_T testing $\hat{\theta}_{\mathbf{g}} = 0$. Column (6) records the reason for exclusion. "Wald" indicates failure of statistical significance at conventional levels; "Econ" indicates the factor's incremental KL falls below the entropy bound. A factor must pass both criteria to enter the model.

The algorithm selects two factors: the EQP in iteration 1 and SMB in iteration 2. The 49 industry portfolios exhibit limited cross-sectional dispersion in expected returns. Therefore, only a few systematic risk sources can be reliably identified from these test assets. Both selected factors enter with highly significant risk prices, and Wald statistics exceed 10. The annualized Sharpe ratios implied by their incremental contributions (0.503 for EQP, 0.588 for SMB) fall within plausible ranges for priced risk factors estimated over this sample. For each selected factor, the incremental KL divergence exceeds the entropy bound of 0.0032. This result implies that an ambiguity-averse investor calibrated to a detection error probability of $\varepsilon = 0.15$ would recognize these factors as

[5] In unreported results, ESS/T can fall below 0.05 before the exponential tilt encounters numerical convergence failures. However, specifications in this range yield implausible implied Sharpe ratios exceeding 10, far beyond any economically defensible bound.

[6] "Equity Premium" (Mkt-RF)

genuine sources of systematic risk rather than artifacts of model misspecification.

The feasible set contains exactly one factor in each of the first two iterations. The algorithm terminates at iteration 3 with an empty feasible set. This pattern illustrates a property of the sequential procedure. As additional factors enter the SDF, the conditional risk price estimates for remaining candidates become less contaminated by omitted-variable bias. At the same time, the bar for incremental contribution rises because the unexplained pricing variation shrinks. The evolution of SMB across iterations is instructive. In iteration 1, SMB does not appear among the top-ranked infeasible candidates. This implies an incremental KL divergence below the 0.002 recorded for HML. In iteration 2, with the market factor included, SMB's incremental KL rises to 0.014. This order-of-magnitude increase shows that size-related cross-sectional variation in industry returns is partially confounded with broad market exposure. Once market risk is absorbed, the independent pricing content of SMB becomes identifiable. The changing ranks of the infeasible set reflect this as well. For example, MOM and RMW reverse order between iterations 2 and 3. The remaining candidates in iterations 2 and 3 (MOM, RMW, CMA) fail the Wald criteria.

The algorithm terminates at iteration 3 with an empty feasible set. At this point, the two highest-ranked infeasible candidates are RMW and MOM, each demonstrating how different constraints can lead to exclusion. Specifically, RMW meets the entropy bound but fails the Wald test, indicating that while an ambiguity-averse investor might view its incremental pricing content as economically meaningful, the associated risk price lacks estimation precision, precluding inclusion. In contrast, MOM fails both the entropy bound and the Wald test, justifying its exclusion on both economic and statistical grounds.

These exclusion patterns reflect the taxonomy from Table I. A factor that passes the entropy bound but fails the Wald test provides economically relevant pricing information, but it is too imprecisely estimated to act on. CMA in iteration 1 and MOM and RMW in iteration 2 show this: their incremental KL divergences exceed $\bar{\eta}$, but risk prices remain statistically indistinguishable from zero. A factor that fails both criteria, such as HML in iteration 1 and MOM in iteration 3, makes neither a statistically reliable nor an economically significant contribution to the SDF, conditional on the selected factors. The table shows no factor passing the Wald test but failing the entropy bound. In such cases, a precisely estimated but economically negligible risk price emerges, which is more likely in a larger cross-section of asset returns with weak but well-measured signals.

The panel at the bottom of Table II reports summary diagnostics for the selected two-factor SDF. The cumulative KL divergence $\sum \hat{D}_{KL}^{\mathbf{g}} = 0.0250$ quantifies the information content of the pricing kernel relative to the risk-neutral benchmark and closely matches the direct estimate $\hat{D}_{KL}(\hat{\mathcal{Q}}^{\mathbb{S}}||\mathcal{P}) = 0.0251$, indicating minimal distortion of the physical measure to price the 49 industry portfolios.

No factor triggered an ESS warning at any iteration. The final two-factor model yields an ESS of 640.3, or 95.0% of the 674-month sample. The pricing kernel fits cross-sectionally by modest, diffuse changes to state probabilities, rather than concentrating mass on a few extreme returns. Economically, market and size factors price the 49 industries by pervasive exposure differences, not

Iter.	Rank	Factor	$\hat{\theta}_{\mathbf{g}}$ (1)	$\Delta D^{\mathbf{g}}_{KL}$ (2)	SR (3)	$\bar{\eta}$ (4)	W_T (5)	Exc. (6)
1				*Feasible*				
	1	EQP	3.081***	0.011*	0.503	0.0032	12.78***	—
			(0.862)	(0.006)	—	—	—	—
				Infeasible				
	1	CMA	-4.446	0.004	0.321	0.0032	2.41	Wald
			(2.864)	(0.006)	—	—	—	—
	2	HML	-2.215	0.002	0.237	0.0032	1.87	Wald, Econ
			(1.618)	(0.003)	—	—	—	—
	⋮							
2				*Feasible*				
	1	SMB	-5.674***	0.014	0.588	0.0032	10.21***	—
			(1.776)	(0.009)	—	—	—	—
				Infeasible				
	1	MOM	2.367	0.005	0.356	0.0032	2.38	Wald
			(1.533)	(0.007)	—	—	—	—
	2	RMW	4.221	0.005	0.335	0.0032	2.47	Wald
			(2.684)	(0.006)	—	—	—	—
	⋮							
3				*Feasible*				
	1.	—	—	—	—	—	—	—
				Infeasible				
	1	RMW	3.232	0.003	0.278	0.0032	1.24	Wald
			(2.898)	(0.005)	—	—	—	—
	2	MOM	1.593	0.003	0.246	0.0032	0.97	Wald, Econ
			(1.620)	(0.005)	—	—	—	—
	⋮							

$T = 674 \quad N = 49 \quad N/T = 0.073$
$\sum \hat{D}^{\mathbf{g}}_{KL} = 0.0250$
HJ-Distance $= 0.2835$
$\hat{D}_{KL}(\hat{\mathcal{Q}}^{\mathbb{S}}||\mathcal{P}) = 0.0251$
ESS $= 640.3 \quad$ ESS$/T = 0.950$

Table II: The table reports the results from Algorithm 1 applied to the Fama-French 49 industry portfolios. Each iteration is partitioned into the feasible and infeasible sets, showing the top 2 factors in each set. The Algorithm was run with the error detection probability $\varepsilon = 0.15$, the Wald significance $\alpha = 0.005$, and minimum ESS at 90%. Standard errors in parentheses are HAC with a Bartlett kernel and 6 lags.
Significance levels: $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

rare-event mechanisms. In later specifications with broader test-asset sets and more factors, this property sometimes fails, making the ESS diagnostic crucial for guarding against overfitting due to extreme probability distortion.

C. The Full Factor Zoo

I now turn to the full candidate universe. The factor set consists of the 151 characteristic-based long-short factors constructed by Jensen et al. (2023), together with the six factors from the illustrative example. This deliberate augmentation introduces intentional redundancy, as several JKP factors overlap with or replicate the Fama-French and momentum factors already included. For example, the JKP factor `ret_12_1` and the standard momentum factor MOM represent the same underlying return predictor, and the algorithm ranks these two factors identically at every iteration of the forward selection. By including near-duplicate factors, I directly test whether the algorithm is robust to redundancy in the candidate set.

The test assets are principal components extracted from 4,686 characteristic-sorted decile portfolios spanning 623 anomaly dimensions. The anomalies are collected from Jensen et al. (2023) and Chen and Zimmermann (2022). I drop firms that have less than 5 years of characteristic data. The full sample contains 17,296 firms spanning from July 1969 to December 2024. To respect the concentration bound of $N/T < 0.1$, I retain the first 73 principal components of the test assets, which collectively explain 95% of the cross-sectional return variation.

The columns of Table III match those of Table II, with two additions. Column (6) reports the projected effective sample size (ESS) that each candidate factor would produce if added to the SDF. Factors with projected ESS below the feasibility threshold are excluded, regardless of their KL divergence or Wald statistic. Each factor name includes its thematic classification as defined by Jensen et al. (2023).

The algorithm selects 14 factors spanning 9 thematic categories. The largest themes are profit growth (3 factors), profitability (2), quality (2), and value (2). Together these four themes account for approximately 62% of the cumulative KL divergence $\sum \hat{D}^g_{KL} = 2.27$. The remaining 5 selections span debt issuance, seasonality, short-term reversal, size, and the equity premium, each contributing a single factor. No momentum, accruals, investment, low leverage, or low risk factor enters the model.

The shift from the six-factor illustrative setting to the full 157-factor and test asset universe changes not only the size of the selected model but the identity of the factors that represent each economic theme. In the illustrative example, EQP enters at iteration 1 with an incremental KL divergence of 0.011, accounting for approximately 44% of the total KL budget against the 49-industry cross-section. In the full specification, EQP enters at iteration 8, after three profit-growth factors, one debt-issuance factor, one seasonality factor, one quality factor, and one value factor have sequentially reduced the unexplained cross-sectional KL. Despite entering later, EQP's incremental contribution in the full model is an order of magnitude larger than in the illustrative setting. The PCA test assets span substantially richer cross-sectional variation than the 49 industry portfo-

Iter.	Factor	$\hat{\theta}_{\mathbf{g}}$ (1)	$\Delta D^{\mathbf{g}}_{KL}$ (2)	SR (3)	$\bar{\eta}$ (4)	W_T (5)	Proj. ESS (6)	Exc. (7)
1	dsale_dinv	49.888***	0.256**	2.477	0.005	56.21***	0.401	
	(Profit Growth)	(6.654)	(0.100)					
2	nfna_gr1a	51.586***	0.163***	1.979	0.005	47.55***	0.562	
	(Debt Issuance)	(7.481)	(0.045)					
3	seas_11_15an	-26.671***	0.132**	1.778	0.005	18.10***	0.399	
	(Seasonality)	(6.269)	(0.054)					
4	saleq_su	-22.788***	0.082**	1.401	0.005	20.17***	0.330	
	(Profit Growth)	(5.074)	(0.033)					
5	cop_at	13.595***	0.091***	1.474	0.005	16.28***	0.568	
	(Quality)	(3.369)	(0.016)					
6	sale_me	17.663***	0.258**	2.487	0.005	37.18***	0.431	
	(Value)	(2.897)	(0.103)					
7	tax_gr1a	-16.384***	0.070***	1.298	0.005	9.48***	0.322	
	(Profit Growth)	(5.321)	(0.027)					
8	EQP	6.251***	0.119***	1.691	0.005	27.53***	0.503	
	(Market)	(1.191)	(0.016)					
9	ret_1_0	-12.693***	0.145*	1.865	0.005	15.74***	0.453	
	(ST Reversal)	(3.200)	(0.078)					
10	ocf_at	23.707***	0.207***	2.231	0.005	13.51***	0.403	
	(Profitability)	(6.450)	(0.030)					
11	ni_me	10.085***	0.288***	2.628	0.005	10.37***	0.383	
	(Value)	(3.131)	(0.012)					
12	at_turnover	13.004***	0.200***	2.193	0.005	7.90***	0.315	
	(Quality)	(4.627)	(0.015)					
13	ami_126d	8.176***	0.214***	2.264	0.005	8.82***	0.386	
	(Size)	(2.753)	(0.009)					
14	o_score	-15.304***	0.044***	1.025	0.005	8.55***	0.466	
	(Profitability)	(5.235)	(0.013)					
15 (rej.)	op_atl1	18.348*	0.055***	1.152		3.28*	0.317	Wald
	(Quality)	(10.136)	(0.016)					
	⋮							

$T = 666 \quad N = 73 \quad N/T = 0.110$
$\sum \hat{D}^{\mathbf{g}}_{KL} = 2.2678$
HJ-Distance $= 0.5549$
$\hat{D}_{KL}(\hat{\mathcal{Q}}^{\mathbb{S}} || \mathcal{P}) = 0.3537$
ESS $= 310.7 \quad$ ESS$/T = 0.466$

Table III: The table reports the results from Algorithm 1 applied to the first 73 principal components of the 4,686 decile sorted portfolios and the 151 JKP + 6 FF factors. The table shows the final selected factor at each iteration. The Algorithm was run with the error detection probability $\varepsilon = 0.10$, the Wald significance $\alpha = 0.005$, and minimum ESS at 30%. Standard errors in parentheses are HAC with a Bartlett kernel and 6 lags.
Significance levels: $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

lios, and the market factor faces stronger competition from characteristic-based factors that price aggregate-return exposure through firm-level sorts. The conditional risk price 6.25 is roughly twice its illustrative value, and the Wald statistic more than doubles. The market premium is therefore robustly priced across both test-asset universes, with its conditional pricing content strengthening once the characteristic factors that compete with it on the PCA test assets are controlled for.

The size factor behaves differently. SMB enters the illustrative model at iteration 2 with 0.014 and an annualized Sharpe ratio of 0.588. In the full specification, SMB does not enter at any iteration. The size theme is represented instead by the Amihud illiquidity measure at iteration 13, which enters with an incremental KL of 0.214 and a Sharpe ratio of 2.264. The Amihud measure loads on the same small-cap premium that drives SMB but captures it through illiquidity rather than through a market-cap-sorted spread Amihud (2002). Its conditional pricing content against the PCA test assets substantially exceeds what SMB contributes at any comparable point in the selection path. The displacement reflects the richer information set of the full test-asset cross-section. Against the PCA test assets, size-related pricing information is summarized more efficiently by a liquidity-based measure than by the market-cap-sorted spread that dominates against the industry portfolios.

The contrast between these two factors points to a general feature of greedy sequential selection. The identity of the representative that captures a given economic theme is determined partly by the ordering of earlier selections, not solely by the unconditional pricing content of the factor. In the illustrative model, SMB is the best available incremental source of size information given EQP. In the full model, twelve other factors enter before the size theme is reached, and a different representative wins the conditional race. The next section formalizes this argument and shows that while the representative standing for each theme may rotate across specifications, the set of themes itself is broadly stable.

D. Path Dependence

The patterns documented in the previous section are not specific to EQP and SMB. They reflect a general property of the greedy sequential algorithm, whereby within-theme path dependence rotates the factor representative but leaves the set of selected themes broadly intact.

At each iteration, the algorithm selects the candidate with the highest incremental KL divergence from the current conditional distribution. Two factors from the same thematic cluster share a large fraction of their pricing information, so whichever factor enters first substantially reduces the conditional KL available to the other. The second factor may then fail the Wald test, not because its unconditional pricing content is weak, but because its marginal contribution conditional on the first representative falls below the significance threshold. The algorithm therefore selects one or two representatives per theme rather than a dense within-theme bundle. Path dependence manifests primarily as rotation among near-substitutes within a theme rather than as substitution across themes.

The ordering of themes in the selection path has more consequence than within-theme rotation.

Themes that enter early absorb a disproportionate share of the KL budget, reducing the conditional KL available to later themes. When two themes are partially correlated, the order in which their representatives enter can determine whether the second theme clears the entropy bound. For example, profitability and quality, which Novy-Marx (2013) and Asness, Frazzini, and Pedersen (2019) show share substantial overlap in their pricing content. Between-theme ordering is itself path-dependent, depending on which factor from which theme attains the highest unconditional KL in the first iteration.

Despite within-theme rotation and between-theme ordering effects, the set of themes represented in the selected model is broadly stable across parameter configurations. To document this formally, I pool all ESS trials from ESS/$T \in \{0.30, 0.40, \ldots, 0.90\}$ at the baseline $\alpha = 0.005$ and $\varepsilon = 0.10$. This yields seven model estimates. For each, I record which themes are represented in the selected set. Pooling across trials averages out within-theme rotation and surfaces the themes that reliably attract a representative regardless of the overfitting threshold.

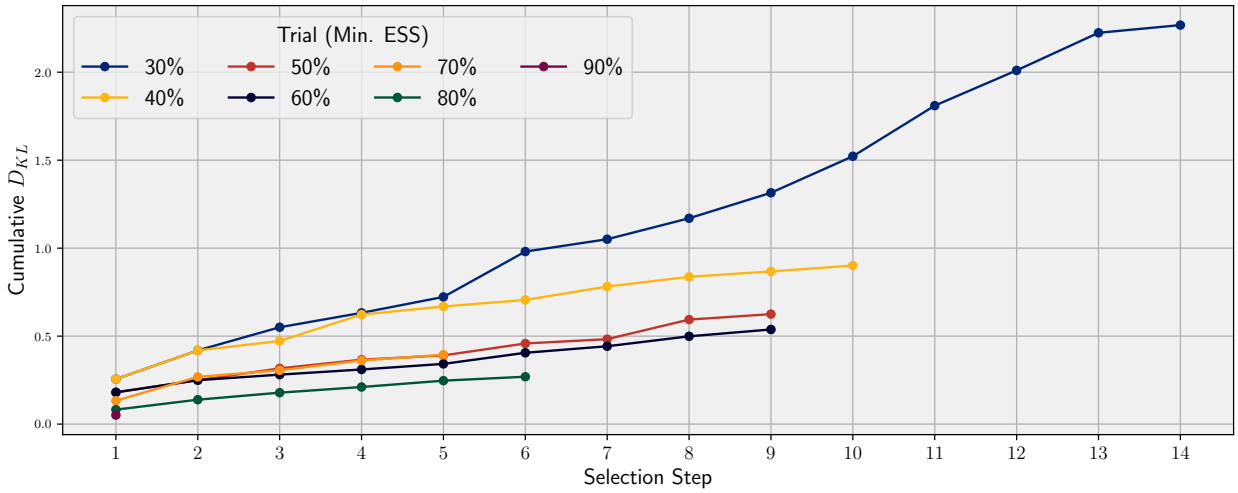


Figure 1. This plot shows the cumulative D_{KL} gain across algorithm steps for the 7 aggregated scenarios. The scenarios differ in the minimum ESS threshold as a percentage of T.

Figure 1 plots the cumulative KL divergence $\sum \hat{D}_{KL}$ as a function of iteration k for each ESS trial. Through the first several iterations the paths are tightly bunched across thresholds, accumulating comparable KL at comparable rates. Trials with ESS between 40% and 80% subsequently decelerate and terminate between iterations 6 and 10, a shape consistent with a diminishing-returns greedy algorithm in which the dominant pricing information is absorbed early, and successive admissible candidates contribute progressively less. The ESS = 30% trial does not exhibit this pattern. Its slope is approximately constant across the full selection path, with the marginal KL contribution in the later iterations comparable to or larger than in the early iterations, and the curve terminates abruptly at iteration 14 with a terminal cumulative KL more than twice that of the next-highest trial.

A possible explanation is that at the loosest thresholds the algorithm admits factors whose effective sample sizes are small enough that their realized KL contributions substantially exceed

their population values. The sampling distribution of $\hat{D}_{KL}$ at small T can produce inflated point estimates, and once the algorithm exhausts the Wald-significant candidates in the common admissible set, the ESS = 30% filter permits short-sample factors whose apparent marginal contribution reflects a mixture of true pricing content and small-sample noise. The abrupt termination of the curve, rather than a smooth asymptotic approach, is consistent with this interpretation: the algorithm continues to find nominally significant additions until the residual set of short-sample candidates is exhausted, at which point selection halts. The tight bunching of the early iterations therefore indicates that the dominant economic content of the model is invariant to the ESS threshold, while the late divergence of the ESS = 30% trial indicates that the apparent KL gains from marginal additions at loose thresholds should not be interpreted as the discovery of new pricing dimensions. As documented in Table IV, the additional factors admitted at low ESS thresholds map almost entirely to themes already represented in higher-ESS selected sets.

Theme	# Factors (1)	Sel. Rate (2)	Models (3)	Median D_{KL} (It.1) (4)	Median D_{KL} (5)
Debt Issuance	7	14.3%	85.7%	0.061	0.010
Quality	16	7.1%	85.7%	0.009	0.005
Profit Growth	10	12.9%	71.4%	0.032	0.010
Seasonality	11	5.2%	57.1%	0.019	0.009
Value	14	7.1%	57.1%	0.013	0.009
EQP	1	42.9%	42.9%	0.035	0.035
Size	5	8.6%	42.9%	0.003	0.004
Short-Term Reversal	6	4.8%	28.6%	0.043	0.009
Profitability	10	5.7%	28.6%	0.016	0.005
Momentum	8	3.6%	28.6%	0.015	0.012
Investment	22	1.3%	28.6%	0.012	0.007
Low Leverage	10	2.9%	28.6%	0.005	0.006
Low Risk	18	0.8%	14.3%	0.001	0.003
Accruals	6	0.0%	0.0%	0.004	0.003

Table IV: This table shows (1) the number of factors in each theme, (2) how often an individual factor×trial results in a selection, (3) the percentage of models the theme is represented in, (4) the median KL divergence on iteration 1, and (5) the median KL divergence across all iterations.

Table IV reports the thematic composition of the selected models aggregated across the seven pooled trials. Column (3) gives the share of trials in which at least one factor from each theme enters the selected set. Columns (4) and (5) report the median KL divergence of the theme's factors on iteration 1 and pooled across all iterations, respectively. These two columns admit distinct interpretations. The iteration-1 median captures the theme's unconditional pricing information, that is, the divergence obtainable when no other factor has been selected. The all-iterations median captures the theme's average marginal contribution, which is mechanically smaller because once one representative is admitted, remaining candidates within the theme compete for residual conditional entropy. A large gap between columns (4) and (5) indicates a theme whose pricing content

is concentrated in a single dimension, so that once the first representative enters, within-theme substitutes add little. A small gap indicates a theme whose factors carry pricing information along distinct economic dimensions and continue to clear the Wald threshold even after a representative is selected.

Debt issuance and quality are the most stable themes in the aggregated results, each entering in six of the seven trials. Their near-uniform selection despite very different within-theme structures is informative about the mechanism producing theme stability. Debt issuance enters with a high iteration-1 median KL of 0.061 that collapses to 0.010 when pooled across all iterations, a pattern consistent with a theme whose pricing content is concentrated in a first-order representative. A possible explanation follows from the external financing literature of Bradshaw, Richardson, and Sloan (2006) and Pontiff and Woodgate (2008): debt issuance aggregates information about managerial incentives, limits to arbitrage, and future operating performance into a small set of highly correlated signals, so that once one representative enters, the incremental KL from additional debt issuance factors falls sharply. The single trial in which debt issuance fails to enter is the ESS = 90% specification, under which quality is the only theme represented in the selected set.

Quality is the mirror image of debt issuance. It contains sixteen candidate factors and enters in six of seven trials, but its iteration-1 median KL of 0.009 is close to its all-iteration median of 0.005. The narrow gap suggests that quality is not dominated by a single informative representative. Several factors appear to capture distinct but economically related pricing dimensions, including profitability persistence, asset composition, and the earnings quality construction of Sloan (1996) as well as the composite quality measure of Asness et al. (2019). The single configuration in which quality fails to enter is ESS = 60%. A possible explanation is that at this specific threshold, the admissible candidate set permits profit growth or profitability factors to attain the highest unconditional KL in an early iteration, absorbing pricing content that quality would otherwise have conveyed. Because quality shares substantial information with profitability and profit growth Novy-Marx (2013), the order of admission under ESS = 60% may foreclose quality through between-theme crowding rather than reflect any absolute weakness in its marginal contribution. This interpretation is consistent with the fact that quality re-enters at both lower and higher ESS thresholds, which corresponds to different admissible sets and therefore to different between-theme orderings.

Profit growth enters in five of seven trials, with an iteration-1 median KL that again substantially exceeds the all-iteration median. The pattern resembles debt issuance. The economic interpretation is that profit growth factors carry concentrated information about earnings dynamics, along the lines of Chan, Jegadeesh, and Lakonishok (1996), with a single representative capturing most of the theme's pricing content. Seasonality contains eleven factors with a median iteration-1 KL of 0.019 that falls to 0.009 in the all-iterations pool, consistent with a theme whose factors capture related but partially distinct seasonal patterns. The pattern is consistent with the evidence in Heston and Sadka (2008) and Keloharju, Linnainmaa, and Nyberg (2016) that monthly return seasonalities constitute a robust anomaly. Value enters with a small iteration-1-to-all-iterations gap, which suggests that the multiple book-to-market and earnings-yield variants in the theme

carry overlapping but not identical pricing information.

Themes in the middle of the distribution include EQP and size, which each enter in three of seven trials. The contrast between them is informative. EQP contains a single factor, and its pricing content does not decrease conditional on other factors having been selected. This suggests that the risk-signal in aggregate market returns is distinct from any other theme. Size contains five factors but enters with a very low median KL, consistent with the documented attenuation of the size premium since Banz (1981) and with the view in Asness, Frazzini, Israel, Moskowitz, and Pedersen (2018) that size becomes informative primarily when interacted with other characteristics. Short-term reversal, profitability, momentum, investment, and low leverage all enter in two of seven trials, which indicates that under most configurations these themes are crowded out by earlier entrants with which they share pricing content.

Low risk and accruals rarely enter or not at all. Low risk enters in only one of seven trials despite containing eighteen candidate factors, and its iteration-1 median KL is 0.001. A possible explanation is that the eighteen candidates function as near-substitutes along a common beta and volatility dimension, so that whichever factor carries the theme's pricing information loses most of it conditional on the market-related representative that typically enters prior. The premia documented in Frazzini and Pedersen (2014) and Baker, Bradley, and Wurgler (2011) load heavily on the first principal component of characteristic-sorted returns, which any broad market factor captures. A complementary explanation concerns the construction of the test-asset universe. Because the test assets are the first 73 principal components of 4,600+ decile-sorted portfolios, pricing content associated with narrow volatility anomalies may load on principal components beyond the truncation threshold, in which case a portion of the theme's pricing information is excluded from the test-asset representation by construction.

Accruals fail to enter in any of the seven trials despite the prominence of the accruals anomaly following Sloan (1996). A possible explanation is that accruals-based pricing information overlaps substantially with the quality theme, which enters in six of seven trials. The accruals anomaly is itself an earnings-quality signal, and the composite quality construction of Asness et al. (2019) explicitly incorporates accruals-related components, so that once a quality representative enters the conditional pricing content of the accruals factors can fall below the Wald threshold. An alternative and not mutually exclusive interpretation is that the accruals premium has attenuated in more recent samples, consistent with the post-publication decay documented in McLean and Pontiff (2015). Neither interpretation implies that accruals carry no pricing information. The observed absence reflects the specific identifying variation available in the PCA-based test-asset representation and the presence of correlated themes in the candidate set.

Taken together, the pooled results support the claim that the set of themes represented in the selected model is broadly stable across ESS configurations. A small core (debt issuance, quality, profit growth) enters reliably, and a larger set of peripheral themes enters contingent on the admissible candidate set. Path dependence operates primarily through within-theme rotation among near-substitutes and, in the case of quality at ESS = 60%, through specific between-theme ordering

interactions. Neither channel overturns the conclusion that the KL selection algorithm identifies a stable economic content rather than an arbitrary subset of factors.

E. Stopping Criteria in Action

Now I examine what each stopping criterion is actually doing in this sample. The analysis relies on the distributions of incremental KL divergences and Wald statistics across all evaluated candidates, pooled over the seven ESS trials. The two criteria were introduced separately in Section II as distinct screens on the admissible factor set, and their empirical relationship in this sample is informative about which screen is doing the binding work.

Figure 2 plots the median incremental KL divergence by thematic category. The left panel aggregates across all evaluated candidates; the right panel restricts to the subset of candidates that pass the Wald feasibility filter. The dashed vertical line in each panel marks the entropy bound $\bar{\eta}$ calibrated to a detection error probability of $\varepsilon = 0.10$.

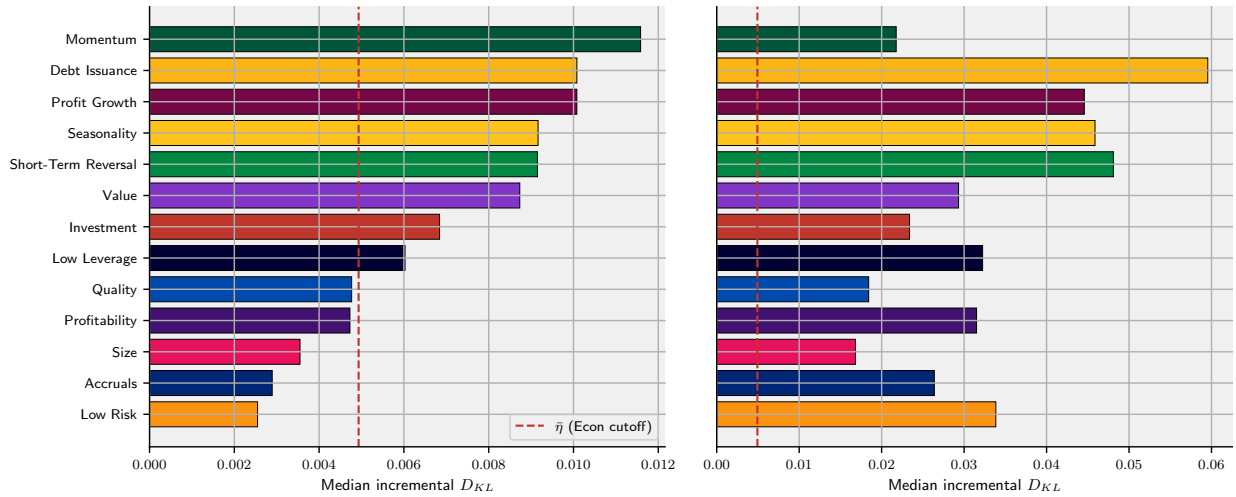


Figure 2. This plot shows the median D_{KL} value for each theme at all iterations. The left panel aggregates across all candidates; the right panel subsets to only feasible candidates.

Two features of Figure 2 warrant comment. The first is the order-of-magnitude difference in scale between the two panels. In the left panel, theme-level median KL values range from roughly 0.002 to 0.012. In the right panel they range from roughly 0.017 to 0.06, approximately five times higher. The Wald filter removes from each theme a long left tail of candidates whose incremental KL is close to zero. This scale shift quantifies the volume of uninformative candidates in the factor zoo. In every theme the median unfiltered candidate has a KL contribution that is economically small, and only 61% of candidate themes survive the significance test. The pattern is consistent with the broader characterization of the factor zoo in Harvey et al. (2015) as dominated by weakly identified signals, where appropriate multiple-testing adjustments eliminate a substantial share of documented factors.

The second feature is the reordering of themes between panels. In the left panel, momentum has

the highest median incremental KL, followed by debt issuance, profit growth, seasonality, and short-term reversal. In the right panel debt issuance moves to the top with a median of approximately 0.06. Momentum drops from first to among the lowest medians, and low risk rises from the bottom of the left panel to an intermediate position in the right panel. The ratio of the feasible-set median to the unfiltered median captures the within-theme distribution of information content rather than its absolute level. A possible explanation for these patterns draws on the theme structures established in the discussion on path dependence. Debt issuance contains a few highly informative candidates concentrated in the right tail of its KL distribution, with a larger mass of near-zero candidates; the Wald filter removes the near-zero mass and shifts the median toward the informative representatives. This is consistent with the concentrated-signal interpretation of the external-financing literature of Bradshaw et al. (2006) and Table IV. Momentum, by contrast, contains candidates whose Wald statistics are relatively uniform and typically identifiable, so conditioning on feasibility produces only a modest multiplicative shift in the median. Low risk lies at the extreme opposite of debt issuance: a high feasibility-conditional median combined with a very low unconditional median indicates that the theme contains a small subset of well-identified representatives embedded within a large mass of noise-dominated alternatives.

Across all candidate evaluations pooled over trials, no factor fails the entropy bound without also failing the Wald test. The entropy bound is therefore never the sole reason for exclusion in this sample. Two mechanisms account for the co-movement. First, the calibration at $\varepsilon = 0.10$ produces a relatively permissive bound. An investor who attributes 10% misclassification probability to the reference model attributes pricing error to omitted priced risk rather than model misspecification. Second, with $T = 666$ monthly observations, any factor with a KL contribution large enough to exceed $\bar{\eta}$ tends to produce a risk price estimate large enough to reject the null at $\alpha = 0.005$. The co-movement is a finite-sample regularity rather than a theoretical necessity. At substantially smaller ε, the bound tightens considerably, and an investor who attributes nearly all pricing error to misspecification would admit far fewer factors. In that region of the parameter space, the entropy bound would begin to bind independently of the Wald test.

Figure 3 presents the same thematic decomposition for Wald statistics. The left panel plots the median Wald statistic by theme for all evaluated candidates; the right panel restricts to feasible candidates only. The dashed vertical line marks the critical value $W_{\alpha=0.005} = 7.88$ for the one-degree-of-freedom chi-square test.

The most notable feature of the left panel is that every theme-level median falls well below the critical value. The highest medians are less than 60% of the threshold, and the lower tail of the distribution sits at or below 2. The interpretation is not that every individual factor in these themes fails significance. The median candidate in every theme produces a conditional risk price estimate that does not clear the multiple-testing-adjusted threshold. This further strengthens the conclusions from Figure 2 and Harvey et al. (2015) that the factor zoo is full of noise factors that are difficult to identify.

The right panel shows that once the Wald filter conditions on feasibility, debt issuance is dis-

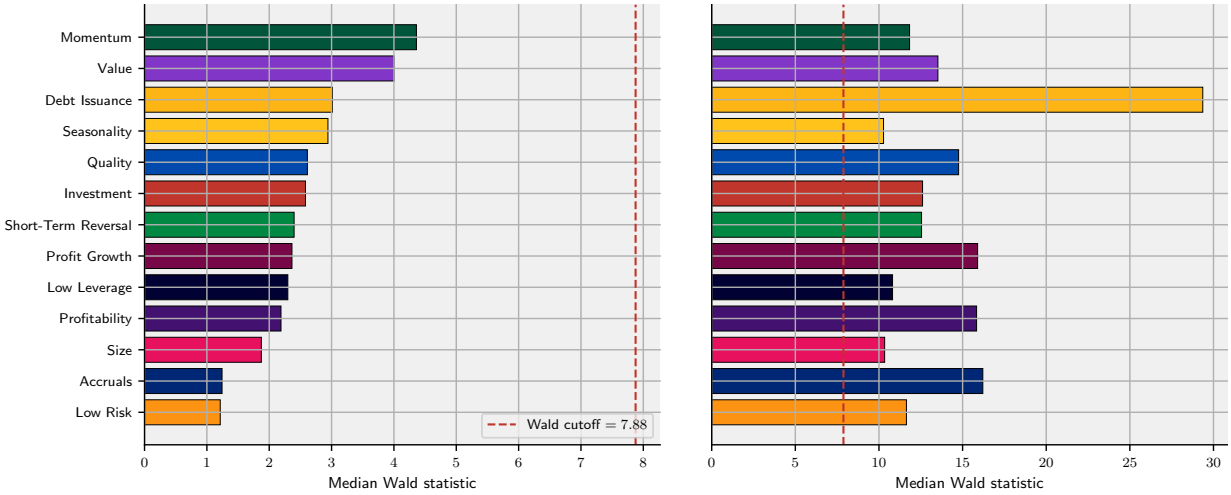


Figure 3. This plot shows the median Wald statistic for each theme at all iterations. The left panel aggregates across all candidates; the right panel subsets to only feasible candidates.

tinctly separated from the remaining themes. Its median Wald statistic of approximately 29 is roughly 80% higher than the next-highest theme. The remaining themes cluster between roughly 10 and 17, producing a relatively uniform distribution above the critical value. Again this points to debt issuance admitting a few highly identifiable factors whose pricing content concentrates along a single economic dimension, so that once the filter selects them, the feasible distribution retains unusually high t-statistics. For other themes the feasible set is more diffuse. Several candidates contribute moderate but broadly comparable pricing signal. This interpretation aligns with the iteration-1 versus all-iterations KL pattern documented in Table IV.

Comparing the two figures, the Wald statistic at $\alpha = 0.005$ separates included from excluded themes at least as cleanly as the entropy bound, and for the peripheral themes the Wald test rules out candidates well before the entropy bound is approached. This reflects a general property of short time-series data. At $T = 666$, the Wald test is powerful enough to detect any factor whose KL contribution is economically material but the factors sample KL divergence may exceed its population level, leaving the entropy bound as a redundant screen in the interior of the parameter space. The entropy bound retains its economic interpretation. However, at these sample lengths, the Wald test is the operative filter. At shorter longer lengths, or at substantially tighter ε values, the two criteria would be expected to diverge, and the entropy bound would begin to bind for candidates whose Wald statistics clear the significance threshold but whose KL contributions are too small to survive the ambiguity-aversion adjustment.

F. Wald v.s. ESS

The ESS and Wald criteria are not independent. Both guard against overfitting, but through different channels, and their interaction determines which marginal factors survive.

The Wald test screens factors whose risk price estimates are imprecise relative to sampling

variation. The ESS criterion screens factors whose inclusion would concentrate the tilted probability measure on too few observations. A tighter Wald threshold may partially substitute for a tighter ESS threshold. Factors that trigger the ESS criterion often do so because their pricing content is driven by a few extreme months, and those same factors tend to have imprecise risk prices once attention is restricted to the remainder of the sample. The converse also holds. A tighter ESS threshold may reject factors that pass the Wald test but whose risk prices are supported by concentrated observation weights. The two criteria therefore operate as partial substitutes within the interior of the parameter space, and the question of interest is the degree of that substitution.

Table V reports a line search holding $\alpha = 0.005$ and $\varepsilon = 0.10$ fixed and varying the ESS threshold from $\text{ESS}/T = 0.10$ to $\text{ESS}/T = 0.90$. The two lowest thresholds illustrate why an effective-sample-size floor is necessary in the first place. At $\text{ESS}/T = 0.10$, the selected SDF implies an annualized Sharpe ratio of 19.3, with an individual factor contributing a Sharpe ratio of 7.42 along the selection path. At $\text{ESS}/T = 0.20$, the implied Sharpe ratio is 10.25, with a within-path maximum of 5.83. Sharpe ratios of this magnitude are roughly an order of magnitude larger than the robust long-sample estimates of aggregate market Sharpe ratios in Fama and French (2002). The interpretation is that at $\text{ESS}/T < 0.30$ the algorithm is fitting a handful of extreme observations rather than identifying pervasive cross-sectional pricing. I therefore treat $\text{ESS}/T = 0.30$ as a hard lower bound throughout. The line search results for $\text{ESS}/T < 0.30$ are reported for completeness and separated from the main specifications by a horizontal rule.

ESS/T (1)	# Factors (2)	# Themes (3)	Baseline themes (4)	$\sum D_{KL}$ (5)	SR (6)	Max SR (7)	Act. ESS/T (8)
10%	10	8	78%	15.519	19.30	7.42	13.2%
20%	14	10	89%	4.381	10.25	5.83	40.2%
30%	14	9	100%	2.268	7.38	2.63	46.6%
40%	10	9	78%	0.901	4.65	2.48	43.3%
50%	9	7	67%	0.625	3.87	2.08	53.3%
60%	9	7	56%	0.538	3.59	2.08	60.7%
70%	5	4	33%	0.394	3.08	1.80	70.1%
80%	6	5	44%	0.269	2.54	1.40	80.1%
90%	1	1	11%	0.051	1.11	1.11	90.3%

Table V: This table shows a line search of different minimum ESS thresholds from 10% to 90% of T. Column (1) shows the minimum ESS threshold, (2) and (3) show the number of factors selected and themes represented in the final model, respectively, (4) shows the percentage of themes represented in the baseline model at ESS = 30%, (5) and (6) are the cumulative KL divergence and implied annualized Sharpe ratio, respectively, (7) is the maximum Sharpe ratio achieved by any included factor, and (8) is the ESS of the final model. The Wald significance and error detection probability are set to the baseline level of $\alpha = 0.005$ and $\epsilon = 0.1$, respectively.

Table V also documents the monotone decline of the implied Sharpe ratio as the ESS floor tightens. The SDF at $\text{ESS}/T = 0.30$ achieves an annualized Sharpe ratio of 7.38 with nine themes represented. At $\text{ESS}/T = 0.90$, the SDF retains a single factor with an annualized Sharpe ratio of 1.11, which aligns with consensus estimates of the efficient-frontier Sharpe ratio under broadly

tradable risk factors. The monotone relationship between the ESS threshold and the implied Sharpe ratio is consistent with the interpretation of ESS as a sample-breadth filter rather than a convergence criterion. Tightening the filter rejects factors whose pricing content is concentrated, and the rejected factors contribute both to the cumulative KL and to the tails of the SDF.

To identify which themes generate these concentrated violations, I examine the distribution of ESS outcomes at the candidate-evaluation level across the full grid of trials. Table VI reports, for each theme, the fraction of candidate evaluations that fail the ESS criterion along with the mean realized KL divergence for the failing and passing subsets.

Theme	Fail (1)	# Fail (2)	ESS | Fail (3)	ESS | Pass (4)	$\bar{D}_{KL}$ | Fail (5)	$\bar{D}_{KL}$ | Pass (6)
Seasonality	35.4%	226	39.7%	61.3%	0.498	0.018
Momentum	33.8%	162	49.3%	61.9%	0.279	0.017
Profit Growth	30.6%	162	47.2%	61.8%	0.298	0.018
Short-Term Reversal	27.3%	97	48.9%	61.8%	0.154	0.019
Value	26.8%	220	54.4%	60.5%	0.045	0.018
Low Risk	24.9%	272	56.0%	62.6%	0.046	0.011
Low Leverage	24.7%	149	53.9%	61.7%	0.037	0.012
Debt Issuance	23.9%	90	55.9%	59.8%	0.093	0.025
FF	23.6%	82	49.8%	63.5%	0.156	0.017
Investment	23.5%	312	50.0%	63.7%	0.050	0.013
Profitability	22.4%	133	58.7%	62.2%	0.044	0.014
Size	20.7%	62	51.3%	64.0%	0.039	0.013
Quality	16.2%	153	58.3%	63.3%	0.036	0.013
Accruals	14.3%	52	50.5%	63.9%	0.126	0.009

Table VI: This table reports, for each theme, (1) the fraction of candidate evaluations that fail the ESS criterion, (2) the number of failing evaluations, (3) the mean realized ESS/T conditional on failure, (4) the mean realized ESS/T conditional on passing, (5) the mean realized D_{KL} conditional on ESS failure, and (6) the mean realized D_{KL} conditional on ESS passing. Evaluations are pooled across the seven trials and all iterations.

Three themes exhibit the highest rates of ESS violation. Seasonality leads the distribution, failing in 35.4% of candidate evaluations, followed by momentum at 33.8% and profit growth at 30.6%. The remaining themes cluster between 14% and 28%. The pattern is not limited to violation frequency. Conditional on failure, seasonality candidates produce a mean $\bar{D}_{KL}$ of 0.498, roughly double the corresponding values for momentum and profit growth. The latter two are indistinguishable from each other and substantially larger than the failure-conditional means of the remaining themes. Conditional on passing, all three themes produce $\bar{D}_{KL}$ values that are indistinguishable from the cross-theme average. The failing-to-passing ratio is between 16 and 28 for these themes, an order of magnitude larger than the ratios observed in the remaining themes. This two-regime behavior is the empirical signature of overfitting on concentrated observations. When a seasonality, momentum, or profit growth factor clears the ESS filter, it behaves as an ordinary well-identified factor. When it fails, it does so by producing an apparent pricing contribution that traces to a

small subset of months.

Economic intuition supports this division. Seasonality factors, as constructed in Heston and Sadka (2008) and Keloharju et al. (2016), exploit same-calendar-month return patterns whose informational content in any given month is supported only by the subset of assets in phase with the signal. When the full history of same-calendar-month observations contains a few extreme periods, the tilted measure concentrates on those periods and the effective sample size collapses. The mechanical narrowness of this conditioning set accounts for seasonality's failing-conditional $\bar{D}_{KL}$ being roughly double that of the other two themes. Momentum factors exhibit an analogous concentration through momentum crashes, documented in Daniel and Moskowitz (2016) and Barroso and Santa-Clara (2015), during which cumulative returns drop sharply within one or two months following market rebounds. Profit growth operates through the same mechanism in a different guise. Novy-Marx (2015) argues that price momentum is largely a manifestation of earnings momentum, which rationalizes the near-identical ESS violation patterns of the two themes in Table VI. All three themes load on signals that refresh at discrete intervals and produce outsized spreads in a small subset of observations, precisely the configuration that causes the tilted measure to collapse onto a narrow support.

The policy implication for the selection algorithm is that seasonality, momentum, and profit growth factors should not be excluded on principle. All three themes enter the baseline model at $\text{ESS}/T = 0.30$ through representatives that pass the ESS filter, and when they pass they behave as the rest of the factor zoo. The ESS filter is required precisely because a substantial fraction of the candidates in these themes would otherwise enter on the strength of concentrated-observation fits. Table VI therefore supplies a quantitative rationale for the qualitative argument that the ESS criterion is not redundant with the Wald test. The two criteria target partially overlapping but distinguishable failure modes, and the themes most affected by ESS screening are precisely those whose economic construction makes concentrated-observation fits most likely.

The ESS line search above holds α fixed. To assess whether the two parameters are substitutable overfitting guards across a wider range of configurations, a full $\alpha \times \text{ESS}$ grid search is forthcoming. The grid varies $\alpha \in \{0.001, 0.005, 0.01, 0.025, 0.05\}$ crossed with the eight ESS levels above, producing 48 specifications. The question of interest is whether iso-theme-count contours in the (α, ESS) plane are downward-sloping or approximately flat in one dimension. Results will be reported in the next draft.

G. Higher-Order-Moments

The KL divergence between the MCE-tilted measure and the physical measure decomposes into a Gaussian component, which captures the mean-variance pricing content a standard SDF would price, and a residual higher-order component, which captures skewness, kurtosis, and other non-Gaussian structure in the factor's pricing information.

For an exponential-family tilt, the decomposition takes the form

$$D_{KL}(\mathcal{Q}^{\mathbb{S}}||\mathcal{P}) = D_{KL}^{\text{Gauss}} + D_{KL}^{\text{HOM}}, \tag{24}$$

where $D_{KL}^{\text{Gauss}} = \frac{1}{2}\frac{\hat{\theta}^2}{\sigma_f^2}$ is the KL divergence that would obtain if the factor were normally distributed with the same mean and variance, and $D_{KL}^{\text{HOM}} \geq 0$ is the residual attributable to the higher-order cumulants of the factor's return distribution. The Gaussian component is fully summarized by the factor's Sharpe ratio and coincides with the pricing content available to a mean-variance investor; the higher-order component reflects the additional information an ambiguity-averse investor extracts from asymmetry and tail events. The non-negativity of the residual follows from the convexity of the cumulant generating function.

As in the ESS analysis, the decomposition is cleanest at iteration 1. Once the algorithm has selected multiple factors, the tilt $\mathcal{Q}^{\mathbb{S}}$ is a joint function of all k included factors, and the marginal higher-order contribution of the k-th factor is confounded with the higher-order structure of the preceding tilts. At iteration 1 the mapping is one-to-one: the decomposition reflects the higher-order structure of a single factor's pricing information, unconditional on any prior selections. I report theme-level averages of the two components, averaging across all candidate factors within a theme at iteration 1 across the ESS trials.

Figure 4 plots the average Gaussian component against the average total KL divergence, with one point per thematic category. The 45° line is the locus of equal Gaussian and total KL divergence. Points above the line have a higher-order component that exceeds the Gaussian component, and points on the line have approximately balanced components. A point lying far above the line indicates that the factor's cross-sectional pricing power derives substantially from non-Gaussian features rather than from mean-variance compensation.

The most striking feature of the figure is that higher-order dominance is the modal pattern. The majority of themes sit above the 45° line, and no theme sits materially below it once sampling noise is taken into account. The cross-section of factor pricing information is therefore not well summarized by Sharpe ratios alone, consistent with the cross-sectional evidence on non-Gaussian features in characteristic-sorted portfolios Harvey and Siddique (2000); Dittmar (2002) and with the structural claim in Section II that the robust investor's concern about model misspecification maps into sensitivity to higher cumulants of the factor return distribution.

Among the selected themes, profitability, profit growth, value, and the equity premium all sit clearly above the line, so their inclusion in the SDF is driven in part by pricing information that a mean-variance investor would not register. Quality, by contrast, sits very close to the 45° line, indicating that its pricing content is nearly Gaussian. Its selection reflects mean-variance compensation that a standard SDF would also capture, consistent with the characterization of the quality premium as compensation for persistent firm fundamentals rather than for tail risk Asness et al. (2019); Novy-Marx (2013).

Profit growth is the most extreme higher-order case. It carries the largest total KL divergence in the sample, paired with only a middling Gaussian component, so its pricing content is dominated by non-Gaussian structure. The economic interpretation is the earnings-news channel: expectations

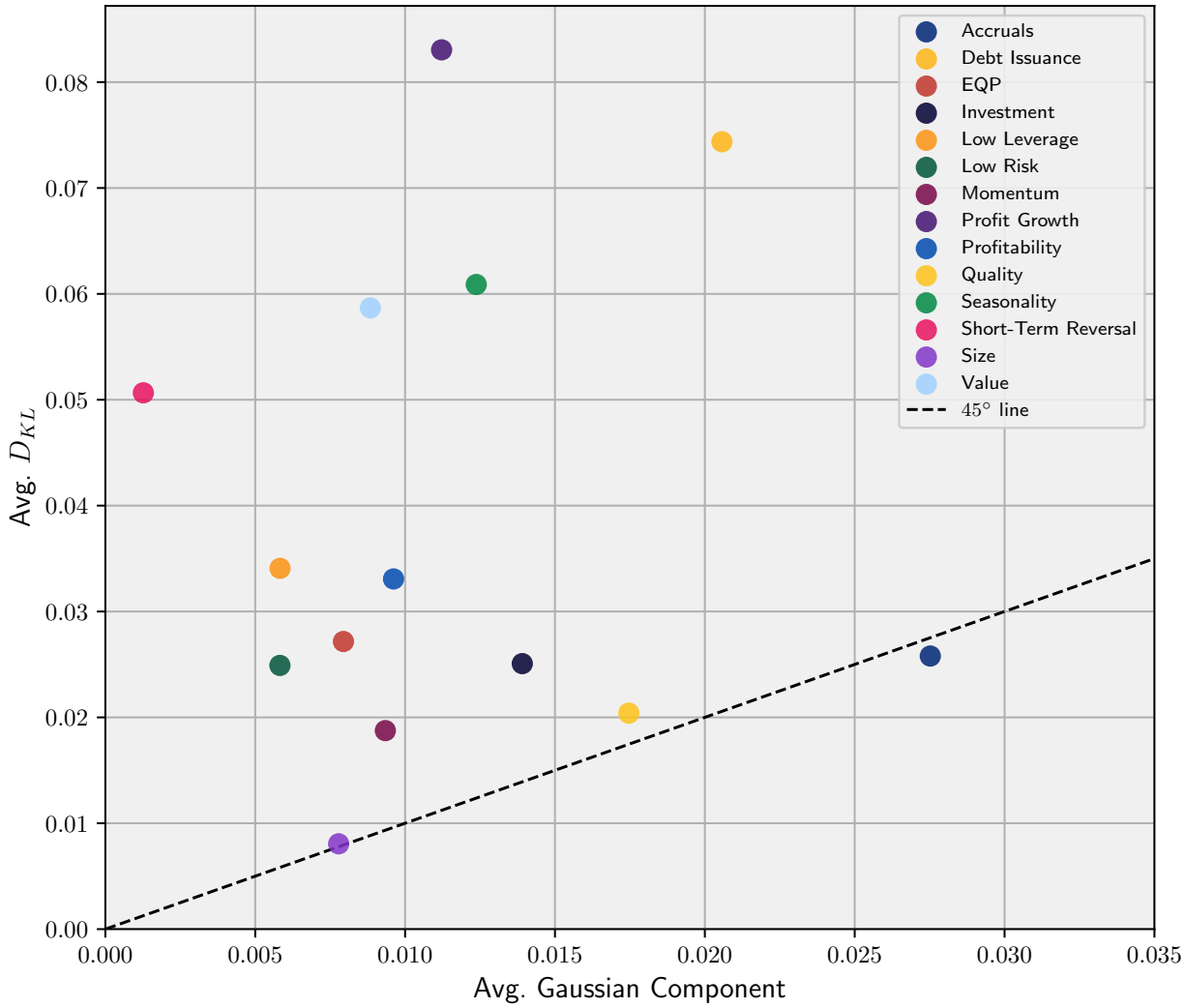


Figure 4. This plot shows the average Gaussian (x-axis) and KL Divergence (y-axis) for each theme at iteration one of the algorithm. Evaluations are pooled across the seven ESS trials restricted to the subset of feasible candidates.

errors and post-announcement drift concentrate returns in episodic, skewed events tied to earnings surprises, and the cross-section of profit-growth-sorted portfolios inherits this non-Gaussian structure Chan et al. (1996); Bouchaud, Krüger, Landier, and Thesmar (2019). Debt issuance sits at the opposite end of the selected themes. Its Gaussian and higher-order components are both large, placing it high on both axes and yielding the second-largest total KL divergence. Issuance activity generates a sizeable mean-variance premium and a sizeable tail component simultaneously, reflecting both a persistent cross-sectional spread in expected returns and the episodic, conditionally skewed corporate events that drive it Pontiff and Woodgate (2008).

Profitability and value occupy intermediate positions above the line. Profitability's higher-order content is consistent with the flight-to-safety behaviour of profitable firms in bad states, and value's is consistent with the episodic negative skewness of value returns documented in the value-crash literature Daniel and Moskowitz (2016). The equity premium sits noticeably above the line despite a

middling Sharpe ratio. The Gaussian component is modest because the Sharpe ratio is modest, but the higher-order component is non-trivial, consistent with the large literature documenting fat tails and rare-disaster exposure in aggregate market returns Barro (2006); Gabaix (2012). The robust investor extracts pricing information from this tail structure that a Sharpe-ratio-based selection criterion would miss.

Accruals and size sit on or very close to the 45° line. In population the residual is non-negative, so the natural interpretation of these points is that the higher-order component is effectively zero and the pricing information these themes do carry is primarily Gaussian.[7]

The higher-order structure documented here connects directly to the ESS line search results. Factors whose KL divergence is higher-order dominated achieve their information gain through non-Gaussian concentration of the probability tilt, which is precisely the concentration the ESS criterion penalizes. Themes lying far above the 45° line in Figure 4 correspond to the contingent themes in Table V that drop out first as the ESS threshold is tightened.

H. Summary

The greedy KL-maximization algorithm selects a parsimonious stochastic discount factor spanning roughly nine thematic categories drawn from the candidate universe. Three themes constitute a stable core. Debt issuance and quality each enter in six of seven ESS trials, and profit growth enters in five. Seasonality, value, the equity premium, and size form a middle tier that enters in three to four trials. Short-term reversal, momentum, profitability, investment, low leverage, and low risk enter in one or two trials, and accruals is absent from every specification despite containing six candidate factors. The set of themes represented is stable across the interior of the parameter space, and the peripheral themes are excluded on statistical or economic rather than on overfitting grounds.

Theme-level stability coexists with substantial factor-level variation. Of the thirty-five factors selected in at least one trial, only one—`cop_at`, the quality theme's cash-profits-to-book representative—is selected in more than four trials. The remaining themes distribute their selections across multiple representatives. Profit growth enters through six distinct factors across seven trials, seasonality through four, and value through four. This within-theme rotation reflects the greedy algorithm's treatment of near-substitute candidates. Between-theme ordering is more consequential and more stable. The dominant themes enter in approximately the same sequence across all trials, which indicates that the algorithm recovers a consistent ranking of the cross-sectional pricing structure even when the within-theme representative rotates.

The entropy bound does not bind singularly at the baseline calibration $\varepsilon = 0.10$, and its role in the final specification is interpretive rather than operative. The Wald test and the ESS criterion jointly govern admissibility, and they operate as partial substitutes within the interior

[7]Equation (24) is an exact population identity, but the empirical decomposition is a plug-in estimator with two sources of sampling error. Points whose distance below the 45° line in Figure 4 is small relative to the cross-sectional spread in KL divergences should therefore be read as approximately Gaussian with sampling noise in the residual.

of the parameter space. The ESS floor at $\mathrm{ESS}/T = 0.30$ is imposed to rule out specifications whose implied Sharpe ratios are inconsistent with economically defensible risk pricing. Below this threshold the algorithm fits concentrated-observation anomalies rather than pervasive cross-sectional variation, and the tilted probability measure places disproportionate weight on a small subset of months. The themes most exposed to this failure mode are seasonality, momentum, and profit growth, whose economic construction makes concentrated-observation fits mechanically likely.

Table VII provides the complete factor-level selection roster. For each factor selected in at least one ESS trial, the table reports its mnemonic, full name, thematic classification, the number of trials in which it was selected, the corresponding selection frequency, the mean iteration at which the factor entered the selection path, the mean incremental KL divergence at selection, the implied annualized Sharpe ratio of the selected factor, and the primary reference documenting its empirical construction.

V. Conclusion

In this paper, I develop an information-geometric framework for selecting priced risk factors from a high-dimensional candidate set. The stochastic discount factor is cast as a point on a statistical manifold of exponential-affine pricing kernels, and the manifold's dual flatness admits an exact Pythagorean decomposition of the Kullback-Leibler divergence between nested models. This decomposition supplies the ranking criterion for a greedy forward-selection algorithm, and the Hansen-Jagannathan distance emerges as the leading term of a curvature expansion of the KL divergence, coinciding with the geometric criterion locally and departing from it in directions governed by the higher cumulants of the factor return distribution. I provide an economic foundation for the procedure through a representative investor with multiplier preferences in the sense of Hansen and Sargent (2001, 2020). The investor's worst-case distortion is an exponential tilt that lands on the same SDF manifold traversed by the geometric algorithm, and the investor's ranking of candidate factors by worst-case welfare loss coincides with the purely statistical ranking by KL divergence. The entropy of the worst-case distortion provides an economic stopping threshold that, combined with a Wald test on the estimated risk price, produces a complete inclusion rule with separate statistical and economic failure modes.

On the estimation side, I implement the full nested Minimum Cross-Entropy optimization over both the Lagrange multipliers and the exponential-affine risk prices under the pricing restrictions. To my knowledge, this is the first implementation of the Kitamura and Stutzer (1997) estimator that solves the complete problem rather than a restricted subproblem. The dual architecture enforces SDF positivity by construction, penalizes concentration of probability mass on extreme observations, and delivers an estimator whose finite-sample behavior reflects the theoretical properties that motivate the exponential-affine specification.

Applying the methodology to 73 principal components of 4,686 characteristic-sorted portfolios

Factor	Full name	Trials Sel. (1)	Freq. (2)	Iter. (3)	$\overline{\Delta D}_{KL}$ (4)	$\overline{\text{SR}}$ (5)	Citation
			Debt Issuance				
nfna_gr1a	Δ Net Financial Assets	4	57%	1.5	0.172	2.03	Richardson, Sloan, Soliman, and Tuna (2005)
fnl_gr1a	Δ Financial Liabilities	2	29%	3.5	0.086	1.37	Richardson et al. (2005)
capex_abn	Abnormal Corporate Investment	1	14%	3.0	0.040	0.98	Titman, Wei, and Xie (2003)
			Quality				
cop_at	Cash Operating Profits-to-Book	6	86%	3.7	0.068	1.23	Ball, Gerakos, Linnainmaa, and Nikolaev (2016)
at_turnover	Capital Turnover	1	14%	12.0	0.200	2.19	Haugen and Baker (1996)
sale_bev	Asset Turnover	1	14%	5.0	0.033	0.88	Soliman (2008)
			Profit Growth				
dsale_drec	Δ Sales - Δ Receivables	3	43%	1.7	0.073	1.33	Abarbanell and Bushee (1997)
dsale_dinv	Δ Sales - Δ Iventory	2	29%	1.0	0.256	2.48	Abarbanell and Bushee (1997)
ocf_at_chg1	Δ Operating Chasflows-to-Assets	1	14%	3.0	0.067	1.27	Bouchaud et al. (2019)
sale_emp_gr1	Labor Force Efficiency	1	14%	4.0	0.050	1.09	Abarbanell and Bushee (1997)
saleq_su	Standardized Revenue Surprise	1	14%	4.0	0.082	1.40	Jegadeesh and Livnat (2006)
tax_gr1a	Tax Expense Surprise	1	14%	7.0	0.070	1.30	Thomas and Zhang (2011)
			Seasonality				
dbnetis_at	Net Debt Issuance	1	14%	1.0	0.133	1.78	Bradshaw et al. (2006)
seas_11_15an	Years 11–15 Lagged Ret.	1	14%	3.0	0.132	1.78	Heston and Sadka (2008)
seas_16_20an	Years 16–20 Lagged Ret.	1	14%	5.0	0.046	1.05	Heston and Sadka (2008)
seas_2_5an	Years 2–5 Lagged Ret.	1	14%	6.0	0.068	1.28	Heston and Sadka (2008)
			Value				
sale_me	Sales-to-Market	3	43%	6.0	0.117	1.53	Barbee, Mukherji, and Raines (1996)
ival_me	Intrinsic Value-to-Market	2	29%	8.5	0.044	1.01	Frankel and Lee (1998)
fcf_me	Free Cashflow-to-Market	1	14%	5.0	0.024	0.76	Lakonishok, Shleifer, and Vishny (1994)
ni_me	Earnings-to-Price	1	14%	11.0	0.288	2.63	Basu (1983)
			Size				
ami_126d	Amihud Measure	1	14%	13.0	0.214	2.26	Amihud (2002)
dolvol_126d	Dollar Trading Volume	1	14%	4.0	0.032	0.88	Brennan, Chordia, and Subrahmanyam (1998)
market_equity	Market Equity	1	14%	9.0	0.031	0.86	Banz (1981)
			Equity Premium				
EQP	Equity Premium	3	43%	6.0	0.062	1.16	Sharpe (1964)
			Investment				
ncoa_gr1a	Δ Noncurrent Operating Assets	1	14%	6.0	0.022	0.73	Richardson et al. (2005)
nncoa_gr1a	Δ Net Noncurrent Operating Assets	1	14%	6.0	0.063	1.23	Richardson et al. (2005)
			Low Leverage				
aliq_mat	Liquidity of Market Assets	2	29%	6.0	0.084	1.40	Ortiz-Molina and Phillips (2014)
			Profitability				
ocf_at	Operating Cashflow-to-Assets	2	29%	7.5	0.120	1.55	Bouchaud et al. (2019)
dolvol_var_126d	Variation in Dollar Trading Volume	1	14%	7.0	0.037	0.94	Chordia, Subrahmanyam, and Anshuman (2001)
o_score	Ohlson O-Score	1	14%	14.0	0.044	1.03	Dichev (1998)
			Momentum				
prc_highprc_252d	Current Price-to-Last Year High Price	1	14%	10.0	0.034	0.90	George and Hwang (2004)
resff3_12_1	6 Month Residual Momentum of FF3	1	14%	3.0	0.031	0.87	Blitz, Huij, and Martens (2011)
			Short-Term Rev.				
ret_1_0	Short-Term Reversal	1	14%	9.0	0.145	1.87	Jegadeesh (1990)
rskew_21d	Total Skewness	1	14%	8.0	0.055	1.15	Bali, Engle, and Murray (2017)
			Low Risk				
beta_dimson_21d	Dimson Beta	1	14%	7.0	0.076	1.35	Dimson (1979)

Table VII: This table reports, for each factor selected in at least one ESS trial, (1) the number of trials in which the factor enters the model, (2) the selection frequency as a percentage of all 7 trials, (3) the average iteration of entry (averaged only over trials in which the factor is selected), (4) the average incremental KL divergence ΔD_{KL} at the iteration of entry, and (5) the average annualized implied Sharpe ratio at entry. Trials span ESS/$T \in \{0.30, 0.40, \ldots, 0.90\}$ with $\alpha = 0.005$ and $\varepsilon = 0.1$. Factors with zero selection frequency across all trials are omitted.

against a candidate factor zoo of 157 factors yields several findings. First, the algorithm selects a parsimonious SDF spanning approximately nine thematic categories. Debt issuance, quality, and profit growth form a stable core that enters in six or more of seven parameter configurations; seasonality, value, the equity premium, and size form a middle tier; and short-term reversal, momentum, profitability, investment, low leverage, and low risk enter contingent on the admissible set. Second, accruals is absent from every specification despite containing six candidate factors, a pattern attributable to overlap with the quality theme. Third, path dependence operates primarily within themes through rotation among near-substitute representatives rather than across themes, indicating that the algorithm recovers a stable economic content rather than an arbitrary subset of factors. Fourth, the selected factors derive a substantial share of their pricing content from non-Gaussian features of the return distribution, consistent with the robust investor's concern about model misspecification mapping into sensitivity to skewness and tail exposure.

Taken together, these results suggest that the factor zoo can be organized by a principled criterion that integrates statistical and economic content rather than by a significance threshold alone. The information-geometric framework supplies the ranking, the robust-control foundation supplies the stopping rule, and the two converge on the same equilibrium selection of priced risks.

Appendix A. Proofs

This appendix provides formal proofs for all propositions, corollaries, and formal results. We adopt the notation from the main text. Let $(\Omega, \mathfrak{F}, \mathcal{P})$ be the probability space of the physical economy, and let $\mathbb{M}$ be the manifold of admissible SDFs parameterized by $\boldsymbol{\theta} \in \mathbb{R}^K$ using the normalized exponential-affine SDF.

Appendix A.A. Proposition 1

Under the exponential-affine SDF with factor $\mathbf{g}^\perp$ as the orthogonal component of $\mathbf{g}$ with respect to $\mathbb{S}$, the Wald statistic $W_T := \hat{\theta}^2_{\mathbf{g}}/\hat{\sigma}^2_{\theta_{\mathbf{g}}}$ satisfies:

1. Under $H_0 : \theta_{\mathbf{g}} = 0, \quad W_T \to \chi^2_1$.
2. Under $H_1 : \theta_{\mathbf{g}} \neq 0, \quad W_T \to \chi^2_1(\delta_{\mathbf{g}})$, with noncentrality

$$\delta_{\mathbf{g}} = \frac{T(\theta^*_{\mathbf{g}})^2}{\sigma^2_{\theta_{\mathbf{g}},\infty}} \tag{A.A.1}$$

   where $\sigma^2_{\theta_g,\infty}$ denotes the long-run asymptotic variance of $\sqrt{T}\hat{\theta}_{\mathbf{g}}$.
3. To second order in $\theta^*_{\mathbf{g}}$, the noncentrality parameter coincides with twice the KL divergence between the augmented and restricted models scaled by the long run variance,

$$\delta_{\mathbf{g}} = \frac{T}{\Omega_{\mathbf{gg}}} \frac{2D_{KL}(\mathcal{Q}^{(\mathbb{S},\mathbf{g})} || \mathcal{Q}^{\mathbb{S}})}{\left(\nabla_{\theta_{\mathbf{g}}} \psi(\boldsymbol{\theta}^{(\mathbb{S},\mathbf{g})})\right)^2}. \tag{A.A.2}$$

Proof. Throughout the main text, $\hat{\sigma}^2_{\theta_{\mathbf{g}}}$ denotes the estimated variance of the estimator $\hat{\theta}_{\mathbf{g}}$ itself, which is of order T^{-1}. Writing $\sigma^2_{\theta_{\mathbf{g}},\infty}$ for the long-run asymptotic variance of $\sqrt{T}\hat{\theta}_{\mathbf{g}}$, the two are related by $\hat{\sigma}^2_{\theta_{\mathbf{g}}} = \hat{\sigma}^2_{\theta_{\mathbf{g}},\infty}/T$. The factor of T in the noncentrality parameter is therefore absorbed into the standard-error normalization appearing in the denominator of W_T, and no explicit T scaling is required when computing the statistic in practice.

Under H_0, the MCE estimator satisfies the asymptotic normality established by Kitamura and Stutzer (1997)

$$\sqrt{T}\hat{\theta}_{\mathbf{g}} \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sigma^2_{\theta_{\mathbf{g}},\infty}\right), \qquad \sigma^2_{\theta_{\mathbf{g}},\infty} = \left(D^\top_{\mathbf{gg}} \Omega^{-1}_{\mathbf{gg}} D_{\mathbf{gg}}\right)^{-1}, \tag{A.A.3}$$

where $D_{\mathbf{gg}} = \nabla_{\theta_{\mathbf{g}}} \mathbf{E}[m_t(\boldsymbol{\theta}) \mathbf{r}^e_t]$ and $\Omega_{\mathbf{gg}}$ is the long-run covariance of the orthogonalized moment condition. The orthogonalization of $\mathbf{g}$ against $\mathbb{S}$ renders the Fisher information matrix block diagonal, so $\hat{\theta}_{\mathbf{g}}$ is asymptotically independent of $\hat{\theta}_{\mathbb{S}}$.

Given any consistent estimator $\hat{\sigma}^2_{\theta_{\mathbf{g}},\infty}$ of $\sigma^2_{\theta_{\mathbf{g}},\infty}$, the Wald statistic admits the scale-invariant representation

$$W_T = \frac{\hat{\theta}^2_{\mathbf{g}}}{\hat{\sigma}^2_{\theta_{\mathbf{g}}}} = \frac{\left(\sqrt{T}\hat{\theta}_{\mathbf{g}}\right)^2}{\hat{\sigma}^2_{\theta_{\mathbf{g}},\infty}} \xrightarrow{\mathcal{D}} \frac{Z^2}{1} = \chi^2_1, \qquad Z \sim \mathcal{N}(0,1), \tag{A.A.4}$$

by the continuous mapping theorem. This is the standard Wald result of Wald (1943); see Engle (1984) for a unified treatment.

Under H_1, local asymptotic normality Kitamura and Stutzer (1997) gives

$$\sqrt{T}\left(\hat{\theta}_{\mathbf{g}} - \theta^*_{\mathbf{g}}\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sigma^2_{\theta_g,\infty}\right), \tag{A.A.5}$$

so that $\sqrt{T}\hat{\theta}_{\mathbf{g}} \xrightarrow{\mathcal{D}} \mathcal{N}\left(\sqrt{T}\theta^*_{\mathbf{g}}, \sigma^2_{\theta_{\mathbf{g}},\infty}\right)$. It follows that

$$W_T = \left(\frac{\sqrt{T}\hat{\theta}_{\mathbf{g}}}{\hat{\sigma}_{\theta_{\mathbf{g}},\infty}}\right)^2 + o_{\mathcal{P}}(1) \xrightarrow{\mathcal{D}} \chi^2_1(\delta^2), \qquad \delta^2 := \frac{T(\theta^*_{\mathbf{g}})^2}{\sigma^2_{\theta_{\mathbf{g}},\infty}}. \tag{A.A.6}$$

We now relate the noncentrality parameter to the KL divergence between the augmented and restricted models. By the Bregman-divergence representation (7),

$$D_{KL}\left(\mathcal{Q}^{(\mathbb{S},\mathbf{g})} || \mathcal{Q}^{\mathbb{S}}\right) = \psi(\boldsymbol{\theta}^{(\mathbb{S},\mathbf{g})}) - \psi(\boldsymbol{\theta}^{\mathbb{S}}) - \nabla\psi(\boldsymbol{\theta}^{(\mathbb{S},\mathbf{g})})^\top \left(\boldsymbol{\theta}^{(\mathbb{S},\mathbf{g})} - \boldsymbol{\theta}^{\mathbb{S}}\right). \tag{A.A.7}$$

Orthogonalization ensures $\boldsymbol{\theta}^{(\mathbb{S},\mathbf{g})} - \boldsymbol{\theta}^{\mathbb{S}} = (0, \ldots, 0, \theta^*_{\mathbf{g}})^\top$, so only the g-block of the Hessian of ψ contributes. A Taylor expansion of ψ around $\boldsymbol{\theta}^{\mathbb{S}}$ yields

$$D_{KL}\left(\mathcal{Q}^{(\mathbb{S},\mathbf{g})} || \mathcal{Q}^{\mathbb{S}}\right) = \frac{1}{2}(\theta^*_{\mathbf{g}})^2 \left[\nabla^2\psi(\boldsymbol{\theta}^{\mathbb{S}})\right]_{\mathbf{gg}} + \mathcal{O}_{\mathcal{P}}\left(\left(\theta^*_{\mathbf{g}}\right)^3\right). \tag{A.A.8}$$

By the information-matrix identity for exponential families, the Hessian of the CGF equals the Fisher information, which in turn coincides (up to the long-run adjustment $\Omega_{\mathbf{gg}}$ that vanishes in the i.i.d. case) with the inverse asymptotic variance

$$\left[\nabla^2\psi(\boldsymbol{\theta}^{\mathbb{S}})\right]_{\mathbf{gg}} = \mathcal{I}_{\mathbf{gg}} = \sigma^{-2}_{\theta_{\mathbf{g}},\infty}. \tag{A.A.9}$$

Combining (A.A.8) and (A.A.9),

$$D_{KL}\left(\mathcal{Q}^{(\mathbb{S},\mathbf{g})} || \mathcal{Q}^{\mathbb{S}}\right) = \frac{(\theta^*_{\mathbf{g}})^2}{2\sigma^2_{\theta_{\mathbf{g}},\infty}} + \mathcal{O}_{\mathcal{P}}\left(\left(\theta^*_{\mathbf{g}}\right)^3\right), \tag{A.A.10}$$

so that

$$\delta^2 = \frac{T(\theta^*_g)^2}{\sigma^2_{\theta_g,\infty}} = \frac{T}{\Omega_{\mathbf{gg}}} \frac{2D_{KL}(\mathcal{Q}^{(\mathbb{S},\mathbf{g})} || \mathcal{Q}^{\mathbb{S}})}{\left(\nabla_{\theta_{\mathbf{g}}}\psi(\boldsymbol{\theta}^{(\mathbb{S},\mathbf{g})})\right)^2} + \mathcal{O}_{\mathcal{P}}\left(\left(\theta^*_{\mathbf{g}}\right)^3\right). \tag{A.A.11}$$

The identity (A.A.11) is a *local* second-order result; it equates the Wald noncentrality with twice the KL divergence up to cubic terms in the displacement $\theta^*_{\mathbf{g}}$. In the Gaussian i.i.d. baseline, the CGF ψ is quadratic in $\boldsymbol{\theta}$, the higher-order remainder vanishes identically, and one recovers the exact equality $\delta^2 = T\left(\theta^*_{\mathbf{g}}\right)^2 \sigma^2_{\mathbf{g}} = 2T \cdot D_{KL}$. Away from the Gaussian case, (A.A.11) should be read as a local approximation whose accuracy is governed by the curvature of ψ along the direction of the orthogonalized factor. □

Appendix A.B. Proposition 2

For any fixed portfolio $\mathbf{w}$, the worst case density distortion that solves the inner problem in (11) is

$$h^*(\mathbf{w}) = \frac{\exp\left(-\vartheta^{-1}u(W_1)\right)}{\mathbf{E}_{\mathcal{P}}[\exp\left(-\vartheta^{-1}u(W_1)\right)]}. \tag{A.B.1}$$

The worst-case measure $\mathcal{Q}^*_{\mathbf{w}} := h^*(\mathbf{w})d\mathcal{P}$ is an exponential tilt of the reference model in the direction of low utility.

Proof. We solve the constrained variational problem

$$V(\mathbf{w}) := \min_{h\geq 0, \mathbf{E}_{\mathcal{P}}[h]=1} \underbrace{\left\{\mathbf{E}_{\mathcal{P}}[hu(W_1)] + \vartheta\mathbf{E}_{\mathcal{P}}[h\ln(h)]\right\}}_{:=J(h)}. \tag{A.B.2}$$

Let $\varsigma \in \mathbb{R}$ be the Lagrange multiplier for the normalization constraint $\mathbf{E}_{\mathcal{P}}[h] = 1$

$$\mathcal{L}(h;\varsigma) := \mathbf{E}_{\mathcal{P}}[hu(W_1)] + \vartheta\mathbf{E}_{\mathcal{P}}[h\ln(h)] - \varsigma\left(\mathbf{E}_{\mathcal{P}}[h] - 1\right). \tag{A.B.3}$$

Note: The constraint $h \geq 0$ will be satisfied automatically by the exponential form of the solution. Taking the Gâteaux derivative w.r.t. the functional h in the direction of an arbitrary perturbation δh

$$\frac{\delta\mathcal{L}}{\delta h}(h;\varsigma) = \lim_{\epsilon\to 0}\frac{1}{\epsilon}\left(\mathcal{L}(h+\epsilon\delta h;\varsigma) - \mathcal{L}(h;\varsigma)\right) = \mathbf{E}_{\mathcal{P}}\left[\delta h u(W_1) + \vartheta\delta h(1+\ln(h)) - \varsigma\delta h\right] \tag{A.B.4}$$

where I used the fact

$$\frac{d}{d\epsilon}\left((h+\epsilon\delta h)\ln(h+\epsilon\delta h)\right)\Big|_{\epsilon=0} = \delta h(1+\ln(h)). \tag{A.B.5}$$

The F.O.C. for all admissible perturbations δh yields the pointwise condition

$$\delta h u(W_1) + \vartheta\delta h(1+\ln(h)) - \varsigma\delta h = 0 \Rightarrow h^*(\mathbf{w}) = \exp\left(\frac{\varsigma-\vartheta}{\vartheta}\right)\exp\left(-\vartheta^{-1}u(W_1)\right) \tag{A.B.6}$$

As promised this expression satisfies the non-negativity constraint for all realizations of W_1. To impose the normalization constraint $\mathbf{E}_{\mathcal{P}}[h] = 1$ we require that

$$\exp\left(\frac{\varsigma-\vartheta}{\vartheta}\right) = \frac{1}{\mathbf{E}_{\mathcal{P}}[\exp\left(-\vartheta^{-1}u(W_1)\right)]}. \tag{A.B.7}$$

Substituting this back into (A.B.6) gives the canonical Gibbs tilt

$$h^*(\mathbf{w}) = \frac{\exp\left(-\vartheta^{-1}u(W_1)\right)}{\mathbf{E}_{\mathcal{P}}[\exp\left(-\vartheta^{-1}u(W_1)\right)]}. \tag{A.B.8}$$

To show this functional is unique note that the second Gâteaux derivative of the entropy penalty

is

$$\frac{\delta^2}{\delta h^2}\mathbf{E}_{\mathcal{P}}[h\ln(h)] = \mathbf{E}_{\mathcal{P}}\left[\frac{(\delta h)^2}{h}\right] > 0 \tag{A.B.9}$$

for all perturbations $\delta h \neq 0$ with $h > 0$. Since the entropy penalty is strictly convex and the distorted expected utility is linear in the distortion h, the objective function $J(h)$ is strictly convex in h when $\vartheta > 0$. By Kiwiel (1997); Nishiyama (2020), because the constraint set $\{h \geq 0 : \mathbf{E}_{\mathcal{P}}[h] = 1\}$ is convex; a strictly convex functional on a convex constraint set has at most one minimizer, so $h^*(\mathbf{w})$ is unique.

Finally, I show that $h^*(\mathbf{w})$ is truly a minimizer and not a maximizer or saddle point. For any feasible perturbation with $\mathbf{E}_{\mathcal{P}}[\delta h] = 0$, a second order expansion gives

$$J(h^* + \epsilon\delta h) - J(h^*) = \frac{\epsilon^2\vartheta}{2}\mathbf{E}_{\mathcal{P}}\left[\frac{(\delta h)^2}{h^*}\right] + \mathcal{O}_{\mathcal{P}}\left(\epsilon^3\right) > 0, \tag{A.B.10}$$

confirming that $h^*(\mathbf{w})$ is a strict minimizer. □

Appendix A.C. Lemma 1

Under the SDF (14) the expected excess return on asset i decomposes

$$\mathbf{E}_{\mathcal{P}}\left[r_i^e\right] = -\left(\frac{\mathrm{Cov}_{\mathcal{P}}\left(m^{\mathbb{S}}, r_i^e\right)}{\mathbf{E}_{\mathcal{P}}\left[m^{\mathbb{S}}\right]} + \frac{\mathrm{Cov}_{\mathcal{P}}\left(h^*, m^{\mathbb{S}} r_i^e\right)}{\mathbf{E}_{\mathcal{P}}\left[h^* m^{\mathbb{S}}\right]}\right) \tag{A.C.1}$$

Proof. Under the misspecification adjusted SDF

$$m := h^*(\mathbf{w}^*)\frac{u'(W_1)}{R_f\mathbf{E}_{\mathcal{P}}\left[u'(W_1)\right]} = h^*(\mathbf{w}^*)m^{\mathbb{S}} \tag{A.C.2}$$

the pricing restriction for asset i is $\mathbf{E}_{\mathcal{P}}\left[h^* m^{\mathbb{S}} r_i^e\right] = 0$. Using the covariance decomposition

$$\mathbf{E}[ABC] = \mathbf{E}[A]\mathbf{E}[BC] + \mathrm{Cov}(A, BC) \tag{A.C.3}$$

and the normalization $\mathbf{E}_{\mathcal{P}}\left[h^*\right] = 1$ we write the pricing anomaly as

$$\mathbf{E}_{\mathcal{P}}\left[m^{\mathbb{S}} r_i^e\right] = -\mathrm{Cov}_{\mathcal{P}}\left(h^*(\mathbf{w}^*), m^{\mathbb{S}} r_i^e\right). \tag{A.C.4}$$

Applying the same covariance decomposition to $\mathbf{E}_{\mathcal{P}}\left[m^{\mathbb{S}} r_i^e\right]$ gives

$$\mathbf{E}_{\mathcal{P}}\left[r_i^e\right] = -\frac{\mathrm{Cov}_{\mathcal{P}}\left(m^{\mathbb{S}}, r_i^e\right)}{\mathbf{E}_{\mathcal{P}}\left[m^{\mathbb{S}}\right]} - \frac{\mathrm{Cov}_{\mathcal{P}}\left(h^*(\mathbf{w}^*), m^{\mathbb{S}} r_i^e\right)}{\mathbf{E}_{\mathcal{P}}\left[m^{\mathbb{S}}\right]}. \tag{A.C.5}$$

□

Appendix A.D. Proposition 3

Under the multiplier preferences (11) and the exponential-affine SDF specification, the worst-case welfare loss from misspecification in the direction of factor $\mathbf{g}$, to leading order, satisfies

$$|\Delta V_{\mathbf{g}}| \leq \sqrt{2\vartheta \mathbf{V}_{\mathcal{Q}^{\mathbb{S}}}(u(W_1^*))}\sqrt{D_{KL}(\mathcal{Q}^{(\mathbb{S},\mathbf{g})}||\mathcal{Q}^{\mathbb{S}})}. \tag{A.D.1}$$

Moreover, the ranking of candidates by worst-case welfare loss is identical to the ranking by KL divergence

$$|\Delta V_{\mathbf{g}}| > \left|\Delta V_{\mathbf{g}'}\right| \iff D_{KL}(\mathcal{Q}^{(\mathbb{S},\mathbf{g})}||\mathcal{Q}^{\mathbb{S}}) > D_{KL}(\mathcal{Q}^{(\mathbb{S},\mathbf{g}')}||\mathcal{Q}^{\mathbb{S}}). \tag{A.D.2}$$

Proof. Let $\mathcal{Q}^{\mathbb{S}}$ denote the baseline risk-neutral measure with Radon-Nikodym derivative

$$\frac{d\mathcal{Q}^{\mathbb{S}}}{d\mathcal{P}} = m_t(\boldsymbol{\theta}^{\mathbb{S}}) = \exp\left(-\left(\boldsymbol{\theta}^{\mathbb{S}}\right)^{\top}\mathbf{f}_t - \psi(\boldsymbol{\theta}^{\mathbb{S}})\right). \tag{A.D.3}$$

Similarly, let $\mathcal{Q}^{(\mathbb{S},\mathbf{g})}$ denote the augmented measure with the augmented risk price vector $\boldsymbol{\theta}^{(\mathbb{S},\mathbf{g})}$. The parameter vectors $\boldsymbol{\theta}^{\mathbb{S}}, \boldsymbol{\theta}^{(\mathbb{S},\mathbf{g})} \in \mathbb{R}^K$; the components corresponding to factors not in $\mathbb{S}$ are zero in $\boldsymbol{\theta}^{\mathbb{S}}$, and the component corresponding to the candidate $\mathbf{g}$ takes the value $\theta^*_{\mathbf{g}}$ in $\boldsymbol{\theta}^{(\mathbb{S},\mathbf{g})}$.

Let

$$\mathbf{g}_t^{\perp} := \mathbf{g}_t - \mathrm{Proj}^{\mathcal{Q}^{\mathbb{S}}}_{\mathrm{span}(\mathbb{S})}(\mathbf{g}_t) \tag{A.D.4}$$

be the orthogonalized factor defined in Lemma 3. By Assumption 1 the moment generating function $\mathbf{E}_{\mathcal{P}}\left[\exp\left(-\boldsymbol{\theta}^{\top}\mathbf{f}_t\right)\right]$ exists and is finite on an open set containing $\Theta \subseteq \mathbb{R}^K$, the factors are non-degenerate under $\mathcal{P}$, and $u(W_1^*)$ has finite variance under $\mathcal{Q}^{\mathbb{S}}$.

The indirect value function under model $\mathcal{Q}$ is

$$V(\mathbf{w};\mathcal{Q}) := -\vartheta \ln\left(\mathbf{E}_{\mathcal{Q}}\left[\exp\left(-\vartheta^{-1}u(W_1)\right)\right]\right). \tag{A.D.5}$$

The welfare damage from misspecification in the direction of $\mathbf{g}$ is

$$\Delta V_{\mathbf{g}} := V(\mathbf{w}^*;\mathcal{Q}^{(\mathbb{S},\mathbf{g})}) - V(\mathbf{w}^*;\mathcal{Q}^{\mathbb{S}}). \tag{A.D.6}$$

The proof proceeds by deriving the upper bound, the equality condition, and the ranking equivalence.

The Donsker-Varadhan sets the upper bound on the welfare damage. The Donsker-Varadhan variational representation Donsker and Varadhan (1975) states that for any two probability measures $\mathcal{Q}_1, \mathcal{Q}_2$ with $\mathcal{Q}_1 \ll \mathcal{Q}_2$ and any measurable function φ with finite moment generating function under $\mathcal{Q}_2$,

$$\ln\left(\mathbf{E}_{\mathcal{Q}_2}\left[e^{\varphi}\right]\right) = \sup_{\mathcal{Q}\ll\mathcal{Q}_2}\left\{\mathbf{E}_{\mathcal{Q}}\left[\varphi\right] - D_{KL}(\mathcal{Q}||\mathcal{Q}_2)\right\}. \tag{A.D.7}$$

Applying (A.D.7) to $t\varphi$ with the sub-Gaussian bound on the cumulant generating function

$$\ln\left(\mathbf{E}_{\mathcal{Q}_2}\left[e^{t(\varphi-\mathbf{E}_{\mathcal{Q}_2}[\varphi])}\right]\right) \leq \tfrac{1}{2}t^2\mathbf{V}_{\mathcal{Q}_2}(\varphi) \tag{A.D.8}$$

see Hoeffding (1963), and optimizing the resulting inequality over $t \in \mathbb{R}$, one obtains

$$|\mathbf{E}_{\mathcal{Q}}[\varphi] - \mathbf{E}_{\mathcal{Q}_2}[\varphi]| \leq \sqrt{2\mathbf{V}_{\mathcal{Q}_2}(\varphi)D_{KL}(\mathcal{Q}||\mathcal{Q}_2)}. \tag{A.D.9}$$

Applying (A.D.9) with $\mathcal{Q} = \mathcal{Q}^{(\mathbb{S},\mathbf{g})}$, $\mathcal{Q}_2 = \mathcal{Q}^{\mathbb{S}}$, and $\varphi = u(W_1^*)$:

$$\left|\mathbf{E}_{\mathcal{Q}^{(\mathbb{S},\mathbf{g})}}[u(W_1^*)] - \mathbf{E}_{\mathcal{Q}^{\mathbb{S}}}[u(W_1^*)]\right| \leq \sqrt{2\mathbf{V}_{\mathcal{Q}^{\mathbb{S}}}\left(u(W_1^*)\right) D_{KL}\left(\mathcal{Q}^{(\mathbb{S},\mathbf{g})}||\mathcal{Q}^{\mathbb{S}}\right)}. \tag{A.D.10}$$

Relating (A.D.10) to the welfare damage (A.D.6). For the augmented model $\mathcal{Q}^{(\mathbb{S},\mathbf{g})}$ we have

$$\mathbf{E}_{\mathcal{Q}^{(\mathbb{S},\mathbf{g})}}\left[\exp\left(-\vartheta^{-1}u(W_1^*)\right)\right] = \mathbf{E}_{\mathcal{Q}^{\mathbb{S}}}\left[\frac{d\mathcal{Q}^{(\mathbb{S},\mathbf{g})}}{d\mathcal{Q}^{\mathbb{S}}}\exp\left(-\vartheta^{-1}u(W_1^*)\right)\right]. \tag{A.D.11}$$

On the exponential-affine manifold, the likelihood ratio between the augmented and selected measures is

$$\frac{d\mathcal{Q}^{(\mathbb{S},\mathbf{g})}}{d\mathcal{Q}^{\mathbb{S}}} = \frac{m_t(\boldsymbol{\theta}^{(\mathbb{S},\mathbf{g})})}{m_t(\boldsymbol{\theta}^{\mathbb{S}})} = \exp\left(-\theta^*_{\mathbf{g}}\mathbf{g}_t^{\perp} - \left(\psi(\boldsymbol{\theta}^{(\mathbb{S},\mathbf{g})}) - \psi(\boldsymbol{\theta}^{\mathbb{S}})\right)\right) \tag{A.D.12}$$

The orthogonality of $\mathbf{g}_t^{\top}$ w.r.t. span($\mathbb{S}$) under $\mathcal{Q}^{\mathbb{S}}$ ensures that the likelihood ratio depends on $\mathbf{g}^{\perp}$ only. Substituting (A.D.12) into (A.D.11)

$$\mathbf{E}_{\mathcal{Q}^{(\mathbb{S},\mathbf{g})}}\left[\exp\left(-\vartheta^{-1}u(W_1^*)\right)\right] = \mathbf{E}_{\mathcal{Q}^{\mathbb{S}}}\left[\exp\left(-\theta^*_{\mathbf{g}}\mathbf{g}_t^{\perp} - \Delta\psi\right)\exp\left(-\vartheta^{-1}u(W_1^*)\right)\right] \tag{A.D.13}$$

$$= e^{-\Delta\psi}\mathbf{E}_{\mathcal{Q}^{\mathbb{S}}}\left[\exp\left(-\theta^*_{\mathbf{g}}\mathbf{g}_t^{\perp} - \vartheta^{-1}u(W_1^*)\right)\right] \tag{A.D.14}$$

where $\Delta\psi := \psi(\boldsymbol{\theta}^{(\mathbb{S},\mathbf{g})}) - \psi(\boldsymbol{\theta}^{\mathbb{S}})$. Define the joint cumulant generating

$$\Upsilon(a,b) := \ln\left(\mathbf{E}_{\mathcal{Q}^{\mathbb{S}}}\left[\exp\left(a(-u(W_1^*)) + b(-\mathbf{g}_t^{\perp})\right)\right]\right). \tag{A.D.15}$$

Υ is well-defined and finite in the neighborhood of the origin by 1. Taking logarithms of (A.D.14) and multiplying by $-\vartheta$ gives

$$V(\mathbf{w}^*;\mathcal{Q}^{(\mathbb{S},\mathbf{g})}) = -\vartheta\ln\left(e^{-\Delta\psi}\exp\left(\Upsilon(\vartheta^{-1},\theta^*_{\mathbf{g}})\right)\right) = \vartheta\Delta\psi - \vartheta\Upsilon(\vartheta^{-1},\theta^*_{\mathbf{g}}). \tag{A.D.16}$$

For the selected model $b = 0$,

$$V(\mathbf{w}^*;\mathcal{Q}^{\mathbb{S}}) = -\vartheta\Upsilon(\vartheta^{-1},0). \tag{A.D.17}$$

The welfare loss is

$$\Delta V_{\mathbf{g}} = \vartheta \left(\Delta\psi + \Upsilon(\vartheta^{-1}, \theta^*_{\mathbf{g}}) - \Upsilon(\vartheta^{-1}, 0) \right). \tag{A.D.18}$$

Expanding (A.D.18) to leading order. The CGF of $-u(W_1^*)$ under measure $\mathcal{Q}$ admits the cumulant expansion (McCullagh, 1987, Equation 2.4)

$$-\vartheta \ln \left(\mathbf{E}_{\mathcal{Q}} \left[e^{-\vartheta^{-1} u(W_1^*)} \right] \right) = -\vartheta \sum_{n=1}^{\infty} \frac{(-\vartheta^{-1})^n}{n!} \kappa^{(n)}_{\mathcal{Q}}(u(W_1^*)) \tag{A.D.19}$$

$$= \sum_{n=1}^{\infty} \frac{(-1)^{n+1} \vartheta^{1-n}}{n!} \kappa^{(n)}_{\mathcal{Q}}(u(W_1^*)) \tag{A.D.20}$$

$$= \mathbf{E}_{\mathcal{Q}} \left[u(W_1^*) \right] - \frac{\vartheta^{-1}}{2} \mathbf{V}_{\mathcal{Q}}(u(W_1^*)) + \cdots \tag{A.D.21}$$

where $\kappa^{(n)}_{\mathcal{Q}}(u(W_1^*))$ is the n-th cumulant of under $\mathcal{Q}$. Differencing between $\mathcal{Q}^{(\mathbb{S},\mathbf{g})}$ and $\mathcal{Q}^{\mathbb{S}}$ gives

$$\Delta V_{\mathbf{g}} = \mathbf{E}_{\mathcal{Q}^{(\mathbb{S},\mathbf{g})}} \left[u(W_1^*) \right] - \mathbf{E}_{\mathcal{Q}^{\mathbb{S}}} \left[u(W_1^*) \right] - \frac{\vartheta^{-1}}{2} \left(\mathbf{V}_{\mathcal{Q}^{(\mathbb{S},\mathbf{g})}}(u(W_1^*)) - \mathbf{V}_{\mathcal{Q}^{\mathbb{S}}}(u(W_1^*)) \right) + \mathcal{O}_{\mathcal{P}} \left(\vartheta^{-2} \Delta\kappa^{(n)} \right) \tag{A.D.22}$$

where $\Delta\kappa^{(n)} := \kappa^{(n)}_{\mathcal{Q}^{(\mathbb{S},\mathbf{g})}}(u(W_1^*)) - \kappa^{(n)}_{\mathcal{Q}^{\mathbb{S}}}(u(W_1^*))$. The change of measure from $\mathcal{Q}^{\mathbb{S}}$ to $\mathcal{Q}^{(\mathbb{S},\mathbf{g})}$ is an exponential tilt with parameter $\theta^*_{\mathbf{g}}$, so each cumulant is a smooth function of $\theta^*_{\mathbf{g}}$ and $\Delta\kappa^{(n)} = \mathcal{O}_{\mathcal{P}}\left(\left|\theta^*_{\mathbf{g}}\right|\right)$, $\forall n \in \mathbb{N}_+$. The leading term is bounded by (A.D.10).

The ranking equivalence result (A.D.2) asserts that $D_{KL}(\mathcal{Q}^{(\mathbb{S},\mathbf{g})} || \mathcal{Q}^{\mathbb{S}})$ provides a sufficient statistic for ordering candidates by welfare damage. We establish this in three parts: expressing the KL divergence as a Bregman divergence, showing that both $|\Delta V_{\mathbf{g}}|$ and D_{KL} are strictly increasing functions of the same pair of sufficient statistics.

From the closed-form derivation in (7), the KL divergence on the exponential-affine manifold equals the Bregman divergence generated by ψ ichi Amari (2020). Since $\boldsymbol{\theta}^{\mathbb{S}}$ and $\boldsymbol{\theta}^{(\mathbb{S},\mathbf{g})}$ differ only in the $\mathbf{g}$-component, and the $\mathbf{g}$-component of $\boldsymbol{\theta}^{\mathbb{S}}$ is zero

$$\nabla\psi(\boldsymbol{\theta}^{(\mathbb{S},\mathbf{g})})^{\top}(\boldsymbol{\theta}^{(\mathbb{S},\mathbf{g})} - \boldsymbol{\theta}^{\mathbb{S}}) = \theta^*_{\mathbf{g}} \frac{\partial\psi}{\partial\theta_{\mathbf{g}}}(\boldsymbol{\theta}^{(\mathbb{S},\mathbf{g})}). \tag{A.D.23}$$

Substituting

$$D_{KL}(\mathcal{Q}^{(\mathbb{S},\mathbf{g})} || \mathcal{Q}^{\mathbb{S}}) = \psi(\boldsymbol{\theta}^{\mathbb{S}}) - \psi(\boldsymbol{\theta}^{(\mathbb{S},\mathbf{g})}) + \theta^*_{\mathbf{g}} \frac{\partial\psi}{\partial\theta_{\mathbf{g}}}(\boldsymbol{\theta}^{(\mathbb{S},\mathbf{g})}). \tag{A.D.24}$$

Differentiating w.r.t. $\theta^*_{\mathbf{g}}$

$$\frac{\partial D_{KL}}{\partial\theta^*_{\mathbf{g}}}(\mathcal{Q}^{(\mathbb{S},\mathbf{g})} || \mathcal{Q}^{\mathbb{S}}) = \theta^*_{\mathbf{g}} \frac{\partial^2\psi}{\partial\theta^2_{\mathbf{g}}}(\boldsymbol{\theta}^{(\mathbb{S},\mathbf{g})}) = \theta^*_{\mathbf{g}} \mathbf{V}_{\mathcal{Q}^{(\mathbb{S},\mathbf{g})}}(\mathbf{g}^{\perp}) \tag{A.D.25}$$

where $\partial^2_{\theta_{\mathbf{g}}}\psi = I_{\mathbf{gg}}(\boldsymbol{\theta}^{(\mathbb{S},\mathbf{g})}) = \mathbf{V}_{\mathcal{Q}^{(\mathbb{S},\mathbf{g})}}(\mathbf{g}^{\perp}) > 0$ by the positive-definiteness of the Fisher information 1. $\partial_{\theta_{\mathbf{g}}} D_{KL} > 0$ when $\theta_{\mathbf{g}} > 0$ and negative when $\theta_{\mathbf{g}} < 0$; so D_{KL} is strictly increasing in $\left|\theta^*_{\mathbf{g}}\right|$ with a minimum at $\theta^*_{\mathbf{g}} = 0$. Also note that the divergence is strictly increasing in the residual variance $\mathbf{V}_{\mathcal{Q}^{\mathbb{S}}}(\mathbf{g}^{\perp})$.

Now, I show that the absolute welfare loss is also monotonic in $\left|\theta^*_{\mathbf{g}}\right|$ and $\mathbf{V}_{\mathcal{Q}^{\mathbb{S}}}(\mathbf{g}^{\perp})$. Differentiating (A.D.18) with respect to $\theta^*_{\mathbf{g}}$

$$\frac{\partial \Delta V_{\mathbf{g}}}{\partial \theta^*_{\mathbf{g}}} = \vartheta \left(\frac{\partial \Delta \psi}{\partial \theta^*_{\mathbf{g}}} + \frac{\partial \Upsilon}{\partial \theta^*_{\mathbf{g}}}(\vartheta^{-1}, \theta^*_{\mathbf{g}}) \right). \tag{A.D.26}$$

The first term is

$$\frac{\partial \Delta \psi}{\partial \theta^*_{\mathbf{g}}} = \frac{\partial \psi}{\partial \theta_{\mathbf{g}}}(\boldsymbol{\theta}^{(\mathbb{S},\mathbf{g})}) = -\mathbf{E}_{\mathcal{Q}^{(\mathbb{S},\mathbf{g})}}[\mathbf{g}_t^{\perp}]. \tag{A.D.27}$$

The second term is given by

$$\frac{\partial \Upsilon}{\partial \theta^*_{\mathbf{g}}}(\vartheta^{-1}, \theta^*_{\mathbf{g}}) = -\mathbf{E}_{\Xi(\vartheta^{-1}, \theta^*_{\mathbf{g}})}[\mathbf{g}_t^{\perp}] \tag{A.D.28}$$

where $\Xi(\vartheta^{-1}, \theta^*_{\mathbf{g}})$ is the exponential tilt of $\mathcal{Q}^{\mathbb{S}}$ with Radon-Nikodym derivative

$$\frac{d\Xi}{d\mathcal{Q}^{\mathbb{S}}}(\vartheta^{-1}, \theta^*_{\mathbf{g}}) = \frac{\exp\left(-\vartheta^{-1} u(W_1^*) - \theta^*_{\mathbf{g}} \mathbf{g}_t^{\perp}\right)}{\mathbf{E}_{\mathcal{Q}^{\mathbb{S}}}\left[\exp\left(-\vartheta^{-1} u(W_1^*) - \theta^*_{\mathbf{g}} \mathbf{g}_t^{\perp}\right)\right]}. \tag{A.D.29}$$

Evaluate at $\theta^*_{\mathbf{g}} = 0$. By the orthogonalization construction, $\mathbf{E}_{\mathcal{Q}^{\mathbb{S}}}\left[\mathbf{g}_t^{\perp}\right] = 0$. The measure $\Xi(\vartheta^{-1}, 0)$ is the tilt in the $-\vartheta^{-1} u(W_1^*)$ direction alone; when $\theta^*_{\mathbf{g}} = 0$ the optimal portfolio does not load on $\mathbf{g}^{\perp}$, so $u(W_1^*)$ and $\mathbf{g}^{\perp}$ are uncorrelated under $\mathcal{Q}^{\mathbb{S}}$ and

$$\left.\frac{\partial \Delta V_{\mathbf{g}}}{\partial \theta^*_{\mathbf{g}}}\right|_{\theta^*_{\mathbf{g}}=0} = 0. \tag{A.D.30}$$

For the second derivative, the map $b \mapsto \Upsilon(\vartheta^{-1}, b)$ is the CGF of $-\mathbf{g}_t^{\perp}$ under $\Xi(\vartheta^{-1}, 0)$, and CGFs of non-degenerate random variables are strictly convex

$$\frac{\partial^2 \Upsilon}{\partial b^2}(\vartheta^{-1}, b) = \mathbf{V}_{\Xi(\vartheta^{-1}, b)}(\mathbf{g}_t^{\perp}) > 0. \tag{A.D.31}$$

The variance is strictly positive because $\mathbf{g}_t^{\perp}$ is non-degenerate under $\mathcal{Q}^{\mathbb{S}}$ and exponential tilting preserves non-degeneracy.

The strict convexity of Υ in b implies that $\Upsilon(\vartheta^{-1}, \theta^*_{\mathbf{g}}) - \Upsilon(\vartheta^{-1}, 0)$ is strictly convex in $\theta^*_{\mathbf{g}}$ with minimum zero at the origin. Similarly, $\Delta\psi$ is strictly convex in $\theta^*_{\mathbf{g}}$ with minimum zero because $I_{\mathbf{gg}} > 0$. The sum $\Delta\psi + \Upsilon(\vartheta^{-1}, \theta^*_{\mathbf{g}}) - \Upsilon(\vartheta^{-1}, 0)$ is therefore strictly convex with minimum zero at $\theta^*_{\mathbf{g}} = 0$. A strictly convex function with minimum at the origin is strictly increasing in the absolute value of its argument, so $|\Delta V_{\mathbf{g}}|$ is strictly increasing in $\left|\theta^*_{\mathbf{g}}\right|$. The dependence on $\mathbf{V}_{\mathcal{Q}^{\mathbb{S}}}(\mathbf{g}^{\perp})$ follows by the same logic: increasing $\mathbf{V}_{\mathcal{Q}^{\mathbb{S}}}(\mathbf{g}^{\perp})$ increases the curvature $\partial_b^2 \Upsilon$ and increases $\Delta\psi$ through $I_{\mathbf{gg}}$.

Define the maps

$$F : (\left|\theta^*_{\mathbf{g}}\right|, \sigma^2_{\mathbf{g}}) \longmapsto D_{KL}(\mathcal{Q}^{(\mathbb{S},\mathbf{g})}||\mathcal{Q}^{\mathbb{S}}), \tag{A.D.32}$$

$$H : (\left|\theta^*_{\mathbf{g}}\right|, \sigma^2_{\mathbf{g}}) \longmapsto |\Delta V_{\mathbf{g}}|, \tag{A.D.33}$$

where $\sigma^2_{\mathbf{g}} := \mathbf{V}_{\mathcal{Q}^{\mathbb{S}}}(\mathbf{g}^{\perp})$. Both F and H are strictly increasing in each argument, and both equal zero if and only if $\theta^*_{\mathbf{g}} = 0$.

When the baseline $\mathbb{S}$ is held fixed, the pair $(\left|\theta^*_{\mathbf{g}}\right|, \sigma^2_{\mathbf{g}})$ constitutes the only quantities that vary across candidates $\mathbf{g} \notin \mathbb{S}$. The risk price $\theta^*_{\mathbf{g}}$ is determined by the m-projection of the physical measure $\mathcal{P}$ onto $\mathbb{M}^{(\mathbb{S},\mathbf{g})}$, and $\sigma^2_{\mathbf{g}}$ is determined by the covariance structure of the factors under $\mathcal{Q}^{\mathbb{S}}$. Both are population quantities independent of the investor's preferences.

For each candidate $\mathbf{g}$, the value $\sigma^2_{\mathbf{g}}$ is fixed. Along the one-dimensional submanifold $\mathbb{M}^{(\mathbb{S},\mathbf{g})}$ parameterized by $\theta_{\mathbf{g}}$ alone, both $\Delta V_{\mathbf{g}}$ and $D_{KL}(\mathcal{Q}^{(\mathbb{S},\mathbf{g})}||\mathcal{Q}^{\mathbb{S}})$ are functions of the single scalar $\theta^*_{\mathbf{g}}$. The maps F and H, restricted to this path, are strictly increasing functions of the same scalar, so the composition $H \circ F^{-1}$ is strictly increasing on $[0, \infty)$. It follows that

$$|\Delta V_{\mathbf{g}}| > \left|\Delta V_{\mathbf{g}'}\right| \iff H(\left|\theta^*_{\mathbf{g}}\right|, \sigma^2_{\mathbf{g}}) > H(\left|\theta^*_{\mathbf{g}'}\right|, \sigma^2_{\mathbf{g}'}) \tag{A.D.34}$$

$$\iff F(\left|\theta^*_{\mathbf{g}}\right|, \sigma^2_{\mathbf{g}}) > F(\left|\theta^*_{\mathbf{g}'}\right|, \sigma^2_{\mathbf{g}'}) \tag{A.D.35}$$

$$\iff D_{KL}(\mathcal{Q}^{(\mathbb{S},\mathbf{g})}||\mathcal{Q}^{\mathbb{S}}) > D_{KL}(\mathcal{Q}^{(\mathbb{S},\mathbf{g})}||\mathcal{Q}^{\mathbb{S}}). \tag{A.D.36}$$

□

Appendix A.E. Lemma 2

The forward selection algorithm is invariant to affine reparameterization of the exponential family.

Proof. Let $\mathcal{Q}(\boldsymbol{\theta})$ be an exponential family of distributions induced by the SDF

$$m_t(\boldsymbol{\theta}) := \exp\left(-\boldsymbol{\theta}^\top \mathbf{f}_t - \psi(\boldsymbol{\theta})\right) \tag{A.E.1}$$

where $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^K$ are the canonical coordinates.

Consider an affine change of coordinates $\eta : \Theta \to \Xi$ defined by $\eta = A\theta + c$, where A is an invertible matrix and c is a constant vector. This transformation defines a new parameterization of the same family of distributions. The selection criterion is the KL divergence between the augmented and restricted models,

$$\Delta D_{KL} = D_{KL}(\mathcal{Q}^{(\mathbb{S},g)}||\mathcal{Q}^{\mathbb{S}}). \tag{A.E.2}$$

which depends only on the Radon-Nikodym derivative $d\mathcal{Q}^{(\mathbb{S},\mathbf{g})}/d\mathcal{Q}^{\mathbb{S}}$ and hence only on the distributions themselves, not on their parameterization. Because the MCE estimator is equivariant

under invertible parameter transformations Zehna (1966), the fitted distributions $\mathcal{Q}^{(\mathbb{S},\mathbf{g})}$ and $\mathcal{Q}^{\mathbb{S}}$ are identical across coordinate systems. It follows that ΔD_{KL}, and therefore the ranking of candidate factors $\mathbf{g}$, is invariant to the choice of coordinates. □

Appendix A.F. Lemma 3

Replacing a candidate factor $\mathbf{g}$ with its orthogonal residual $\mathbf{g}^{\perp} := \mathbf{g} - \text{Proj}^{\mathcal{Q}^{\mathbb{S}}}_{\text{span}(\mathbb{S})}(\mathbf{g})$ under the current risk-neutral measure $\mathcal{Q}^{\mathbb{S}}$ leaves the augmented SDF manifold and the KL divergence unchanged.

Proof. Let $\mathbb{S} = \{\mathbf{f}_1, \ldots, \mathbf{f}_k\}$ be the set of currently selected factors. The manifold of SDFs spanned by $\mathbb{S}$ is

$$\mathbb{M}^{\mathbb{S}} := \left\{ \mathcal{Q}(\boldsymbol{\theta}) : \frac{d\mathcal{Q}}{d\mathcal{P}} = \exp\left(-\sum_{i=1}^{k} \theta_i f_{i,t} - \psi(\boldsymbol{\theta}) \right) \right\} \tag{A.F.1}$$

Consider a new candidate factor $\mathbf{g}_t$. The augmented manifold $\mathbb{M}^{(\mathbb{S},\mathbf{g})}$ is spanned by the linear combination of $\mathbb{S} \cup \{\mathbf{g}\}$. We define the orthogonalized factor $\mathbf{g}_t^{\perp}$ as the residual of $\mathbf{g}_t$ after projecting onto the span of $\mathbb{S}$

$$\mathbf{g}_t^{\perp} = \mathbf{g}_t - \text{Proj}^{\mathcal{Q}^{\mathbb{S}}}_{\text{span}(\mathbb{S})}(\mathbf{g}_t) = \mathbf{g}_t - \sum_{i=1}^{k} \alpha_i f_{i,t} \tag{A.F.2}$$

where α_i are the projection coefficients under $\mathcal{Q}^{\mathbb{S}}$. For any vector $\mathbf{v}$ in the augmented span

$$\mathbf{v} = \sum_{i=1}^{k} \beta_i \mathbf{f}_i + \lambda \mathbf{g} = \sum_{i=1}^{k} \beta_i \mathbf{f}_i + \lambda \left(\mathbf{g}^{\perp} + \sum_{j=1}^{k} \alpha_j \mathbf{f}_j \right) = \sum_{i=1}^{k} (\beta_i + \lambda \alpha_i) \mathbf{f}_i + \lambda \mathbf{g}^{\perp}. \tag{A.F.3}$$

Since the set of reachable vectors is identical, the set of admissible probability distributions is identical

$$\mathbb{M}^{(\mathbb{S},\mathbf{g})} \equiv \mathbb{M}^{(\mathbb{S},\mathbf{g}^{\perp})}. \tag{A.F.4}$$

The information projection is unique. The estimation of the SDF parameters constitutes an m-projection of the true data-generating process $\mathcal{Q}^{\dagger}$ onto the model manifold. Since the two manifolds coincide, the projections coincide $\mathcal{Q}^{(\mathbb{S},\mathbf{g})} = \mathcal{Q}^{(\mathbb{S},\mathbf{g}^{\perp})}$. Therefore, the KL divergence contribution is identical

$$D_{KL}(\mathcal{Q}^{(\mathbb{S},\mathbf{g})} || \mathcal{Q}^{\mathbb{S}}) = D_{KL}(\mathcal{Q}^{(\mathbb{S},\mathbf{g}^{\perp})} || \mathcal{Q}^{\mathbb{S}}). \tag{A.F.5}$$

□

Appendix B. Derivations

Appendix B.A. Closed Form Kullback-Leibler Divergence

This appendix details the derivation of equation (7), which provides the closed-form expression for the Kullback-Leibler (KL) divergence between two exponential-affine pricing models.

Let the exponential-affine SDF be defined as

$$m_t(\boldsymbol{\theta}) = \exp\left(-\boldsymbol{\theta}^\top \mathbf{f}_t - \psi(\boldsymbol{\theta})\right) \tag{B.A.1}$$

where $\psi(\boldsymbol{\theta})$ is the cumulant generating function defined as

$$\psi(\boldsymbol{\theta}) = \ln\left(\mathbf{E}_{\mathcal{P}}\left[\exp\left(-\boldsymbol{\theta}^\top \mathbf{f}_t\right)\right]\right), \tag{B.A.2}$$

ensuring that $\mathbf{E}_{\mathcal{P}}[m_t(\boldsymbol{\theta})] = 1$.

We consider the divergence between a restricted model $\mathcal{Q}^{\mathbb{S}}$ (parameterized by $\theta^{\mathbb{S}}$) and an augmented model $\mathcal{Q}^{(\mathbb{S},\mathbf{g})}$ (parameterized by $\theta^{(\mathbb{S},\mathbf{g})}$). Both parameter vectors $\boldsymbol{\theta} \in \mathbb{R}^K$ are defined over the full factor space, where coefficients for non-selected factors are zero.

The KL divergence is defined as

$$D_{KL}\left(\mathcal{Q}^{(\mathbb{S},\mathbf{g})} || \mathcal{Q}^{\mathbb{S}}\right) = \mathbf{E}_{\mathcal{Q}^{(\mathbb{S},\mathbf{g})}}\left[\ln\left(\frac{m_t(\boldsymbol{\theta}^{(\mathbb{S},\mathbf{g})})}{m_t(\boldsymbol{\theta}^{\mathbb{S}})}\right)\right]. \tag{B.A.3}$$

Switching the expectation to the physical measure $\mathcal{P}$, we have

$$D_{KL}\left(\mathcal{Q}^{(\mathbb{S},\mathbf{g})} || \mathcal{Q}^{\mathbb{S}}\right) = \mathbf{E}_{\mathcal{P}}\left[m_t\left(\boldsymbol{\theta}^{(\mathbb{S},\mathbf{g})}\right)\ln\left(\frac{m_t(\boldsymbol{\theta}^{(\mathbb{S},\mathbf{g})})}{m_t(\boldsymbol{\theta}^{\mathbb{S}})}\right)\right]. \tag{B.A.4}$$

Substituting the exponential functional form into the log ratio

$$\ln\left(\frac{m_t(\boldsymbol{\theta}^{(\mathbb{S},\mathbf{g})})}{m_t(\boldsymbol{\theta}^{\mathbb{S}})}\right) = \left(-\left(\boldsymbol{\theta}^{(\mathbb{S},\mathbf{g})}\right)^\top \mathbf{f}_t - \psi\left(\boldsymbol{\theta}^{(\mathbb{S},\mathbf{g})}\right)\right) - \left(-\left(\boldsymbol{\theta}^{\mathbb{S}}\right)^\top \mathbf{f}_t - \psi\left(\boldsymbol{\theta}^{\mathbb{S}}\right)\right) \tag{B.A.5}$$

$$= \psi\left(\boldsymbol{\theta}^{\mathbb{S}}\right) - \psi\left(\boldsymbol{\theta}^{(\mathbb{S},\mathbf{g})}\right) - \mathbf{f}_t^\top\left(\boldsymbol{\theta}^{(\mathbb{S},\mathbf{g})} - \boldsymbol{\theta}^{\mathbb{S}}\right) \tag{B.A.6}$$

Substituting this back into the expectation

$$D_{KL}\left(\mathcal{Q}^{(\mathbb{S},\mathbf{g})} || \mathcal{Q}^{\mathbb{S}}\right) = \mathbf{E}_{\mathcal{P}}\left[m_t\left(\boldsymbol{\theta}^{(\mathbb{S},\mathbf{g})}\right)\left(\psi\left(\boldsymbol{\theta}^{\mathbb{S}}\right) - \psi\left(\theta^{(\mathbb{S},\mathbf{g})}\right) - \mathbf{f}_t^\top\left(\boldsymbol{\theta}^{(\mathbb{S},\mathbf{g})} - \boldsymbol{\theta}^{\mathbb{S}}\right)\right)\right]. \tag{B.A.7}$$

Using the linearity of expectations and the normalization condition $\mathbf{E}_{\mathcal{P}}[m_t(\boldsymbol{\theta}^{(\mathbb{S},\mathbf{g})})] = 1$

$$D_{KL}\left(\mathcal{Q}^{(\mathbb{S},\mathbf{g})} || \mathcal{Q}^{\mathbb{S}}\right) = \psi\left(\boldsymbol{\theta}^{\mathbb{S}}\right) - \psi\left(\boldsymbol{\theta}^{(\mathbb{S},\mathbf{g})}\right) - \mathbf{E}_{\mathcal{P}}\left[m_t(\boldsymbol{\theta}^{(\mathbb{S},\mathbf{g})})\mathbf{f}_t^\top\right]\left(\boldsymbol{\theta}^{(\mathbb{S},\mathbf{g})} - \boldsymbol{\theta}^{\mathbb{S}}\right). \tag{B.A.8}$$

Next, we relate the expected factor payoff to the gradient of the cumulant generating function. Taking the gradient of $\psi(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$

$$\nabla_{\boldsymbol{\theta}}\psi(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \ln \mathbf{E}_{\mathcal{P}}\left[\exp\left(-\mathbf{f}_t^\top \boldsymbol{\theta}\right)\right] \tag{B.A.9}$$

$$= \mathbf{E}_{\mathcal{P}}\left[\exp\left(-\mathbf{f}_t^\top \boldsymbol{\theta}\right)\right]^{-1} \mathbf{E}_{\mathcal{P}}\left[-\mathbf{f}_t^\top \exp\left(-\mathbf{f}_t^\top \boldsymbol{\theta}\right)\right] \tag{B.A.10}$$

$$= -\exp\left(-\psi(\boldsymbol{\theta})\right) \mathbf{E}_{\mathcal{P}}\left[m_t(\boldsymbol{\theta})\mathbf{f}_t^\top\right] \exp\left(\psi(\boldsymbol{\theta})\right) \tag{B.A.11}$$

$$= -\mathbf{E}_{\mathcal{P}}\left[m_t(\boldsymbol{\theta})\mathbf{f}_t^\top\right] \tag{B.A.12}$$

Plugging this back into the KL divergence expression gives

$$D_{KL}\left(\mathcal{Q}^{(\mathbb{S},\mathbf{g})} || \mathcal{Q}^{\mathbb{S}}\right) = \psi\left(\boldsymbol{\theta}^{\mathbb{S}}\right) - \psi\left(\boldsymbol{\theta}^{(\mathbb{S},\mathbf{g})}\right) + \nabla_{\boldsymbol{\theta}}\psi\left(\boldsymbol{\theta}^{(\mathbb{S},\mathbf{g})}\right)\left(\boldsymbol{\theta}^{(\mathbb{S},\mathbf{g})} - \boldsymbol{\theta}^{\mathbb{S}}\right). \tag{B.A.13}$$

Appendix B.B. Minimum Cross Entropy Estimator

This appendix details the derivation of the Minimum Cross-Entropy (MCE) estimator used in Section III. We follow the information-theoretic approach of Kitamura and Stutzer (1997), transforming the constrained primal problem into a computationally tractable unconstrained dual problem.

Let $p(\boldsymbol{\gamma}, \boldsymbol{\theta}) := \frac{d\mathcal{P}_{\boldsymbol{\gamma},\boldsymbol{\theta}}}{\mathcal{P}^\dagger}$; the constrained minimization uses a Lagrangian with multipliers ξ for $\int_\Omega p(\boldsymbol{\gamma}, \boldsymbol{\theta}) d\mathcal{P}^\dagger = 1$ and $\boldsymbol{\gamma} \in \mathbb{R}^N$ for the N pricing conditions

$$L(p(\boldsymbol{\gamma}, \boldsymbol{\theta}); \xi, \boldsymbol{\gamma}) := \int_\Omega \left(p(\boldsymbol{\gamma}, \boldsymbol{\theta}) \ln p(\boldsymbol{\gamma}, \boldsymbol{\theta}) - \xi(p(\boldsymbol{\gamma}, \boldsymbol{\theta}) - 1) - \boldsymbol{\gamma}^\top m_t(\boldsymbol{\theta})\mathbf{r}_t^e p(\boldsymbol{\gamma}, \boldsymbol{\theta})\right) d\mathcal{P}^\dagger \tag{B.B.1}$$

Taking the Gateaux derivative w.r.t. $p(\boldsymbol{\gamma}, \boldsymbol{\theta})$ gives the F.O.C.

$$\frac{\partial L}{\partial p(\boldsymbol{\gamma}, \boldsymbol{\theta})}(p(\boldsymbol{\gamma}, \boldsymbol{\theta}); \xi, \boldsymbol{\gamma}) = \ln p(\boldsymbol{\gamma}, \boldsymbol{\theta}) + 1 - \xi - \boldsymbol{\gamma}^\top m_t(\boldsymbol{\theta})\mathbf{r}_t^e = 0. \tag{B.B.2}$$

Which will have a solution of the form

$$p(\boldsymbol{\gamma}, \boldsymbol{\theta}) \propto \exp\left(\boldsymbol{\gamma}^\top m_t(\boldsymbol{\theta})\mathbf{r}_t^e\right) \tag{B.B.3}$$

Normalizing B.B.3, gives the exponential/Gibbs tilt

$$\frac{d\mathcal{P}_{\boldsymbol{\gamma},\boldsymbol{\theta}}}{d\mathcal{P}^\dagger} = \frac{\exp\left(\boldsymbol{\gamma}^\top m_t(\boldsymbol{\theta})\mathbf{r}_t^e\right)}{\mathbf{E}_{\mathcal{P}^\dagger}\left[\exp\left(\boldsymbol{\gamma}^\top m_t(\boldsymbol{\theta})\mathbf{r}_t^e\right)\right]} \tag{B.B.4}$$

where we choose $\boldsymbol{\gamma}$ so that the pricing conditions hold under $\mathcal{P}_{\boldsymbol{\gamma},\boldsymbol{\theta}}$, $\mathbf{E}_{\mathcal{P}_{\boldsymbol{\gamma},\boldsymbol{\theta}}}\left[m_t(\boldsymbol{\theta})\mathbf{r}_t^e\right] = \mathbf{0}$. This is a challenging problem to solve in function space, but under the empirical measure, it becomes a finite-dimensional convex optimization problem.

The dual objective is the cumulant-generating-function of $\boldsymbol{\gamma}^\top m_t(\boldsymbol{\theta})\mathbf{r}_t^e$

$$\phi_T(\boldsymbol{\gamma};\boldsymbol{\theta}) := \ln\left(\frac{1}{T}\sum_{t=1}^{T}\exp\left(\boldsymbol{\gamma}^\top m_t(\boldsymbol{\theta})\mathbf{r}_t^e\right)\right). \tag{B.B.5}$$

This can be seen by plugging in the Gibbs tilt probability density $\mathcal{P}_{\boldsymbol{\gamma},\boldsymbol{\theta}}$ back into the Lagrangian B.B.1 and minimizing over densities.

Define softmax weights

$$w_t(\boldsymbol{\gamma},\boldsymbol{\theta}) := \frac{\exp\left(\boldsymbol{\gamma}^\top m_t(\boldsymbol{\theta})\mathbf{r}_t^e\right)}{\sum_{\tau=1}^{T}\exp\left(\boldsymbol{\gamma}^\top m_t(\boldsymbol{\theta})\mathbf{r}_t^e\right)}. \tag{B.B.6}$$

Computing the gradient and Hessian of ϕ_T w.r.t. $\boldsymbol{\gamma}$ gives

$$\nabla_{\boldsymbol{\gamma}}\phi_T(\boldsymbol{\gamma},\boldsymbol{\theta}) = \sum_{t=1}^{T} w_t(\boldsymbol{\gamma},\boldsymbol{\theta}) m_t(\boldsymbol{\theta})\mathbf{r}_t^e \tag{B.B.7}$$

and

$$\nabla^2_{\boldsymbol{\gamma}}\phi_T(\boldsymbol{\gamma},\boldsymbol{\theta}) = \sum_{t=1}^{T} w_t(\boldsymbol{\gamma},\boldsymbol{\theta})\left(m_t(\boldsymbol{\theta})\mathbf{r}_t^e - \nabla_{\boldsymbol{\gamma}}\phi_T(\boldsymbol{\gamma},\boldsymbol{\theta})\right)\left(m_t(\boldsymbol{\theta})\mathbf{r}_t^e - \nabla_{\boldsymbol{\gamma}}\phi_T(\boldsymbol{\gamma},\boldsymbol{\theta})\right)^\top. \tag{B.B.8}$$

The optimization problem in $\boldsymbol{\gamma}$ is globally convex; the Hessian $\nabla^2_{\boldsymbol{\gamma}}\phi_T(\boldsymbol{\gamma},\boldsymbol{\theta})$ is a weighted covariance, hence semidefinite and when $\{m_t(\boldsymbol{\theta})\mathbf{r}_t^e\}$ spans $\mathbb{R}^N$ it is positive definite. I solve for γ by optimizing the moment conditions $\nabla_{\boldsymbol{\gamma}}\phi_T(\boldsymbol{\gamma}^*,\boldsymbol{\theta}) = \mathbf{0}$, i.e., the tilted moments are zero. When $\nabla^2_{\boldsymbol{\gamma}}\phi_T(\boldsymbol{\gamma},\boldsymbol{\theta})$ is positive semidefinite, ϕ_T is convex in $\boldsymbol{\gamma}$, the minimizer exists and is unique Boyd and Vandenberghe (2009). Under standard smoothness assumptions and a Wolfe line search the Limited Memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) optimizer converges rapidly Liu and Nocedal (1989); Jin, Jiang, and Mokhtari (2025); Politis and Romano (1994).

After optimizing for $\boldsymbol{\gamma}^*$ we optimize over $\boldsymbol{\theta}$; define the profiled objective

$$J_T(\boldsymbol{\theta}) := \phi_T(\boldsymbol{\gamma}^*(\boldsymbol{\theta}),\boldsymbol{\theta}). \tag{B.B.9}$$

By the Envelope Theorem,

$$\nabla_{\boldsymbol{\theta}} J_T(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}\phi_T(\boldsymbol{\gamma}^*(\boldsymbol{\theta}),\boldsymbol{\theta}) = \sum_{t=1}^{T} w_t(\boldsymbol{\gamma}^*(\boldsymbol{\theta}),\boldsymbol{\theta})\boldsymbol{\gamma}^\top\nabla_{\boldsymbol{\theta}} m_t(\boldsymbol{\theta})\mathbf{r}_t^e. \tag{B.B.10}$$

Taking the gradient of the pricing restrictions

$$\nabla_{\boldsymbol{\theta}} m_t(\boldsymbol{\theta})\mathbf{r}_t^e = -m_t(\boldsymbol{\theta})\mathbf{r}_t^e\left(\mathbf{f}_t - \frac{1}{T}\sum_{\tau=1}^{T} m_\tau(\boldsymbol{\theta})\mathbf{f}_\tau\right)^\top, \tag{B.B.11}$$

So the objective for the outer-$\boldsymbol{\theta}$ loop is

$$\nabla_{\boldsymbol{\theta}} J_T(\boldsymbol{\theta}) = -\sum_{t=1}^{T} w_t(\boldsymbol{\gamma}^*(\boldsymbol{\theta}), \boldsymbol{\theta}) m_t(\boldsymbol{\theta}) \boldsymbol{\gamma}^*(\boldsymbol{\theta})^\top \mathbf{r}_t^e \left(\mathbf{f}_t - \frac{1}{T} \sum_{\tau=1}^{T} m_\tau(\boldsymbol{\theta}) \mathbf{f}_\tau \right) \tag{B.B.12}$$

Each loop completes different parts of the optimization; the inner $\boldsymbol{\gamma}$ loop chooses tilted probabilities $w_t(\boldsymbol{\gamma}, \boldsymbol{\theta})$ that make the pricing restrictions hold, while the outer $\boldsymbol{\theta}$ loop changes the SDF weights to reduce the Kullback-Leibler divergence between the candidate pricing kernel and the true pricing kernel.

Appendix B.C. Asymptotic Standard Errors

This appendix details the derivation of the standard errors for the parameter estimates and the Kullback-Leibler (KL) divergence, discusses the testing procedure for nested models, and establishes the conditions for parameter stability. We assume throughout that the stochastic process generating the data is stationary and ergodic and that the relevant moment conditions satisfy the standard regularity conditions for identification and differentiability.

We employ the estimator of Kitamura and Stutzer (1997), which minimizes the Kullback-Leibler (KL) divergence between the estimated probability measure and the empirical measure. This is done under the exponential tilting.

Let $\mathbf{x}_t$ be a stationary, ergodic sequence of random vectors for $t = 1, \ldots, T$. Let $\mathbf{h}(\mathbf{x}_t, \boldsymbol{\theta}) := m_t(\boldsymbol{\theta}) \mathbf{r}_t^e$ denote the N-dimensional vector of moment conditions dependent on parameters $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^k$. The estimator is defined by the saddle-point problem

$$\hat{\boldsymbol{\theta}} := \arg \min_{\boldsymbol{\theta} \in \Theta} \max_{\boldsymbol{\gamma} \in \mathbb{R}^N} \phi_T(\boldsymbol{\theta}, \boldsymbol{\gamma}) \equiv \ln \left(\frac{1}{T} \sum_{t=1}^{T} \exp \left(\boldsymbol{\gamma}^\top \mathbf{h}(\mathbf{x}_t, \boldsymbol{\theta}) \right) \right) \tag{B.C.1}$$

where $\boldsymbol{\gamma}$ is the vector of Lagrange multipliers. Kitamura and Stutzer (1997) establish that the asymptotic variance of $\hat{\boldsymbol{\theta}}$ is identical to that of the optimal two-step efficient GMM estimator

$$\sqrt{T} \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \right) \xrightarrow{\mathcal{D}} \mathcal{N} \left(0, (D^\top \Omega^{-1} D)^{-1} \right) \tag{B.C.2}$$

where $D := \mathbf{E}_{\mathcal{P}} \left[\nabla_{\boldsymbol{\theta}} \mathbf{h}(\mathbf{x}_t, \boldsymbol{\theta}_0) \right]$ is the Jacobian of the moment conditions, and Ω is the long-run covariance matrix of the moments. To account for serial correlation and heteroskedasticity in the data, Ω is estimated using a HAC estimator with bandwidth L

$$\hat{\Omega} := \hat{\Gamma}_0 + \sum_{j=1}^{L} w(j, L) (\hat{\Gamma}_j + \hat{\Gamma}_j^\top) \tag{B.C.3}$$

where

$$\hat{\Gamma}_j := \frac{1}{T} \sum_{t=j+1}^{T} \mathbf{h}(\mathbf{x}_t, \hat{\boldsymbol{\theta}}) \mathbf{h}(\mathbf{x}_t, \hat{\boldsymbol{\theta}})^\top \tag{B.C.4}$$

and $w(j, L)$ are kernel weights.

We derive the standard errors for the minimized KL divergence statistic itself. The sample objective function be the one derived in Equation B.C.1. Let $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}})$ denote the sample estimates, and $(\boldsymbol{\theta}^*, \boldsymbol{\gamma}^*)$ denote the pseudo-true population parameters that minimize the population KL divergence. We seek the asymptotic distribution of $\hat{K} = \phi_T(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}})$. A first order expansion of $\hat{K}$ around the pseudo-true values $(\boldsymbol{\theta}^*, \boldsymbol{\gamma}^*)$

$$\hat{K} - K_0 \approx \phi_T(\boldsymbol{\theta}^*, \boldsymbol{\gamma}^*) - K_0 + \nabla_{\boldsymbol{\theta}} \phi_T(\boldsymbol{\theta}^*, \boldsymbol{\gamma}^*)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) + \nabla_{\boldsymbol{\gamma}} \phi_T(\boldsymbol{\theta}^*, \boldsymbol{\gamma}^*)(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*). \tag{B.C.5}$$

By the Envelope Theorem, the variation in the objective function with respect to the optimal parameters is zero. The FOCs for the optimization problem imply that at the optimum

$$\mathbf{E}_{\mathcal{P}} \left[\nabla_{\boldsymbol{\theta}} \phi_T(\boldsymbol{\theta}^*, \boldsymbol{\gamma}^*)\right] = 0 \quad \text{and} \quad \mathbf{E}_{\mathcal{P}} \left[\nabla_{\boldsymbol{\gamma}} \phi_T(\boldsymbol{\theta}^*, \boldsymbol{\gamma}^*)\right] = 0. \tag{B.C.6}$$

Under the assumption that the model is pseudo-true, the gradient terms with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ vanish asymptotically ($o_{\mathcal{P}}(1)$). Thus, the variation in the KL divergence is driven solely by the randomness of the moment conditions at the fixed pseudo-true parameters.

Unlike standard Maximum Likelihood or GMM, where the objective function is a sum of logs the exponential tilt objective function involves the log of a sum. Therefore, we must use the delta method to derive the asymptotic standard errors.

Define $u_t(\boldsymbol{\theta}, \boldsymbol{\gamma}) := \exp\left(\boldsymbol{\gamma}^\top \mathbf{h}(\mathbf{x}_t, \boldsymbol{\theta})\right)$; the population expectation is $\mu := \mathbf{E}_{\mathcal{P}}\left[u_t(\boldsymbol{\theta}, \boldsymbol{\gamma})\right]$ with sample average $\overline{u}_T(\boldsymbol{\theta}, \boldsymbol{\gamma}) := T^{-1} \sum_{t=1}^{T} u_t(\boldsymbol{\theta}, \boldsymbol{\gamma})$. The estimated KL divergence is $\hat{K} = \ln\left(\overline{u}_T(\boldsymbol{\theta}, \boldsymbol{\gamma})\right)$ and the true divergence is $K_0 = \ln(\mu)$. Applying the delta method to $\ln(\overline{u}_T)$ and expanding around μ gives

$$\ln\left(\overline{u}_T(\boldsymbol{\theta}, \boldsymbol{\gamma})\right) \approx \ln(\mu) + \frac{1}{\mu}(\overline{u}_T(\boldsymbol{\theta}, \boldsymbol{\gamma}) - \mu) \tag{B.C.7}$$

this implies that

$$\sqrt{T}(\hat{K} - K_0) = \sqrt{T}(\ln(\overline{u}_T(\boldsymbol{\theta}, \boldsymbol{\gamma})) - \ln(\mu)) \approx \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left(\frac{u_t(\boldsymbol{\theta}, \boldsymbol{\gamma})}{\overline{u}_T(\boldsymbol{\theta}, \boldsymbol{\gamma})} - 1 \right). \tag{B.C.8}$$

Under the null hypothesis that the model is correctly specified, where $\boldsymbol{\gamma}^* = \mathbf{0}$, this simplifies to the sample mean of the moments. Under misspecification, it follows a central limit theorem for the scalar random variable

$$d_t(\boldsymbol{\theta}^*, \boldsymbol{\gamma}^*) := \frac{u_t(\boldsymbol{\theta}^*, \boldsymbol{\gamma}^*)}{\mathbf{E}_{\mathcal{P}}\left[u_t(\boldsymbol{\theta}^*, \boldsymbol{\gamma}^*)\right]} - 1 \tag{B.C.9}$$

giving

$$\sqrt{T}(\hat{K} - K_0) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sigma_K^2\right) \tag{B.C.10}$$

where σ_K^2 is estimated using the HAC estimator on $d_t(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}})$

$$\hat{\sigma}_K^2 := \frac{1}{T}\sum_{j=1}^{T} \tilde{d}_t(\boldsymbol{\theta}, \boldsymbol{\gamma})\tilde{d}_{t-j}(\boldsymbol{\theta}, \boldsymbol{\gamma}) + \frac{2}{T}\sum_{j=1}^{L}\sum_{t=j+1}^{T} w(j, L)\tilde{d}_t(\boldsymbol{\theta}, \boldsymbol{\gamma})\tilde{d}_{t-j}(\boldsymbol{\theta}, \boldsymbol{\gamma}) \tag{B.C.11}$$

where $\tilde{d}_t(\boldsymbol{\theta}, \boldsymbol{\gamma}) := d_t(\boldsymbol{\theta}, \boldsymbol{\gamma}) - T^{-1}\sum_{t=1}^{T} d_t(\boldsymbol{\theta}, \boldsymbol{\gamma})$ and $w(j, L)$ are kernel weights.

Appendix B.D. Evaluating the Significance of a Candidate Factor

We evaluate the significance of including an additional factor $\mathbf{g}$ into the benchmark model. Let the benchmark model $\mathbb{M}^{\mathbb{S}}$ be strictly nested within the augmented model $\mathbb{M}^{(\mathbb{S},\mathbf{g})}$. Vuong (1989) establishes the limiting distribution of the likelihood ratio statistic $\Lambda_T = T(\hat{K}_1 - \hat{K}_2)$ based on the relationship between the competing models. They define the variance of the difference in the log-likelihoods as

$$\omega^2 := \mathbf{V}\left(\ln\left(\frac{d\mathcal{Q}^{(\mathbb{S},\mathbf{g})}}{d\mathcal{Q}^{\mathbb{S}}}\right)\right) = \mathbf{V}\left(\ln\left(\frac{m_t\left(\boldsymbol{\theta}^{(\mathbb{S},\mathbf{g})}\right)}{m_t\left(\boldsymbol{\theta}^{\mathbb{S}}\right)}\right)\right) \tag{B.D.1}$$

If models are non-nested, Vuong shows

$$T^{-1/2}\Lambda_T \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \omega^2\right). \tag{B.D.2}$$

If the models are nested $\mathbb{M}^{\mathbb{S}} \subset \mathbb{M}^{(\mathbb{S},\mathbf{g})}$, under the null hypothesis that the restricted model is the pseudo-true DGP $\mathbb{M}^{\mathbb{S}} \equiv \mathbb{M}^{(\mathbb{S},\mathbf{g})}$, the likelihoods are identical almost surely. Consequently, the variance degenerates $\omega^2 = 0$.

Vuong demonstrates that the normalized t-statistic is undefined in this regime. Instead, the asymptotic distribution of the likelihood ratio is a weighted sum of Chi-square distributions. Calculating the weights for the general misspecified case is computationally burdensome. However, classical asymptotic theory applies to the parameter estimates. We therefore employ the asymptotically equivalent Wald Test on the parameter estimate associated with the additional factor.

In the greedy forward selection algorithm, we orthogonalize the candidate factor $\mathbf{g}$ with respect to the span of the benchmark model factors before estimation. This orthogonality ensures that the Fisher Information matrix and the HAC covariance matrix are asymptotically block-diagonal

$$V^{(\mathbb{S},\mathbf{g})} = \begin{pmatrix} V^{(\mathbb{S})} & 0 \\ 0 & \sigma_{\mathbf{g}}^2 \end{pmatrix} \tag{B.D.3}$$

Consequently, the covariance between the estimates of the existing parameters and the new parameter vanishes $\mathrm{Cov}(\hat{\boldsymbol{\theta}}^{\mathbb{S}}, \hat{\boldsymbol{\theta}}^{(\mathbb{S},\mathbf{g})}) = 0$. This simplifies the Wald statistic from a matrix quadratic form to the square of the t-statistic for the new parameter. We test $H_0 : \theta_{\mathbf{g}} = 0$ using

$$W_T := \frac{\hat{\theta}_g^2}{\hat{\sigma}_g^2} \xrightarrow{\mathcal{D}} \chi_1^2. \tag{B.D.4}$$

Appendix C. Numerical Implementation

This section describes the numerical techniques used to keep the MCE estimator and the greedy forward selection well-behaved in the low-signal to noise ratio regime of financial returns.

Appendix C.A. Bounding the Concentration Ratio

Valid estimation of Ω and especially Ω^{-1} requires the covariance matrix of the moment conditions to be well-conditioned. We rely on Random Matrix Theory (RMT) Tropp (2011) to establish a bound on the ratio of moments N to observations T. Consider the sample covariance matrix of N i.i.d. variables with variance σ^2. As $N, T \to \infty$ with $N/T \to c \in \mathbb{U}^o$, the Marchenko-Pastur law Marchenko and Pastur (1967) states that the eigenvalues λ are distributed with support

$$\lambda \in \left[\sigma^2\left(1-\sqrt{\frac{N}{T}}\right)^2, \sigma^2\left(1+\sqrt{\frac{N}{T}}\right)^2\right]. \tag{C.A.1}$$

While the assumptions for the eigenvalues to follow this law do not seem applicable, Laloux, Cizeau, Bouchaud, and Potters (1999) find that the bulk of the eigenvalues of a financial correlation matrix fits the Marchenko-Pastur distribution (noise), while only the eigenvalues outside the upper bound represent true signal. To ensure the signal eigenvalues are distinguishable from the noise bulk and to prevent instability in matrix inversion, we impose the constraint

$$\frac{N}{T} < 0.1 \tag{C.A.2}$$

Under this requirement the condition number of the noise distribution is

$$\kappa \approx \frac{(1+\sqrt{0.1})^2}{(1-\sqrt{0.1})^2} = 3.71. \tag{C.A.3}$$

Restricting $N/T < 0.1$ keeps the condition number of the noise component low ($\kappa < 4$), ensuring that the parameter estimates $\hat{\boldsymbol{\theta}}$ and standard errors remain numerically stable. With monthly data $T = 674$ the maximum number of test assets that can be safely used is $N = 67$, for weekly data $N = 268$, and for daily data $N = 2,022$.

Appendix C.B. Numerical Implementation of the MCE Estimator

This subsection describes the numerical implementation of the Minimum Cross-Entropy (MCE) estimator in detail. While the information-theoretic approach to SDF estimation is theoretically well-grounded, its practical execution is non-trivial. The low signal-to-noise ratio inherent in financial return data produces an objective surface that is nearly flat over broad regions of the parameter space and riddled with shallow local minima. Standard gradient-based solvers fail in this environment: gradients provide little directional signal away from the solution, quasi-Newton Hessian approximations deteriorate rapidly under flat curvature, and the exponential structure of

the SDF can concentrate probability mass on isolated observations in a manner indistinguishable from optimization progress. The implementation addresses these challenges through a layered architecture combining data preconditioning, a structured nested solver, rigorous numerical health monitors, a probe-and-exact evaluation protocol, and a multi-start initialization strategy with local and global restart mechanisms.

Before estimation, I pre-condition the factor panel and the test-asset return panel using zero-phase component analysis (ZCA) whitening Kessy, Lewin, and Strimmer (2018). For a data matrix $X \in \mathbb{R}^{T\times N}$ with sample covariance $\hat{\Sigma}_X$, the ZCA whitening operator is the symmetric inverse square root

$$\hat{\Sigma}_X^{-1/2} = U_X \Lambda_X^{-1/2} U_X^\top, \tag{C.B.1}$$

where $\hat{\Sigma}_X = U_X \Lambda_X U_X^\top$ is the eigendecomposition. To guard against near-singular covariance matrices, eigenvalues below $\max(10^{-10},\ 10^{-6}\bar{\lambda})$ are clipped before inversion, where $\bar{\lambda}$ is the mean eigenvalue. The whitened series are

$$\tilde{\mathbf{f}}_t := \hat{\Sigma}_F^{-1/2}(\mathbf{f}_t - \bar{\mathbf{f}}), \qquad \tilde{\mathbf{r}}_t^e := \hat{\Sigma}_R^{-1/2}\mathbf{r}_t^e. \tag{C.B.2}$$

The asymmetric treatment of the two panels is deliberate. The factor panel is demeaned before whitening. The cumulant generating function $\psi(\boldsymbol{\theta}) := \ln T^{-1}\sum_t \exp\left(-\boldsymbol{\theta}^\top \mathbf{f}_t\right)$ already absorbs the unconditional level of the SDF, so the factor mean contributes a redundant location component that competes with $\psi(\boldsymbol{\theta})$ and inflates the condition number of the Newton system without contributing identifying information. Removing it focuses identification on factor innovations. Test-asset returns are whitened without demeaning. The Euler condition $\mathbf{E}_{\mathcal{P}}[m_t(\boldsymbol{\theta}) r_{i,t}^e] = 0$ targets the level of the risk premium, not a mean-adjusted return; subtracting the return mean would shift the pricing target and distort inference on $\boldsymbol{\theta}$.

The symmetry of the ZCA transformation has a further advantage. Among all linear whitening transformations that map the data to unit covariance, ZCA minimizes the mean-squared distortion of the original factor directions Kessy et al. (2018). This property preserves the economic interpretability of the whitened coordinates: the whitened factors remain as close as possible to the original factors in the L^2 sense. Crucially, it ensures that the optimization landscape is spherical rather than elliptical. Because the condition number of the Hessian of the inner Newton problem scales directly with the condition number of the factor covariance matrix $\kappa(\hat{\Sigma}_F)$, pre-whitening reduces the effective condition number of the Newton system from $\kappa(\hat{\Sigma}_F)$ to approximately unity, dramatically accelerating convergence of the inner solver and reducing its susceptibility to numerical breakdown.

After estimation, the whitened risk price vector $\tilde{\boldsymbol{\theta}}$ is mapped back to the original factor space via

$$\hat{\boldsymbol{\theta}} = \hat{\Sigma}_F^{-1/2}\tilde{\boldsymbol{\theta}}, \tag{C.B.3}$$

so that all reported estimates and standard errors are expressed in the units of the original, unwhitened factors.

The MCE estimator minimizes the profiled dual objective $J_T(\boldsymbol{\theta})$ through a two-loop nested structure. The outer loop iterates over the risk price vector $\boldsymbol{\theta} \in \mathbb{R}^K$ using the Limited-Memory BFGS (L-BFGS) algorithm Liu and Nocedal (1989), which approximates the inverse Hessian of J_T using the most recent $m = 12$ gradient-difference pairs. This limited-memory scheme avoids the $O(K^2)$ storage and $O(K^3)$ factorization costs of full BFGS while retaining its superlinear local convergence rate. For each $\boldsymbol{\theta}$, the inner loop solves for the Lagrange multipliers $\boldsymbol{\gamma} \in \mathbb{R}^N$ that satisfy the pricing constraints by minimizing the convex subproblem

$$\phi_T(\boldsymbol{\gamma}; \boldsymbol{\theta}) := \ln \left(\frac{1}{T} \sum_{t=1}^{T} \exp \left(m_t(\boldsymbol{\theta}) \mathbf{r}_t^{e\top} \boldsymbol{\gamma} \right) \right), \tag{C.B.4}$$

where $m_t(\boldsymbol{\theta}) = \exp\left(-\boldsymbol{\theta}^\top \mathbf{f}_t - \psi(\boldsymbol{\theta})\right)$ is the normalized SDF. The profiled objective is $J_T(\boldsymbol{\theta}) = \phi_T(\boldsymbol{\gamma}^*(\boldsymbol{\theta}); \boldsymbol{\theta})$, where $\boldsymbol{\gamma}^*(\boldsymbol{\theta}) = \arg\min_{\boldsymbol{\gamma}} \phi_T(\boldsymbol{\gamma}; \boldsymbol{\theta})$.

Both the cumulant $\psi(\boldsymbol{\theta})$ and the inner objective $\phi_T(\boldsymbol{\gamma}; \boldsymbol{\theta})$ require computing the logarithm of a sample average of exponentials. Direct evaluation is numerically unsafe: arguments that differ by more than the floating-point dynamic range cause overflow or underflow in the exponentials, producing $\pm\infty$ or NaN. I apply the standard log-sum-exp identity

$$\ln \left(\frac{1}{T} \sum_{t} \exp\left(z_t\right) \right) = \max_t \{z_t\} + \ln \left(\frac{1}{T} \sum_{t} \exp\left(z_t - \max_j \{z_j\} \right) \right), \tag{C.B.5}$$

shifting all arguments by the running maximum before exponentiation. The shift reduces the largest argument to zero and keeps all exponential inputs non-positive, eliminating overflow. The shifted quantities remain representable in double precision even when the raw arguments span hundreds of natural-log units, as is common when $\boldsymbol{\theta}$ is far from the optimum during early iterations or when factor realizations are extreme.

For fixed $\boldsymbol{\theta}$, the inner subproblem is strictly convex Kitamura and Stutzer (1997), and I solve it using damped Newton's method with adaptive Levenberg-Marquardt regularization Levenberg (1944); Marquardt (1963). The gradient and Hessian of ϕ_T with respect to $\boldsymbol{\gamma}$ are

$$\nabla_{\boldsymbol{\gamma}} \phi_T = \frac{1}{T} \sum_{t=1}^{T} w_t m_t \mathbf{r}_t^e \equiv R^\top (\mathbf{w} \odot \mathbf{m}), \tag{C.B.6}$$

$$\nabla^2_{\boldsymbol{\gamma}} \phi_T = R^\top \operatorname{diag}(\mathbf{w} \odot \mathbf{m}^2) R - \left(R^\top (\mathbf{w} \odot \mathbf{m}) \right) \left(R^\top (\mathbf{w} \odot \mathbf{m}) \right)^\top, \tag{C.B.7}$$

where $w_t \propto \exp\left(m_t \mathbf{r}_t^{e\top} \boldsymbol{\gamma}\right)$ are the normalized exponential tilting weights, $R \in \mathbb{R}^{T \times N}$ is the matrix of excess returns, and $\odot$ denotes elementwise multiplication. The Hessian is a rank-one downward correction to a weighted outer product sum, and its efficient evaluation and inversion depend on the relative magnitudes of T and N.

When the number of test assets is moderate ($N \leq 512$), I assemble the $N \times N$ Hessian explicitly and factorize it via Cholesky decomposition using a standard dense linear algebra routine. A Levenberg-Marquardt damping term λI_N is added to the Hessian before factorization to ensure positive definiteness. The damping parameter λ is initialized at 10^{-4} and governed by an adaptive schedule Nocedal and Wright (2006): λ increases by a factor of ten when the ratio of actual to predicted objective reduction falls below 0.1 or when three consecutive backtracking steps occur; it decreases toward the floor 10^{-8} when two successive Newton steps achieve a gradient reduction factor of 0.7 or better. This schedule engages strong regularization precisely when the curvature is ill-determined and releases it as the solver approaches a well-conditioned region.

For Hessians that remain ill-conditioned even after damping (condition number exceeding 10^7), I compute the full eigendecomposition of the pre-damping Hessian and restrict the Newton step to the subspace spanned by eigenmodes with eigenvalues above a floor $\varepsilon_{\text{floor}} = \max(10^{-9},\ 10^{-7}\lambda_{\max})$, where $\lambda_{\max}$ is the largest eigenvalue. This truncated Newton step Dembo, Eisenstat, and Steihaug (1982) discards directions of near-zero curvature, which would otherwise produce Newton steps of unbounded norm and cause the inner solver to overshoot the solution by large amounts. If the retained rank is too small to form a reliable direction, the solver falls back to a gradient step scaled by the reciprocal of the largest eigenvalue.

When N is large, assembling and factorizing an $N \times N$ matrix at each Newton step becomes computationally prohibitive. I instead exploit the Woodbury matrix identity to solve the damped Newton system implicitly using only $T \times T$ computations. Writing the damped Hessian as $H + \lambda I_N$, where $H = R^\top D R - \boldsymbol{\mu}\boldsymbol{\mu}^\top$ with $D = \operatorname{diag}(\mathbf{w} \odot \mathbf{m}^2)$ and $\boldsymbol{\mu} = R^\top(\mathbf{w} \odot \mathbf{m})$, define the scaled return matrix $\tilde{R} = D^{1/2} R \in \mathbb{R}^{T \times N}$ and the $T \times T$ kernel $G = I_T + \lambda^{-1} \tilde{R}\tilde{R}^\top$. The action of the

inverse on a vector $\mathbf{v}$ is then

$$(R^\top D R + \lambda I_N)^{-1}\mathbf{v} = \frac{\mathbf{v}}{\lambda} - \frac{1}{\lambda^2}\tilde{R}^\top G^{-1}\tilde{R}\mathbf{v}, \tag{C.B.8}$$

where G is solved by a $T \times T$ Cholesky factorization. Because $T \ll N^2$ in all empirically relevant configurations, this reduces the dominant computational cost from $O(N^3)$ to $O(T^2 N + T^3)$ per Newton step. The rank-one correction from the $\boldsymbol{\mu}\boldsymbol{\mu}^\top$ term is handled by a second application of the Sherman-Morrison formula Nocedal and Wright (2006), with a positivity guard ensuring the effective denominator does not change sign.

The inner solver terminates when the gradient infinity norm falls below the tolerance $\|\nabla_{\boldsymbol{\gamma}} \phi_T\|_\infty \leq \varepsilon_\gamma$ (default $\varepsilon_\gamma = 10^{-8}$), or when the Newton decrement

$$\delta := \sqrt{\nabla_{\boldsymbol{\gamma}}^\top (H + \lambda I_N)^{-1} \nabla_{\boldsymbol{\gamma}}} \tag{C.B.9}$$

falls below 10^{-6}. The Newton decrement provides a scale-free stopping criterion that accounts for the local curvature of the objective and correctly measures the remaining reducible gap regardless of the problem scale Boyd and Vandenberghe (2009). A hard budget of 800 inner iterations is imposed as a fallback.

The gradient of the profiled objective with respect to $\boldsymbol{\theta}$ is computed via the envelope theorem. Because $\boldsymbol{\gamma}^*(\boldsymbol{\theta})$ satisfies the first-order condition $\nabla_{\boldsymbol{\gamma}}\phi_T = \mathbf{0}$ at optimality, the total derivative of $J_T(\boldsymbol{\theta}) = \phi_T(\boldsymbol{\gamma}^*(\boldsymbol{\theta}); \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ equals the partial derivative of ϕ_T evaluated at $\boldsymbol{\gamma}^*(\boldsymbol{\theta})$, with no contribution from the implicit dependence of $\boldsymbol{\gamma}^*$ on $\boldsymbol{\theta}$:

$$\nabla_{\boldsymbol{\theta}} J_T(\boldsymbol{\theta}) = -\sum_{t=1}^{T} w_t^* m_t(\boldsymbol{\theta})(\boldsymbol{\gamma}^*)^\top \mathbf{r}_t^e \left(\mathbf{f}_t - \frac{1}{T}\sum_{\tau=1}^{T} m_\tau(\boldsymbol{\theta})\mathbf{f}_\tau \right). \tag{C.B.10}$$

This expression is efficient to evaluate given the inner solver output $(\boldsymbol{\gamma}^*, \{w_t^*\})$ and requires no additional linear system solves beyond those already performed in the inner loop.

Each L-BFGS outer iteration computes a descent direction $d^{(\ell)}$ and selects a step size α via a strong Wolfe line search Wolfe (1969); Jin et al. (2025). The search enforces two conditions simultaneously. The Armijo sufficient-decrease condition Armijo (1966) requires

$$J_T\left(\boldsymbol{\theta}^{(\ell)} + \alpha d^{(\ell)}\right) \leq J_T\left(\boldsymbol{\theta}^{(\ell)}\right) + c_1 \alpha \nabla J_T^{(\ell)\top} d^{(\ell)}, \quad c_1 = 10^{-4}, \tag{C.B.11}$$

and the strong curvature condition requires

$$\left| \nabla J_T\left(\boldsymbol{\theta}^{(\ell)} + \alpha d^{(\ell)}\right)^\top d^{(\ell)} \right| \leq c_2 \left| \nabla J_T^{(\ell)\top} d^{(\ell)} \right|, \quad c_2 = 0.9. \tag{C.B.12}$$

The search proceeds in two phases. A bracketing phase doubles the candidate step size, starting from the unit step, until the Armijo condition is violated or α reaches a ceiling of 64.0. A zoom phase then applies cubic interpolation within the bracket to locate a step satisfying both conditions, falling back to bisection if interpolation produces an iterate outside the bracket. Up to 30 zoom iterations per outer step are permitted.

The outer loop terminates when the gradient satisfies $\|\nabla_{\boldsymbol{\theta}} J_T\|_\infty \leq \varepsilon_\theta$ (default $\varepsilon_\theta = 5 \times 10^{-4}$). A hard budget of 400 outer iterations is imposed. Because the outer problem is generally non-convex, gradient convergence does not guarantee a global minimum; the multi-start and restart procedures described below mitigate this risk.

The most distinctive feature of this implementation is the embedding of numerical health monitors—hereafter *guards*—directly into the line search protocol. A candidate step α is not committed unless it simultaneously satisfies the following conditions. If any guard triggers, the line search backtracks and attempts a shorter step, forcing the optimizer to find a direction that respects the structure of the estimation problem.

In the exponential-family setting, an observation with an extreme factor realization can attract a tilting weight near unity while all other weights collapse toward zero. This degenerate configuration reduces the effective sample size (ESS) Martino et al. (2017) to one and produces an SDF that prices assets by fitting a single period of data rather than the full distribution. I monitor the ESS

$$\text{ESS} := \frac{1}{\sum_{t=1}^{T} w_t^2}, \tag{C.B.13}$$

which ranges from one (all weight on a single observation) to T (uniform weights). A step is rejected if $\mathrm{ESS} < \max(0.15T,\ 0.5 \cdot \mathrm{ESS}_{\mathrm{prev}})$, where $\mathrm{ESS}_{\mathrm{prev}}$ is the ESS at the last accepted iterate. The relative floor $0.5 \cdot \mathrm{ESS}_{\mathrm{prev}}$ prevents rapid weight concentration even when the absolute floor has not yet been breached.

As a complement to the ESS, I monitor the ratio of the largest tilting weight to the median weight. A step is rejected if $\max_t(w_t)/\mathrm{median}(w_t) > 50$. This guard detects configurations in which a few observations capture a disproportionate fraction of probability mass without necessarily reducing the ESS below the absolute floor, catching concentration patterns that the ESS test alone may miss.

Low signal-to-noise ratios and multicollinear factors cause the inner Hessian $\nabla^2_{\boldsymbol{\gamma}}\phi_T$ to become numerically singular even after Levenberg-Marquardt damping. I evaluate the condition number of the inner Hessian at each candidate step and reject any step for which $\kappa(H + \lambda I) > 10^7$. I additionally require that the minimum eigenvalue satisfies $\lambda_{\min} \geq 10^{-8}$. These conditions ensure that the Newton step returned by the inner solver is numerically well-defined and that the envelope gradient passed to the outer L-BFGS loop is not contaminated by errors that originate in an ill-conditioned Newton system.

The inner solver seeks $\boldsymbol{\gamma}^*$ such that the weighted Euler equation residuals vanish: $R^\top(\mathbf{w} \odot \mathbf{m}) = \mathbf{0}$. If the inner solver exhausts its iteration budget before convergence, the residuals may be non-negligible, producing an incorrect envelope gradient. Any step for which $\|R^\top(\mathbf{w} \odot \mathbf{m})\|_\infty > 5 \times 10^{-3}$ is rejected. The threshold is intentionally loosened from the inner solver convergence tolerance to accommodate the probe evaluations described in the next subsection.

Evaluating $J_T(\boldsymbol{\theta})$ at a candidate step requires running the inner Newton solver to tight convergence, which is expensive when the problem is near-singular. During the line search, I use a two-stage evaluation protocol to reduce this cost without sacrificing solution quality.

In the *probe* stage, each candidate step is evaluated using the inner solver with a loosened tolerance $\varepsilon_\gamma^{\mathrm{probe}} = 10^{-3}$ and a hard budget of eight inner Newton iterations. The probe evaluation is sufficiently accurate to determine whether a step satisfies the Wolfe conditions and passes all guards. Only the step selected by the line search, the single step that survives both the Wolfe filter and the guard system, is then re-evaluated in the *exact* stage using the full inner solver at the target tolerance $\varepsilon_\gamma = 10^{-8}$. This splitting reduces the number of full inner solves per outer iteration from the number of line-search function evaluations (typically three to ten) to one, with a corresponding reduction in computation time.

Before committing any accepted step to the L-BFGS history, I apply a pre-commit filter that rejects microscopic steps offering negligible objective improvement. A step $(\Delta\boldsymbol{\theta}, \Delta J)$ is rejected if $\|\Delta\boldsymbol{\theta}\| \leq 10^{-6}(1 + \|\boldsymbol{\theta}\|)$ and $|\Delta J| \leq 10^{-12}(1 + |J|)$ simultaneously, or if the quasi-Newton curvature condition $\Delta\boldsymbol{\gamma}^\top\Delta\boldsymbol{\theta} \leq 10^{-12}\|\Delta\boldsymbol{\gamma}\|\|\Delta\boldsymbol{\theta}\|$ is violated, which would cause the L-BFGS update to introduce negative curvature information into the Hessian approximation. Rejected steps trigger an immediate trust-radius reduction rather than being appended to the L-BFGS history. This prevents the optimizer from accumulating degenerate curvature pairs that would corrupt subsequent

search directions.

The strong Wolfe line search can fail to find a valid step in three ways: the Wolfe conditions may be unsatisfiable along the current direction, the guards may reject all candidate steps, or the pre-commit filter may discard the selected step. A Nocedal and Wright (2006) trust-region strategy is layered on top of the line search to recover from each of these failures.

When the line search fails, the trust radius is contracted by a factor of 0.25, bounding the maximum permitted step norm and forcing the optimizer to take a shorter move. If failures persist across two consecutive outer iterations, a *local restart* is triggered: the L-BFGS curvature history is flushed, and the current iterate is displaced by a random perturbation $\delta\boldsymbol{\theta} = \sigma\mathbf{v}/\|\mathbf{v}\|$, where $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, I_K)$ and $\sigma \approx 0.02\|\tilde{\boldsymbol{\theta}}\|$ (roughly two percent of the current iterate norm in whitened coordinates). Flushing the L-BFGS history forces the next iteration to use a steepest-descent direction, which is always a valid descent direction regardless of how accurately the stored curvature pairs approximate the true Hessian. Up to three local restarts are permitted per optimization run. If local restarts are exhausted and the gradient norm remains above ε_θ, a single *global restart* is attempted from a fresh starting point, chosen as either $\mathbf{0}$ or a uniform random draw from $[-0.1, 0.1]^K$.

Conversely, if the trust radius collapses to its minimum floor while all guards report a healthy region, indicating that the optimizer is making negligible progress due to an artificially tight step bound rather than a genuine curvature deficit, the radius is restored to a larger value to escape the low-curvature plateau.

The non-convexity of the outer objective motivates running independent optimization trials from a diverse set of starting points. All trials are executed in parallel using a thread pool, and the global optimum is taken as the trial that achieves the lowest final dual objective value among all runs that satisfy the guard conditions at termination.

The starting points are constructed as follows. The first deterministic start is $\boldsymbol{\theta}^{(0)} = \mathbf{0}$, corresponding to the risk-neutral SDF and representing the null hypothesis of the model. The second deterministic start is an informed risk premia estimate. The trimmed mean of factor realizations (trimming five percent from each tail) serves as a robust proxy for the risk premium vector, $\hat{\boldsymbol{\lambda}} = \text{TrimMean}(F, 0.05)$. This is mapped to the whitened coordinate system via the safe inverse square root of the factor covariance matrix: $\boldsymbol{\theta}_{\text{rp}}^{(0)} = \hat{\Sigma}_F^{-1/2}\hat{\boldsymbol{\lambda}}$, where the inverse square root is computed from the clipped eigendecomposition described above. For problems in which $\boldsymbol{\theta}_{\text{rp}}^{(0)}$ has near-zero norm (below 10^{-12}), perturbations along the top-K principal directions of the factor panel are substituted.

Supplementing these two anchors, I add coordinate-axis perturbations $\pm s e_k$ for $k = 1, \ldots, K$ and scales $s \in \{0.01, 0.05, 0.1\}$, applied around both the zero start and the risk premia start. Finally, I add up to two orthogonal jitter directions chosen to be maximally orthogonal to $\boldsymbol{\theta}_{\text{rp}}^{(0)}$ in the span of the principal factor directions, increasing coverage of parameter regions that the systematic starts may under-explore. If no run converges successfully, the optimizer returns a null result and a non-convergence flag is recorded.

REFERENCES

Abarbanell, Jeffery S., and Brian J. Bushee, 1997, Abnormal returns to a fundamental analysis strategy, *The Accounting Review* 72, 1–31.

Akaike, Hirotugu, 1974, A new look at the statistical model identification, *IEEE Transactions on Automatic Control* 19, 716–723.

Amihud, Yakov, 2002, Illiquidity and stock returns: Cross-section and time-series effects, *Journal of Financial Markets* 5, 31–56.

Armijo, Larry, 1966, Minimization of functions having Lipschitz continuous first partial derivatives, *Pacific Journal of Mathematics* 16, 1–3.

Asness, Clifford, Andrea Frazzini, Ronen Israel, Tobias J. Moskowitz, and Lasse H. Pedersen, 2018, Size matters, if you control your junk, *Journal of Financial Economics* 129, 479–509.

Asness, Clifford S., Andrea Frazzini, and Lasse Heje Pedersen, 2019, Quality minus junk, *Review of Accounting Studies* 24, 34–112.

Baker, Malcolm, Brendan Bradley, and Jeffrey Wurgler, 2011, Benchmarks as limits to arbitrage: Understanding the low-volatility anomaly, *Financial Analysts Journal* 67, 40–54.

Bali, Turan, Robert Engle, and Scott Murray, 2017, *Empirical Asset Pricing: The Cross Section of Stock Returns: An Overview*, 1–8 (John Wiley & Sons, Ltd).

Ball, Ray, Joseph Gerakos, Juhani T. Linnainmaa, and Valeri Nikolaev, 2016, Accruals, cash flows, and operating profitability in the cross section of stock returns, *Journal of Financial Economics* 121, 28–45.

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

Barbee, William C., Sandip Mukherji, and Gary A. Raines, 1996, Do salesâ€“price and debtâ€“equity explain stock returns better than bookâ€“market and firm size?, *Financial Analysts Journal* 52, 56–60.

Barro, Robert J., 2006, Rare disasters and asset markets in the twentieth century, *The Quarterly Journal of Economics* 121, 823–866.

Barroso, Pedro, and Pedro Santa-Clara, 2015, Momentum has its moments, *Journal of Financial Economics* 116, 111–120.

Basu, Sanjoy, 1983, The relationship between earnings' yield, market value, and return for nyse common stocks: Further evidence, *Journal of Financial Economics* 12, 129–156.

Blitz, David, Joop Huij, and Martin Martens, 2011, Residual momentum, *Journal of Empirical Finance* 18, 506–521.

Bouchaud, Jean- Philippe, Philipp Krüger, Augustin Landier, and David Thesmar, 2019, Sticky expectations and the profitability anomaly, *Journal of Finance* 74, 639–674.

Boyd, Stephen, and Lieven Vandenberghe, 2009, *Convex Optimization* (Cambridge University Press).

Bradshaw, Mark T., Scott A. Richardson, and Richard G. Sloan, 2006, The relation between corporate financing activities, analysts' forecasts and stock returns, *Journal of Accounting and Economics* 42, 53–85, Conference Issue on Implications of Changing Financial Reporting Standards.

Brennan, Michael J., Tarun Chordia, and Avanidhar Subrahmanyam, 1998, Alternative factor specifications, security characteristics, and the cross-section of expected stock returns, *Journal of Financial Economics* 49, 345–373.

Chan, Louis K. C., Narasimhan Jegadeesh, and Josef Lakonishok, 1996, Momentum strategies, *The Journal of Finance* 51, 1681–1713.

Chen, Andrew Y., and Tom Zimmermann, 2022, Open source cross-sectional asset pricing, *Critical Finance Review* 11, 207–264.

Chordia, Tarun, Avanidhar Subrahmanyam, and Ravi Anshuman, 2001, Trading activity and expected stock returns, *Journal of Financial Economics* 59, 3–32.

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

Daniel, Kent, and Tobias J. Moskowitz, 2016, Momentum crashes, *Journal of Financial Economics* 122, 221–247.

Dembo, Ron S., Stanley C. Eisenstat, and Trond Steihaug, 1982, Inexact Newton methods, *SIAM Journal on Numerical Analysis* 19, 400–408.

Dichev, Ilia D., 1998, Is the risk of bankruptcy a systematic risk?, *The Journal of Finance* 53, 1131–1147.

Dimson, Elroy, 1979, Risk measurement when shares are subject to infrequent trading, *Journal of Financial Economics* 7, 197–226.

Dittmar, Robert F., 2002, Nonlinear pricing kernels, kurtosis preference, and evidence from the cross section of equity returns, *The Journal of Finance* 57, 369–403.

Donsker, M. D., and S. R. S. Varadhan, 1975, Asymptotic evaluation of certain markov process expectations for large time, i, *Communications on Pure and Applied Mathematics* 28, 1–47.

Engle, Robert F., 1984, Wald, likelihood ratio, and lagrange multiplier tests in econometrics, in Zvi Griliches, and Michael D. Intriligator, eds., *Handbook of Econometrics*, volume 2, 775–826 (North-Holland, New York).

Fama, Eugene F., and Kenneth R. French, 2002, The equity premium, *The Journal of Finance* 57, 637–659.

Feng, Guanhao, Stefano Giglio, and Dacheng Xiu, 2020, Taming the factor zoo: A test of new factors, *The Journal of Finance* 75, 1327–1370.

Frankel, Richard, and Charles M.C. Lee, 1998, Accounting valuation, market expectation, and cross-sectional stock returns, *Journal of Accounting and Economics* 25, 283–319.

Frazzini, Andrea, and Lasse Heje Pedersen, 2014, Betting against beta, *Journal of Financial Economics* 111, 1–25.

Gabaix, Xavier, 2012, Variable rare disasters: An exactly solved framework for ten puzzles in macro-finance *, *The Quarterly Journal of Economics* 127, 645–700.

George, Thomas J., and Chuan-Yang Hwang, 2004, The 52-week high and momentum investing, *The Journal of Finance* 59, 2145–2176.

Ghosh, Anisha, Christian Julliard, and Alex P. Taylor, 2016, What is the consumption-capm missing? an information-theoretic framework for the analysis of asset pricing models, *The Review of Financial Studies* 30, 442–504.

Gilboa, Itzhak, and David Schmeidler, 1989, Maxmin expected utility with non-unique prior, *Journal of Mathematical Economics* 18, 141–153.

Gram, J. P., 1883, Ueber die entwicklung reeller funktionen in reihen mittels der methode der kleinsten quadrate, *Journal für die reine und angewandte Mathematik* 94, 41–73.

Hansen, Lars Peter, and Ravi Jagannathan, 1997, Assessing specification errors in stochastic discount factor models, *The Journal of Finance* 52, 557–590.

Hansen, Lars Peter, and Thomas J. Sargent, 2001, Robust control and model uncertainty, *The American Economic Review* 91, 60–66.

Hansen, Lars Peter, and Thomas J. Sargent, 2010, Fragile beliefs and the price of uncertainty, *New York University* .

Hansen, Lars Peter, and Thomas J. Sargent, 2020, Structured ambiguity and model misspecification, Technical report, New York University.

Harvey, Campbell R., Yan Liu, and Heqing Zhu, 2015, … and the cross-section of expected returns, *The Review of Financial Studies* 29, 5–68.

Harvey, Campbell R., and Akhtar Siddique, 2000, Conditional skewness in asset pricing tests, *The Journal of Finance* 55, 1263–1295.

Haugen, Robert A., and Nardin L. Baker, 1996, Commonality in the determinants of expected stock returns, *Journal of Financial Economics* 41, 401–439.

Heston, Steven L., and Ronnie Sadka, 2008, Seasonality in the cross-section of stock returns, *Journal of Financial Economics* 87, 418–445.

Hoeffding, Wassily, 1963, Probability inequalities for sums of bounded random variables, *Journal of the American Statistical Association* 58, 13–30.

ichi Amari, Shun, 2020, *Information Geometry and Its Applications*, volume 194 of *Applied Mathematical Sciences* (Springer).

Jegadeesh, Narasimhan, 1990, Evidence of predictable behavior of security returns, *The Journal of Finance* 45, 881–898.

Jegadeesh, Narasimhan, and Joshua Livnat, 2006, Revenue surprises and stock returns, *Journal of Accounting and Economics* 41, 147–171.

Jensen, Theis Ingerslev, Bryan Kelly, and Lasse Heje Pedersen, 2023, Is there a replication crisis in finance?, *The Journal of Finance* 78, 2465–2518.

Jin, Qiujiang, Ruichen Jiang, and Aryan Mokhtari, 2025, Non-asymptotic global convergence analysis of bfgs with the armijo-wolfe line search .

Keloharju, Matti, Juhani T. Linnainmaa, and Peter Nyberg, 2016, Return seasonalities, *The Journal of Finance* 71, 1557–1590.

Kessy, Agnan, Alex Lewin, and Korbinian Strimmer, 2018, Optimal whitening and decorrelation, *The American Statistician* 72, 309–314.

Kitamura, Yuichi, and Michael Stutzer, 1997, An information-theoretic alternative to generalized method of moments estimation, *Econometrica* 65, 861–874.

Kiwiel, Krzysztof C., 1997, Proximal minimization methods with generalized bregman functions, *SIAM Journal on Control and Optimization* 35, 1142–1168.

Lakonishok, Josef, Andrei Shleifer, and Robert W. Vishny, 1994, Contrarian investment, extrapolation, and risk, *The Journal of Finance* 49, 1541–1578.

Laloux, Laurent, Pierre Cizeau, Jean-Philippe Bouchaud, and Marc Potters, 1999, Noise dressing of financial correlation matrices, *Physical Review Letters* 83, 1467–1470.

Lettau, Martin, and Markus Pelger, 2020, Estimating latent asset-pricing factors, *Journal of Econometrics* 218, 1–31.

Levenberg, Kenneth, 1944, A method for the solution of certain non-linear problems in least squares, *Quarterly of Applied Mathematics* 2, 164–168.

Liu, Dong C., and Jorge Nocedal, 1989, On the limited memory bfgs method for large scale optimization, *Mathematical Programming* 45, 503–528.

Maccheroni, Fabio, Massimo Marinacci, and Aldo Rustichini, 2006, Ambiguity aversion, robustness, and the variational representation of preferences, *Econometrica* 74, 1447–1498.

Marchenko, V. A., and L. A. Pastur, 1967, Distribution of eigenvalues for some sets of random matrices, *Mathematics of the USSR-Sbornik* 1, 457–483.

Marquardt, Donald W., 1963, An algorithm for least-squares estimation of nonlinear parameters, *Journal of the Society for Industrial and Applied Mathematics* 11, 431–441.

Martino, Luca, Víctor Elvira, and Francisco Louzada, 2017, Effective sample size for importance sampling based on discrepancy measures, *Signal Processing* 131, 386–401.

McCullagh, Peter, 1987, *Tensor Methods in Statistics*, volume 29 of *Monographs on Statistics and Applied Probability* (Chapman and Hall/CRC, London).

McLean, R. David, and Jeffrey Pontiff, 2015, Does academic research destroy stock return predictability?, *The Journal of Finance* 71, 5–32.

Nishiyama, Tomohiro, 2020, Convex optimization on functionals of probability densities.

Nocedal, Jorge, and Stephen J. Wright, 2006, *Numerical Optimization*, Springer Series in Operations Research, second edition (Springer).

Novy-Marx, Robert, 2013, The other side of value: The gross profitability premium, *Journal of Financial Economics* 108, 1–28.

Novy-Marx, Robert, 2015, Fundamentally, momentum is fundamental momentum, NBER Working Papers 20984, National Bureau of Economic Research, Inc.

Ortiz-Molina, Hernán, and Gordon M. Phillips, 2014, Real asset illiquidity and the cost of capital, *The Journal of Financial and Quantitative Analysis* 49, 1–32.

Politis, Dimitris N., and Joseph P. Romano, 1994, The stationary bootstrap, *Journal of the American Statistical Association* 89, 1303–1313.

Pontiff, Jeffrey, and Artemiza Woodgate, 2008, Share issuance and cross-sectional returns, *The Journal of Finance* 63, 921–945.

Richardson, Scott A., Richard G. Sloan, Mark T. Soliman, and Irem Tuna, 2005, Accrual reliability, earnings persistence and stock prices, *Journal of Accounting and Economics* 39, 437–485.

Schmidt, Erhard, 1907, Zur theorie der linearen und nichtlinearen integralgleichungen. i. entwicklung willkürlicher funktionen nach systemen vorgeschriebener, *Mathematische Annalen* 63, 433–476.

Shannon, C. E., 1948, A mathematical theory of communication, *Bell System Technical Journal* 27, 379–423.

Sharpe, William F., 1964, Capital asset prices: A theory of market equilibrium under conditions of risk, *The Journal of Finance* 19, 425–442.

Sloan, Richard G., 1996, Do stock prices fully reflect information in accruals and cash flows about future earnings?, *The Accounting Review* 71, 289–315.

Soliman, Mark T., 2008, The use of dupont analysis by market participants, *The Accounting Review* 83, 823–853.

Thomas, Jacob, and Frank X. Zhang, 2011, Tax expense momentum, *Journal of Accounting Research* 49, 791–821.

Titman, Sheridan, K.C. John Wei, and Feixue Xie, 2003, Capital investments and stock returns, *NBER Working Paper Series* 9951.

Tropp, Joel A., 2011, User-friendly tail bounds for sums of random matricies, *The Journal of the Society for the Foundations of Computational Mathematics* 12, 389–434.

Vuong, Quang H., 1989, Likelihood ratio tests for model selection and non-nested hypotheses, *Econometrica* 57, 307–333.

Wald, Abraham, 1943, Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations Is Large, *Transactions of the American Mathematical Society* 54, 426–482.

Wolfe, Philip, 1969, Convergence conditions for ascent methods, *SIAM Review* 11, 226–235.

Zehna, Peter W., 1966, Invariance of maximum likelihood estimators, *The Annals of Mathematical Statistics* 37, 744–744.