Topics in Statistics and Financial Econometrics: Penalized Estimators and Stochastic Discount Factors

Quaini, Alberto

Collaborators: Korsaye, Sofonias Alemu

Topics in Statistics and Financial Econometrics

Penalized Estimators and Stochastic Discount Factors

Author: **Alberto Quaini**

Thesis director: Prof. Fabio Trojani

President of the Jury: Prof. Olivier Scaillet

Jury: Prof. Elvezio Ronchetti, Prof. Patrick Gagliardini

A thesis presented for the degree of

Doctor of Philosophy

UNIVERSITÉ
DE GENÈVE

GSEM, GFRI, RCS

University of Geneva

Switzerland

13 December 2021

Introduction

This thesis is composed of two parts. The first part is based on a paper co-authored with Fabio Trojani, entitled "A Unifying Convex Analysis Framework for Penalized Least Squares". In this paper, we establish a framework for studying the statistical properties of Penalized Least Squares Estimators (PLSEs)with convex penalties, which is applicable both under a regular and a singular design. Our approach borrows from a reinterpretation of PLSEs as proximity operators and Moreau's decomposition of these operators. This allows for a general characterization of the asymp- totic properties of PLSEs, which only depends on suitable functional transformations of the PLSE limit penalty. Exploiting our approach, we propose convenient Oracle PLSEs for singular designs exhibiting the grouping effect, and valid bootstrap approximations for the associated asymptotic distributions. The second part consists in a paper co-authored with Sofonias Korsaye and Fabio Trojani, entitled "Smart Stochastic Discount Factors", which proposes a novel no-arbitrage framework exploiting convex asset pricing constraints to study investors' marginal utility of wealth or, more generally, Stochastic Discount Factors (SDFs). We establish a duality between minimum dispersion SDFs and penalized portfolio selection problems, building the foundation for characteriz- ing the feasible tradeoffs between a SDF's pricing accuracy and its comovement with systematic risks. Empirically, a minimum variance CAPM–SDF produces a Pareto optimal tradeoff. This SDF only depends on two distinct risk factors: A traded mar- ket factor and a minimum variance excess return that bounds the mispricing of risks unspanned by market shocks.

Aknowledgements

First and foremost, I am grateful to Fabio Trojani for the invaluable guidance he offered me throughout the entirety of my PhD. I have much enjoyed and benefited from researching and teaching under his supervision. I regard him as one of the most creative, competent and dedicated professors I had the honor to get to know, and I hope to have assimilated some of his qualities.

I thank my friend, colleague and co-author Sofonias Korsaye, who has shared with me all the joys and the struggles of the PhD life.

I am also grateful to Elvezio Ronchetti, Olivier Scaillet and Patrick Gagliardini for their precious comments that deeply improved my thesis and for the terrific lectures I had the pleasure to attend during my studies, that inspired me to pursue a career in research.

I am especially indebted to Elvezio Ronchetti for all the support he gave me from the beginning of my PhD, and for helping me find the right path forward.

Finally, I thank my colleagues Nicola Gnecco, Andrea Maino, Cesare Miglioli and Alban Moor for the useful discussions related to research, and more importantly for their friendship.

Declaration

I confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged.

Alberto Quaini

Part 1: Penalized Estimators

A UNIFYING CONVEX ANALYSIS FRAMEWORK FOR PENALIZED LEAST SQUARES

ALBERTO QUAINI and FABIO TROJANI*

First version: March 2021. This version: December 8, 2021

Abstract

We introduce a unifying convex analysis framework for studying the statistical properties of Penalized Least Squares Estimators (PLSEs) with convex penalties, which is applicable both under a regular and a singular design. Our approach borrows from a reinterpretation of PLSEs as proximity operators and Moreau's decomposition of these operators. This allows for a general characterization of the asymptotic properties of PLSEs, which only depends on suitable functional transformations of the PLSE limit penalty. Exploiting our approach, we propose convenient Oracle PLSEs for singular designs exhibiting the grouping effect, and valid bootstrap approximations for the associated asymptotic distributions.

Keywords: Penalized Least Squares, asymptotics, Oracle property, singular design, bootstrap.

*Alberto Quaini (email: Alberto.Quaini@unige.ch) is with the University of Geneva, Geneva Finance Research Institute. Fabio Trojani (email: Fabio.Trojani@unige.ch) is with the University of Geneva, the University of Turin and the Swiss Finance Institute. For useful comments, we thank Elvezio Ronchetti, Olivier Scaillet, Patrick Gagliardini, Marco Avella Medina, Sofonias Alemu Korsaye and Lorenzo Camponovo. All errors are ours.

1 Introduction

This paper introduces a unifying convex analysis framework for studying the statistical properties of Penalized Least Squares Estimators (PLSEs) with convex penalties, which is applicable both under a regular and a singular design. PLSEs aim at estimating a parameter of interest $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ in a linear regression model of the form:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}\,, \tag{1.1}$$

where random variables $\boldsymbol{X}$ and $\boldsymbol{\varepsilon}$ take values in $\mathbb{R}^{n\times p}$ and $\mathbb{R}^n$, respectively, and are defined on probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The well-known Least Squares Estimator (LSE) of parameter $\boldsymbol{\beta}_0$ is defined by:

$$\hat{\boldsymbol{\beta}}_n^{ls} := \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \left\{ \frac{1}{2n} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 \right\}\,, \tag{1.2}$$

where $\|\cdot\|_2$ is the Euclidean norm in $\mathbb{R}^n$. For a regular design, matrix $\boldsymbol{X}'\boldsymbol{X}$ is $\mathbb{P}-$almost surely positive definite and this estimator is the best unbiased linear estimator of parameter $\boldsymbol{\beta}_0$, under standard assumptions on linear regression model (1.1). Under a singular design, the LSE is not uniquely defined, because a multiplicity of solutions to LS problem (1.2) exists. PLSEs aim to improve on the bias-variance tradeoff implied by LSEs, while additionally enabling variable selection, when this is desirable and feasible. They are defined by the solution of a penalized Least Squares problem of the form:

$$\hat{\boldsymbol{\beta}}_n := \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \left\{ \frac{1}{2n} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda_n f_n(\boldsymbol{\beta}) \right\}, \tag{1.3}$$

for some (possibly stochastic) penalty function f_n and a penalty parameter $\lambda_n > 0$. Examples of estimators of the form (1.3) based on a non-stochastic penalty $\lambda_n f_n$ are, e.g., the Ridge (Hoerl and Kennard (1970)), the Lasso (Tibshirani (1996)), the Elastic Net (Zou and Hastie (2005)) and the Group Lasso (Yuan and Lin (2006)). Examples of penalized estimators based on a stochastic penalty are the Adaptive Lasso (Zou (2006)) and the Adaptive Elastic Net (Zou and Zhang (2009)). Under a regular design, PLSEs are well-defined for a wide class of penalties f_n. Under a singular design, they are well-defined only for penalties ensuring existence of a unique solution to optimization problem (1.3).

Our framework for studying PLSEs of the form (1.3) starts from a generally valid reinterpretation of PLSEs as proximal operators evaluated at a linear estimator. We develop our theory for a broad class of convex penalties in the family $\Gamma(\mathbb{R}^p)$ of convex, lower-semicontinuous and proper functions, defined on

$\mathbb{R}^p$ and taking values in $(-\infty, \infty]$.[1] In this setting, we obtain general characterizations of the asymptotic properties of PLSEs, such as their asymptotic distributions and Oracle properties under local alternatives, based exclusively on suitable functional transformations of the PLSE limit penalty. Among these transformations, proximal operators and convex conjugation play a prominent role. For functions $f \in \Gamma(\mathbb{R}^p)$ and an inner product $\langle \cdot, \cdot \rangle$ on $\mathbb{R}^p$ with associated norm $\|\cdot\| := \sqrt{\langle \cdot, \cdot \rangle}$, the convex conjugate $f^* \in \Gamma(\mathbb{R}^p)$ and the proximity (or proximal) operator $\text{prox}_f : \mathbb{R}^p \to \mathbb{R}^p$, are defined by:

$$f^*(\boldsymbol{\theta}) := \sup_{\boldsymbol{\beta} \in \mathbb{R}^p} \{\langle \boldsymbol{\theta}, \boldsymbol{\beta} \rangle - f(\boldsymbol{\beta})\} \quad ; \quad \text{prox}_f(\boldsymbol{\theta}) := \underset{\beta \in \mathbb{R}^p}{\text{argmin}} \left\{ \frac{1}{2} \|\boldsymbol{\theta} - \boldsymbol{\beta}\|^2 + f(\boldsymbol{\beta}) \right\} .$$

Section 2 shows that PLSEs with penalty function $f_n \in \Gamma(\mathbb{R}^p)$ are proximal operators induced by a corresponding inner product. Equivalently, they are given by a conjugate operator involving the proximity operator of the convex conjugate of penalty $\lambda_n f_n$, which for well-known PLSEs, such as the Lasso, Adaptive Lasso or Adaptive Elastic Net, is a projection operator on a particular polyhedron.

Section 3 builds on the properties of functions $f_n \in \Gamma(\mathbb{R}^p)$ with respect to (i) epiconvergence in probability and in distribution (Salinetti and Wets (1981) and Salinetti and Wets (1986)) and (ii) convex conjugation, in order to characterize the asymptotic properties of PLSEs under a fixed dimension of the parameter space. Under fairly general high-level assumptions, we obtain two conjugate expressions for the asymptotic distribution of PLSEs.[2] They are given by appropriate limit operators evaluated at a Gaussian random vector reproducing the asymptotic distribution of the underlying linear estimator. The primal limit operator is the proximity operator of the directional derivative of the PLSE limit penalty at the parameter of interest. The conjugate limit operator is the residual of a projection on the subgradient of the PLSE limit penalty at the parameter of interest. We next characterize necessary and sufficient conditions for a PLSE Oracle property to hold, which are formulated exclusively in terms of the subgradients of the PLSEs' defining penalties. Since the sufficient conditions are independent of the inner product in the proximal operator underlying a particular PLSE, they ensure the Oracle property only in terms of functional features of the PLSE penalties. We further exploit Moreau (1962) decomposition of proximity operators, which uniquely links a PLSE to the underlying linear estimator via its conjugate operator, to obtain Oracle asymptotic distributions implying a nontrivial power for testing local hypotheses that depend on both active and inactive components of the

[1] We allow penalty function f_n in equation (1.3) to be extended real-valued, in order to (i) naturally accommodate adaptive penalties that may converge to an extended real-valued limit penalty as $n \to \infty$ and (ii) embed penalties modelling convex parameter constraints with the characteristic function of a set in constrained Least Squares estimation problems (Liew (1976)).

[2] In addition to standard assumptions ensuring valid LLNs and CLTs for the underlying data, these asymptotic representations only require the existence of a limit in epigraph to the sequence of PLSEs' defining penalties.

underlying parameter of interest.

While our theory naturally covers under a unifying framework existing statistical characterizations of PLSEs for regular designs, such as the Lasso, the Adaptive Lasso and the Adaptive Elastic Net, it is naturally applicable to singular designs as well. The construction of well-behaved PLSEs for singular designs is an open problem in the literature and our framework naturally identifies the key challenges that need to be overcome to solve the problem: the specification of a well-behaved LSE for singular designs and the definition of an appropriate inner product, inducing a well-behaved proximal operator for singular designs. In Section 4, we design convenient Oracle PLSEs for singular designs exhibiting the grouping effect. Such Oracle PLSEs arise from suitable proximal operators applied to a minimum Euclidean norm LSE and are defined with appropriate modifications of adaptive penalties in the literature.

Section 5 borrows from our earlier asymptotic results to specify a transparent set of high-level conditions that ensure validity of feasible bootstrap approximations for a PLSE's asymptotic distribution under regular and singular designs. These conditions require (i) existence of a valid bootstrap approximation for the asymptotic distribution of the underlying LSE and (ii) existence of a consistent bootstrap estimator for the directional derivative of the PLSE's limit penalty at the parameter of interest. We systematically characterize valid boostrap approximations under such conditions, based on suitable bootstrap penalties converging in epigraph to the target limit. Finally, all proofs of results in the paper and additional auxiliary findings are collected in the Online Appendix.

2 A convex analysis framework for PLSEs

For vectors $\boldsymbol{\beta}, \boldsymbol{\theta} \in \mathbb{R}^p$, let $\langle \boldsymbol{\beta}, \boldsymbol{\theta} \rangle_{\boldsymbol{M}} := \langle \boldsymbol{\beta}, \boldsymbol{M}\boldsymbol{\theta} \rangle$ be their inner product induced by a symmetric positive definite matrix $\boldsymbol{M} \in \mathbb{R}^{p \times p}$. The associated inner product norm is denoted by $\|\cdot\|_{\boldsymbol{M}}$. Starting point of this work, formalized in the next proposition, is the observation that PLSEs (1.3) with penalties $f_n \in \Gamma(\mathbb{R}^p)$ satisfy two equivalent representations summarized by two corresponding proximal operators evaluated at the LSE.

Proposition 1. *Let matrix $\boldsymbol{Q}_n := \boldsymbol{X}'\boldsymbol{X}/n$ be positive definite and $f_n \in \Gamma(\mathbb{R}^p)$, $\mathbb{P}-$almost surely. Then, PLSE (1.3) is given by following proximal operator:*

$$\hat{\boldsymbol{\beta}}_n = \operatorname{prox}^{\boldsymbol{Q}_n}_{\lambda_n f_n}(\hat{\boldsymbol{\beta}}^{ls}_n) := \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \frac{1}{2} \left\| \hat{\boldsymbol{\beta}}^{ls}_n - \boldsymbol{\beta} \right\|^2_{\boldsymbol{Q}_n} + \lambda_n f_n(\boldsymbol{\beta}) \right\}. \tag{2.1}$$

Equivalently,

$$\hat{\boldsymbol{\beta}}_n = \left(Id - \text{prox}^{\boldsymbol{Q}_n}_{(\lambda_n f_n)^*} \right) (\hat{\boldsymbol{\beta}}^{ls}_n) \, , \tag{2.2}$$

with the proximal operator:

$$\text{prox}^{\boldsymbol{Q}_n}_{(\lambda_n f_n)^*}(\hat{\boldsymbol{\beta}}^{ls}_n) = \underset{\boldsymbol{\theta} \in \mathbb{R}^p}{\text{argmin}} \left\{ \frac{1}{2} \left\| \hat{\boldsymbol{\beta}}^{ls}_n - \boldsymbol{\theta} \right\|^2_{\boldsymbol{Q}_n} + \lambda_n f^*_n(\boldsymbol{\theta}/\lambda_n) \right\} \, , \tag{2.3}$$

where f^*_n is the convex conjugate of f_n under inner product $\langle \cdot, \cdot \rangle_{\boldsymbol{Q}_n}$.

In Proposition 1, PLSE (1.3) is the proximity operator of penalty $\lambda_n f_n$ evaluated at $\hat{\boldsymbol{\beta}}^{ls}_n$, under inner product $\langle \cdot, \cdot \rangle_{\boldsymbol{Q}_n}$. From Moreau's decomposition $Id = \text{prox}_f + \text{prox}_{f^*}$ (Bauschke et al., 2016, Thm. 14.3), this estimator is equivalently given by the difference between the LSE and the proximity operator of conjugate penalty $(\lambda_n f_n)^* = \lambda_n f^*_n(\cdot/\lambda_n)$ evaluated at the LSE.

Remark 1. A useful property of PLSE (2.1) and its conjugate estimator (2.3), which follows directly from their interpretations as proximity operators, is that they are Lipschitz continuous, hence Lebesgue almost everywhere differentiable functions on $\mathbb{R}^p$ (Bauschke et al., 2016, Prop. 12.29). Moreover, both these proximal operators can be interpreted as penalized Asymptotic Least Squares estimators, in which matrix $\boldsymbol{Q}_n$ acts as a weighting matrix in the Least Squares criterion. The optimal choice of such weighting matrix directly depends on the asymptotic properties of LSE $\hat{\boldsymbol{\beta}}^{ls}_n$, and choice $\boldsymbol{Q}_n$ is optimal under the standard assumption of homoskedasticity in lineal model (1.1). More generally, the optimal weighting matrix in such penalized estimators is given by a consistent estimator of the inverse asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}^{ls}_n$, i.e., the optimal choice of the inner product in proximal operators (2.1) and (2.3) is intrinsically related to the asymptotic distribution of the LSE itself.

Convex conjugate $\lambda_n f^*_n(\cdot/\lambda_n)$ in equation (2.3) is defined under inner product $\langle \cdot, \cdot \rangle_{\boldsymbol{Q}_n}$. Equivalently, one can use the penalty $\lambda_n f^*_n(\boldsymbol{Q}_n \cdot /\lambda_n)$, with convex conjugation defined under the Euclidean inner product. Table 2.1 collects various well-known examples of penalty functions f_n and corresponding conjugate penalties f^*_n computed under the Euclidean product. Such penalties may depend on (i) l_r-norms, defined by $\|\boldsymbol{\beta}\|_r := \left(\sum_{j=1}^p |\beta_j|^r \right)^{1/r}$ for $r \geq 1$ and by $\|\boldsymbol{\beta}\|_r := \max\{|\beta_j| \ : \ j = 1, \ldots, p\}$ for $r = \infty$, (ii) a $n^\gamma-$consistent estimator $\tilde{\boldsymbol{\beta}}_n$ ($\gamma > 0$) of $\boldsymbol{\beta}_0$ and (iii) an indicator function of a closed convex set S, defined by $\iota_S(\boldsymbol{\theta}) = 0$ for $\boldsymbol{\theta} \in S$ and $\iota_S(\boldsymbol{\theta}) = \infty$ else. In the Group lasso, the parameter of interest is partitioned as

Penalty	$f_n(\boldsymbol{\beta})$	$\lambda_n f_n^*(\boldsymbol{\theta}/\lambda_n)$
Ridge	$\frac{1}{2}\|\boldsymbol{\beta}\|_2^2$	$\frac{1}{2\lambda_n}\|\boldsymbol{\theta}\|_2^2$
Lasso	$\|\boldsymbol{\beta}\|_1$	$\iota_{B_n}(\boldsymbol{\theta}),\ \ B_n := \bigcap_{i=1}^p \{\boldsymbol{\theta} \ : \ \|\theta_j\| \leq \lambda_n\}$
Adaptive Lasso	$\sum_{j=1}^p \|\beta_j\|/\|\tilde{\beta}_{nj}\|$	$\iota_{\tilde{B}_n}(\boldsymbol{\theta}),\ \ \tilde{B}_n := \bigcap_{i=1}^p \left\{\boldsymbol{\theta} \ : \ \|\theta_j\| \leq \lambda_n/\|\tilde{\beta}_{nj}\|\right\}$
Group Lasso	$\sum_{k=1}^K \|\boldsymbol{\beta}_k\|_2$	$\iota_{G_n}(\boldsymbol{\theta}),\ \ G_n := \bigcap_{k=1}^K \{\boldsymbol{\theta} \ : \ \|\boldsymbol{\theta}_k\|_2 \leq \lambda_n\}$
Naïve Elastic Net	$w\|\boldsymbol{\beta}\|_1 + \frac{1-w}{2}\|\boldsymbol{\beta}\|_2^2,\ \ w \in (0,1)$	$\frac{1}{2\lambda_n(1-w)} d^2_{w\lambda_n \mathcal{B}_\infty}(\boldsymbol{\theta})$
Constrained LS	$\iota_C(\boldsymbol{\beta})$	$\sigma_C(\boldsymbol{\theta})$

Table 2.1: **Penalty functions and corresponding convex conjugates under Euclidean inner product** $\langle \cdot, \cdot \rangle$

$\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_K')'$, using corresponding subvectors $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$ of varying dimension. In the constrained Least Squares setting, $C \subset \mathbb{R}^p$ is a nonempty closed convex set defining a family of convex constraints on parameter $\boldsymbol{\theta}$. Conjugate penalties f_n^* may depend on (i) dual norms of l_r-norms, (ii) the Euclidean distance from a closed convex set S, defined by $d_S(\boldsymbol{\theta}) := \inf_{\boldsymbol{\beta} \in S} \|\boldsymbol{\theta} - \boldsymbol{\beta}\|_2$, and (iii) the support function of a set $S \subset \mathbb{R}^p$, defined as $\sigma_S(\boldsymbol{\theta}) := \sup_{\boldsymbol{\beta} \in S}\{\langle \boldsymbol{\theta}, \boldsymbol{\beta} \rangle\}$.

From conjugate characterization (2.2), we obtain following more explicit expressions for various PLSEs with penalties reported in Table 2.1.

Proposition 2. *(i) For the Lasso penalty, PLSE (2.1) satisfies the projection formula:*

$$\hat{\boldsymbol{\beta}}_n = \left(Id - P_{C_n}^{\boldsymbol{Q}_n}\right)(\hat{\boldsymbol{\beta}}_n^{ls}), \tag{2.4}$$

with the projection operator:

$$P_{C_n}^{\boldsymbol{Q}_n}(\hat{\boldsymbol{\beta}}_n^{ls}) := \underset{\boldsymbol{\theta} \in C_n}{\operatorname{argmin}} \left\{ \frac{1}{2} \left\| \hat{\boldsymbol{\beta}}_n^{ls} - \boldsymbol{\theta} \right\|_{\boldsymbol{Q}_n}^2 \right\},$$

and the polyhedron:

$$C_n := C_n(\lambda_n) := \bigcap_{j=1}^p \{\boldsymbol{\theta} \ : \ |\langle \boldsymbol{e}_j, \boldsymbol{\theta} \rangle_{\boldsymbol{Q}_n}| \leq \lambda_n\}, \tag{2.5}$$

where $\boldsymbol{e}_j$ denotes the $j-$th unit vector in $\mathbb{R}^p$.

(ii) For the Adaptive Lasso penalty, the PLSE satisfies projection formula (2.4) with following polyhedron:

$$C_n := C_n(\lambda_n, \tilde{\boldsymbol{\beta}}_n) := \bigcap_{j=1}^{p} \{\boldsymbol{\theta} \ : \ |\langle \boldsymbol{e}_j, \boldsymbol{\theta} \rangle_{\boldsymbol{Q}_n}| \leq \lambda_n / |\tilde{\beta}_{nj}|\} \ . \tag{2.6}$$

(iii) For the naïve Elastic Net penalty, the PLSE satisfies the projection formula:

$$\hat{\boldsymbol{\beta}}_n = \left(Id - P^{\boldsymbol{Q}_n(\lambda_{2n})}_{C_n(\lambda_n, w)} \right) (\hat{\boldsymbol{\beta}}^{ls}_n(\lambda_{2n})) \ , \tag{2.7}$$

with $\lambda_{1n} := \lambda_n w$, $\lambda_{2n} := \lambda_n (1-w)/2$, linear estimator:

$$\hat{\boldsymbol{\beta}}^{ls}_n(\lambda_{2n}) := [\boldsymbol{Q}_n(\lambda_{2n})]^{-1} \left(\frac{\boldsymbol{X}'\boldsymbol{Y}}{n} \right) := [\lambda_{2n} \boldsymbol{I}_n + \boldsymbol{Q}_n]^{-1} \left(\frac{\boldsymbol{X}'\boldsymbol{Y}}{n} \right) \ , \tag{2.8}$$

and polyhedron:

$$C_n(\lambda_n, w) := \bigcap_{j=1}^{p} \{\boldsymbol{\theta} : |\langle \boldsymbol{e}_j, \boldsymbol{\theta} \rangle_{\boldsymbol{Q}_n(\lambda_{2n})}| \leq \lambda_{1n}\} \ . \tag{2.9}$$

Proposition 2 clarifies the tight connection between well-known PLSEs and projection operators on corresponding polyhedrons. The Lasso and the Adaptive Lasso estimators are given by the residual of a projection of LSE $\hat{\boldsymbol{\beta}}^{ls}_n$ on polyhedron (2.5) and (2.6), respectively. In orthonormal designs ($\boldsymbol{Q}_n = \boldsymbol{I}_n$), these projection formulas have the well-known soft-thresholding form, as they are the residual of an Euclidean projection of LSE $\hat{\boldsymbol{\beta}}^{ls}_n$ on the $l_\infty-$ball of radius λ_n and the scaled $l_\infty-$ball of length $2\lambda_n/|\tilde{\beta}_{nj}|$ along the $j-$th coordinate, respectively.

The naïve Elastic Net estimator is given by the projection residual on polyhedron (2.9) of Ridge estimator (2.8), which is induced by modified design matrix $\boldsymbol{Q}(\lambda_{2n})$. Under an orthonormal design, these projection formulas have a well-known soft-thresholding form, in which the naïve Elastic Net estimator is reproduced by a Lasso estimator with penalization parameter λ_{1n}, which is shrank toward zero by shrinking coefficient $1/(1+\lambda_{2n})$. The naïve Elastic Net estimator is known to produce an excessive amount of shrinkage, which is why Zou and Hastie (2005) propose to rescale it by scaling factor $1+\lambda_{2n}$, in order to obtain a more accurate Elastic Net estimator with respect to the associated bias-variance trade-off. Such Elastic Net estimator satisfies as well a projection formula of the form (2.7), but with an inner

product $\langle \cdot, \cdot \rangle_{\check{\boldsymbol{Q}}_n(\lambda_{2n})}$ induced by rescaled matrix $\check{\boldsymbol{Q}}_n(\lambda_{2n}) := \frac{1}{1+\lambda_{2n}} \boldsymbol{Q}_n(\lambda_{2n})$ and rescaled Ridge estimator $\check{\boldsymbol{\beta}}_n^{ls}(\lambda_{2n}) := [\check{\boldsymbol{Q}}_n(\lambda_{2n})]^{-1}(\boldsymbol{X}'\boldsymbol{Y}/n)$. Therefore, it also directly follows from the proof of Proposition 2 that an Adaptive Elastic Net estimator, which is defined with proximal operator $\hat{\boldsymbol{\beta}}_n = \text{prox}_{\lambda_{1n} f_n}^{\check{\boldsymbol{Q}}_n(\lambda_{2n})}(\check{\boldsymbol{\beta}}_n^{ls}(\lambda_{2n}))$ using an Adaptive Lasso penalty f_n, satisfies a corresponding projection formula given by:

$$\hat{\boldsymbol{\beta}}_n = \left(Id - P_{C_n(\lambda_n, w, \tilde{\boldsymbol{\beta}}_n)}^{\check{\boldsymbol{Q}}_n(\lambda_{2n})} \right) (\check{\boldsymbol{\beta}}_n^{ls}(\lambda_{2n})) \,, \tag{2.10}$$

for a corresponding closed-form polyhedron:

$$C_n(\lambda_n, w, \tilde{\boldsymbol{\beta}}_n) := \bigcap_{j=1}^{p} \{ \boldsymbol{\theta} \,:\, |\langle \boldsymbol{e}_j, \boldsymbol{\theta} \rangle_{\check{\boldsymbol{Q}}_n(\lambda_{2n})}| \leq \lambda_n w / |\tilde{\beta}_{nj}| \} \,. \tag{2.11}$$

More generally, out theory implies that PLSE's dual characterization (2.2) admits a characterization with a projection formula if and only if penalty $f_n \in \Gamma(\mathbb{R}^p)$ is sublinear, since in this case there always exists a nonempty closed and convex set C such that $f_n = \sigma_C$ (Hiriart-Urruty and Lemaréchal, 2004, Thm. 3.1.1) and $f_n^* = \iota_C$.

3 Asymptotic properties of PLSEs

The framework in Proposition 1 enables a systematic functional characterization of the asymptotic properties of PLSEs with penalties in $\Gamma(\mathbb{R}^p)$, based on the limit behavior of proximal operator (2.1) or, equivalently, proximal operator (2.3). As these operators are defined via the minimizers of coercive objective functions in class $\Gamma(\mathbb{R}^p)$, the appropriate notion of convergence implying a well-defined limit proximal operator is the notion of epigraph convergence (or epiconvergence); (Rockafellar and Wets, 2009, Thm. 7.33). In particular, in order to obtain well-behaved stochastic limit proximal operators we make use of the notions of epigraph convergence in probability or in distribution; see, e.g., Salinetti and Wets (1981), Salinetti and Wets (1986), Geyer (1994) and Knight (1999), among many others.

3.1 Main assumptions and basic asymptotic properties

We adopt two sets of high-level assumptions ensuring epiconvergence of the objective functions defining proximal operators (2.1) or (2.3) to a strictly convex coercive objective function in class $\Gamma(\mathbb{R}^p)$. Our first high-level assumption assumes validity of suitable Laws of Large Numbers and Central Limit Theorems

for the variables appearing in linear model (1.1), which is a weak assumption typically needed to ensure consistency and asymptotic normality of LSEs.

Assumption 1. *Denote by $\boldsymbol{X}_i'$ the $i-$th row of matrix $\boldsymbol{X}$. $\{(\boldsymbol{X}_i, \varepsilon_i) : i \in \mathbb{N}\}$ is a second-order stationary stochastic process such that $\mathbb{E}[\boldsymbol{X}_1 \varepsilon_1] = \mathbf{0}$ and satisfying following properties:*

(i) $\boldsymbol{Q}_n$ is $\mathbb{P}-$almost surely positive definite for each $n \in \mathbb{N}$ and $\boldsymbol{Q}_n \to_{\Pr} \boldsymbol{Q}_0 := \mathbb{E}[\boldsymbol{X}_1 \boldsymbol{X}_1']$, where matrix $\boldsymbol{Q}_0$ is positive definite;

(ii) $\boldsymbol{X}'\boldsymbol{\varepsilon}/n \to_{\Pr} \mathbf{0}$;

(iii) $\boldsymbol{X}'\boldsymbol{\varepsilon}/\sqrt{n} \to_d \boldsymbol{Z}$, where $\boldsymbol{Z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}_0)$ for some positive definite matrix $\boldsymbol{\Omega}_0$.

Assumption 1, (i) and (ii) implies following uniform convergence on compact sets:

$$\left\| \hat{\boldsymbol{\beta}}_n^{ls} - \cdot \right\|_{\boldsymbol{Q}_n}^2 \to_{\Pr} \|\boldsymbol{\beta}_0 - \cdot\|_{\boldsymbol{Q}_0}^2 \ .$$

This property is needed to obtain consistency of a PLSE. Similarly, Assumption 1, (i) and (iii) implies following uniform convergence on compact sets:

$$\left\| \sqrt{n}(\hat{\boldsymbol{\beta}}_n^{ls} - \boldsymbol{\beta}_0) - \cdot \right\|_{\boldsymbol{Q}_n}^2 \to_d \left\| \boldsymbol{Q}_0^{-1}\boldsymbol{Z} - \cdot \right\|_{\boldsymbol{Q}_0}^2 \ .$$

This property is needed to obtain the asymptotic distribution of a PLSE.

Our second high-level assumption concerns the properties of penalties f_n and f_n^* in the definition of proximal operators (2.1) and (2.3). Here, what is needed is the epigraph convergence in probability to a limit penalty f_0 and f_0^*, respectively. However, since epigraph convergence is preserved under convex conjugation (Mosco, 1971, Thm. 1) and $f = f^{**}$ for any $f \in \Gamma(\mathbb{R}^p)$ (Bauschke et al., 2016, Cor. 13.38), it is sufficient to assume epigraph convergence in probability of penalties f_n, which is a weak assumption satisfied, e.g., by all penalties in Table 2.1. Finally, in order to ensure that parameter of interest $\boldsymbol{\beta}_0$ is the unique solution of the emerging limit optimization problem of a PLSE, we require it to be in the domain of the associated penalties.

Assumption 2. *Sequence of penalties $\{f_n : n \in \mathbb{N}\} \subset \Gamma(\mathbb{R}^p)$ satisfies following properties:*

(i) There exists proper function $f_0 : \mathbb{R}^p \to (-\infty, \infty]$ such that $f_n \to_{\Pr} f_0$ in epigraph;

(ii) $\boldsymbol{\beta}_0 \in \bigcap_{n\in\mathbb{N}}(\mathrm{dom}(f_0) \cap \mathrm{dom}(f_n))$.

The next straightforward proposition establishes the limit in probability of proximal operators (2.1) and (2.3) under the standard assumption of a converging penalty parameter λ_n. As intuitively expected, this limit is itself a proximal operator evaluated at $\boldsymbol{\beta}_0$, which only depends on scalar product $\langle\cdot,\cdot\rangle_{\boldsymbol{Q}_0}$, limit penalties f_0, f_0^* and limit penalty parameter λ_0.

Proposition 3. *Let Assumptions 1(i), 1(ii) and 2 be satisfied.*

(i) If $\lambda_n \to \lambda_0 > 0$*, then:*

$$\begin{pmatrix} \mathrm{prox}^{\boldsymbol{Q}_n}_{\lambda_n f_n}(\boldsymbol{\beta}_n^{ls}) \\ \mathrm{prox}^{\boldsymbol{Q}_n}_{(\lambda_n f_n)^*}(\boldsymbol{\beta}_n^{ls}) \end{pmatrix} \to_{\mathrm{Pr}} \begin{pmatrix} \mathrm{prox}^{\boldsymbol{Q}_0}_{\lambda_0 f_0}(\boldsymbol{\beta}_0) \\ \mathrm{prox}^{\boldsymbol{Q}_0}_{(\lambda_0 f_0)^*}(\boldsymbol{\beta}_0) \end{pmatrix}, \tag{3.1}$$

with convex conjugate f_0^* *under inner product* $\langle\cdot,\cdot\rangle_{\boldsymbol{Q}_0}$*. Moreover,*

$$\mathrm{prox}^{\boldsymbol{Q}_0}_{\lambda_0 f_0}(\boldsymbol{\beta}_0) = \left(Id - \mathrm{prox}^{\boldsymbol{Q}_0}_{(\lambda_0 f_0)^*}\right)(\boldsymbol{\beta}_0)\,.$$

(ii) If $\lambda_n \to \lambda_0 = 0$ *and* $\lambda_n f_n \to_{\mathrm{Pr}} \iota_{\mathrm{dom}(f_0)}$ *in epigraph, then:*

$$\begin{pmatrix} \mathrm{prox}^{\boldsymbol{Q}_n}_{\lambda_n f_n}(\boldsymbol{\beta}_n^{ls}) \\ \mathrm{prox}^{\boldsymbol{Q}_n}_{(\lambda_n f_n)^*}(\boldsymbol{\beta}_n^{ls}) \end{pmatrix} \to_{\mathrm{Pr}} \begin{pmatrix} \boldsymbol{\beta}_0 \\ \mathbf{0} \end{pmatrix}. \tag{3.2}$$

In Proposition 3, consistency of PLSE (2.1) directly follows under asymptotic regime (ii). Instead, when $\lambda_0 > 0$, $-\mathrm{prox}^{\boldsymbol{Q}_0}_{(\lambda_0 f_0)^*}(\boldsymbol{\beta}_0)$ reproduces the asymptotic bias of a PLSE, which is consistently estimated directly with proximal estimator (2.3). The asymptotic bias depends on the underlying probability $\mathbb{P}$ only via the dependence of matrix $\boldsymbol{Q}_0$ and parameter of interest $\boldsymbol{\beta}_0 = \boldsymbol{Q}_0^{-1}\boldsymbol{\delta}_0 := \boldsymbol{Q}_0^{-1}\mathbb{E}[\boldsymbol{X}_1 Y_1]$ on $\mathbb{P}$. Therefore, $\mathrm{prox}^{\boldsymbol{Q}_0}_{(\lambda_0 f_0)^*}$ also provides a direct definition of the asymptotic bias functional associated to PLSEs of the form (2.1). Obviously, when penalty f_n does not depend on n, one has $f_0 = f_n$, $f_0^* = f_n^*$. Examples of such settings include, e.g., the Ridge, Lasso, Group Lasso and Elastic Net, which all imply $\mathrm{dom}(f_0) = \mathbb{R}^p$ and $\iota_{\mathrm{dom}(f_0)} = 0$ for asymptotic regime (ii). In other cases, e.g., for several adaptive penalties, the probability limit in epigraph of $\lambda_n f_n$ or $(\lambda_n f_n)^*$ needs to be computed, but this is usually an easy task. For instance,

the Adaptive Lasso penalties imply a limit penalty given by:[3]

$$f_0(\boldsymbol{\beta}) = \sum_{j=1}^{p} \left[\frac{|\beta_j|}{|\beta_{0j}|} I(\beta_{0j} \neq 0) + \iota_{\{0\}}(\beta_j) I(\beta_{0j} = 0) \right] , \tag{3.3}$$

i.e., $\mathrm{dom}(f_0) = \cap_{\{j:\beta_{0j}=0\}} \{\boldsymbol{\beta} \in \mathbb{R}^p : \beta_j = 0\}$. Under asymptotic regime (ii) in Proposition 3 it thus follows:

$$\mathrm{prox}^{\boldsymbol{Q}_n}_{\lambda_n f_n}(\boldsymbol{\beta}^{ls}_n) \to_{\mathrm{Pr}} \mathrm{prox}^{\boldsymbol{Q}_0}_{\iota_{\mathrm{dom}(f_0)}}(\boldsymbol{\beta}_0) = P^{\boldsymbol{Q}_0}_{\mathrm{span}\{\boldsymbol{e}_j : \beta_{0j} \neq 0\}}(\boldsymbol{\beta}_0) ,$$

which shows that the Adaptive Lasso asymptotic functional is the Euclidean Projection (under inner product $\langle \cdot, \cdot \rangle_{\boldsymbol{Q}_0}$) on the subspace spanned by canonical basis vectors $\boldsymbol{e}_j$ indexed by the active (nonzero) components in parameter vector $\boldsymbol{\beta}_0$.

3.2 Asymptotic distribution

To characterize the asymptotic distribution of PLSE (2.1), we introduce the directional derivative of limit penalty f_0 at $\boldsymbol{\beta}_0$, which is defined for any $\boldsymbol{b} \in \mathbb{R}^p$ by:

$$\rho_{\boldsymbol{\beta}_0}(\boldsymbol{b}) := \lim_{\alpha \downarrow 0} \frac{f_0(\boldsymbol{\beta}_0 + \alpha \boldsymbol{b}) - f_0(\boldsymbol{\beta}_0)}{\alpha} .$$

To characterize the asymptotic distribution of conjugate estimator (2.3), we make use of the subgradient of f_0 at $\boldsymbol{\beta}_0$, which under the Euclidean inner product is defined by:

$$\partial f_0(\boldsymbol{\beta}_0) := \bigcap_{\boldsymbol{\beta} \in \mathbb{R}^p} \{\boldsymbol{t} \in \mathbb{R}^p : f_0(\boldsymbol{\beta}) - f_0(\boldsymbol{\beta}_0) - \langle \boldsymbol{\beta} - \boldsymbol{\beta}_0, \boldsymbol{t} \rangle \geq 0\} . \tag{3.4}$$

Under Assumption 2, both $\rho_{\boldsymbol{\beta}_0}(\boldsymbol{b})$ and $\partial f_0(\boldsymbol{\beta}_0)$ are well-defined, i.e., $\rho_{\boldsymbol{\beta}_0}$ exists in the extended-real line and $\partial f_0(\boldsymbol{\beta}_0)$ is nonempty (Bauschke et al., 2016, Thm. 17.2 and Remark 16.2). These two objects are intrinsically related by the identity (Bauschke et al., 2016, Prop. 17.17) $\rho_{\boldsymbol{\beta}_0} = (\iota_{\partial f_0(\boldsymbol{\beta}_0)})^* = \sigma_{\partial f_0(\boldsymbol{\beta}_0)}$, i.e., the directional derivative of f_0 at $\boldsymbol{\beta}_0$ is the support function of $\partial f_0(\boldsymbol{\beta}_0)$. Moreover, for any closed convex set C, the directional derivative at $\boldsymbol{\beta}_0$ of characteristic function $f_0 = \iota_C$ is the support function of the normal

[3] See the Online Appendix for details.

cone of set C at $\boldsymbol{\beta}_0$, which is defined by:

$$N_C(\boldsymbol{\beta}_0) := \begin{cases} \{\boldsymbol{\theta} \ : \ \sup_{\boldsymbol{z} \in C} \langle \boldsymbol{\theta}, \boldsymbol{z} - \boldsymbol{\beta}_0 \rangle \leq 0\} & \boldsymbol{\beta}_0 \in C \\ \emptyset & \text{otherwise} \end{cases} .$$

This directional derivative expression follows from (Bauschke et al., 2016, Example 16.13) and (Bauschke et al., 2016, Prop. 17.17), using the identities $\rho_{\boldsymbol{\beta}_0} = \sigma_{\partial \iota_C(\boldsymbol{\beta}_0)} = \sigma_{N_C(\boldsymbol{\beta}_0)}$. Finally, note that while in equation (3.4) subgradients are defined with respect to the Euclidean inner product, the resulting subgradient under inner product $\langle \cdot, \cdot \rangle_{\boldsymbol{Q}_0}$ is $\partial f_0(\boldsymbol{Q}_0 \cdot)$.

The asymptotic distribution functionals of PLSE (2.1) and its conjugate estimator (2.3) are characterized via the limit penalty's directional derivative and subgradient, based on corresponding proximal operators detailed in the next proposition.

Proposition 4. *Let Assumptions 1 (i), 1 (iii) and 2 be satisfied.*

(i) If $\lambda_n \sqrt{n} \to \lambda_0 > 0$, it follows:

$$\sqrt{n} \begin{pmatrix} \text{prox}^{\boldsymbol{Q}_n}_{\lambda_n f_n}(\hat{\boldsymbol{\beta}}^{ls}_n) - \boldsymbol{\beta}_0 \\ \text{prox}^{\boldsymbol{Q}_n}_{(\lambda_n f_n)^*}(\hat{\boldsymbol{\beta}}^{ls}_n) - \boldsymbol{\beta}_0 \end{pmatrix} \to_d \begin{pmatrix} \text{prox}^{\boldsymbol{Q}_0}_{\lambda_0 \rho_{\boldsymbol{\beta}_0}}(\boldsymbol{W}) \\ P^{\boldsymbol{Q}_0}_{\boldsymbol{Q}_0^{-1}(\lambda_0 \partial f_0(\boldsymbol{\beta}_0))}(\boldsymbol{W}) \end{pmatrix} , \tag{3.5}$$

with random vector $\boldsymbol{W} := \boldsymbol{Q}_0^{-1} \boldsymbol{Z}$. Moreover,

$$\text{prox}^{\boldsymbol{Q}_0}_{\lambda_0 \rho_{\boldsymbol{\beta}_0}}(\boldsymbol{W}) = \left(Id - P^{\boldsymbol{Q}_0}_{\boldsymbol{Q}_0^{-1}(\lambda_0 \partial f_0(\boldsymbol{\beta}_0))} \right)(\boldsymbol{W}) . \tag{3.6}$$

(ii) If $\lambda_n \sqrt{n} \to \lambda_0 = 0$ and $\lambda_n \sqrt{n} f_n \to_{Pr} \iota_{\text{dom}(f_0)}$ in epigraph, the above limits in distribution hold with $\lambda_0 \rho_{\beta_0}$ and $\lambda_0 \partial f_0(\boldsymbol{\beta}_0)$ replaced by $\sigma_{N_{\text{dom}(f_0)}(\boldsymbol{\beta}_0)}$ and $N_{\text{dom}(f_0)}(\boldsymbol{\beta}_0)$, respectively.

In Proposition 4, the asymptotic distribution of PLSE (2.1) is given directly by the first row of the limit in equation (3.5), as the distribution of a proximal operator with penalty $\lambda_0 \rho_{\beta_0}$ or $\sigma_{N_{\text{dom}(f_0)}(\boldsymbol{\beta}_0)}$, under asymptotic regimes (i) or (ii), respectively. Conversely, the asymptotic distribution of conjugate PLSE (2.3) is given by the second row of the limit in equation (3.5), as the distribution of a projection on the preimage under $\boldsymbol{Q}_0$ of subgradients $\lambda_0 \partial f(\boldsymbol{\beta}_0)$ or $N_{\text{dom}(f_0)}(\boldsymbol{\beta}_0)$, for asymptotic regimes (i) or (ii), respectively. These preimages are the subgradients of $\lambda_0 f_0$ or $\iota_{\text{dom}(f_0)}$ under inner product $\langle \cdot, \cdot \rangle_{\boldsymbol{Q}_0}$, respectively. All these

proximal operators are applied to a zero-mean normally distributed random vector $\boldsymbol{W} = \boldsymbol{Q}_0^{-1}\boldsymbol{Z}$ having covariance matrix $\boldsymbol{Q}_0^{-1}\boldsymbol{\Omega}_0\boldsymbol{Q}_0^{-1}$, which reproduces the asymptotic distribution of the LSE under the standard Assumption 1. Finally, the asymptotic distribution of PLSE (2.1) is also always given in equation (3.6) by a conjugate formula reproducing the residual of a projection of random vector $\boldsymbol{W}$ on the subgradient of $\lambda_0 f_0$ or $\iota_{\mathrm{dom}(f_0)}$, under asymptotic regimes (i) and (ii), with subgradients defined under inner product $\langle\cdot,\cdot\rangle_{\boldsymbol{Q}_0}$.

From asymptotic distribution (3.6), it immediately follows that a PLSE is asymptotically normally distributed if and only if $\partial f_0(\boldsymbol{\beta}_0)$ or $N_{\mathrm{dom}(f_0)}(\boldsymbol{\beta}_0)$ are affine subspaces of $\mathbb{R}^p$, under asymptotic regimes (i) and (ii), respectively. The second situation always arises for real-valued penalties that do not depend on index n.[4]

3.2.1 Asymptotic distribution of benchmark PLSEs

Asymptotic distribution (3.5) covers at once several direct asymptotic distribution characterizations in the literature, including, e.g., all PLSEs with penalties given in Table 2.1. Table 3.1 reports the directional derivatives $\rho_{\boldsymbol{\beta}_0}(\boldsymbol{b})$ for the family of penalties in Table 2.1. For the Group Lasso, the notation $b_k^{(j)}$ denotes the j^{th} component of subvector $\boldsymbol{b}_k$, where $\boldsymbol{b} = (\boldsymbol{b}_1', \ldots, \boldsymbol{b}_K')' \in \mathbb{R}^p$. For the constrained Least Squares, the directional derivative equals the support function $\sigma_{N_C(\boldsymbol{\beta}_0)}$ of the normal cone of set C at $\boldsymbol{\beta}_0$; see again (Bauschke et al., 2016, Example 16.13) and (Bauschke et al., 2016, Prop. 17.17). The directional derivative of the naïve Elastic Net penalty is not given in Table 3.1, since it is readily obtained as a convex combination with weights $w \in (0,1)$ and $1-w$ of the directional derivatives of the Lasso and the Ridge.

Conjugate asymptotic distribution characterization (3.6) is different from existing characterizations in the literature, as it relies on a projection residual fully computable from the limit penalty's subgradient. These subgradients are usually known in closed-form and for the widely used penalties in Table 2.1 they are reported in Table 3.2. The subgradient of the naïve Elastic Net penalty, which is is not reported, is given by the Cartesian product of the subgradient of the Lasso penalty scaled by $w \in (0,1)$ and the subgradient of the Ridge penalty scaled by $1-w$, respectively.

While we collect for brevity in the Online Appendix further examples of closed-form asymptotic distributions from Proposition 4 for PLSEs with penalties in Table 2.1, it is useful to investigate in more detail

[4]For instance, asymptotic normality always arises in Table 3.2 for the Ridge and the Adaptive Lasso estimators, but not for the Lasso estimator when $\lambda_0 > 0$.

Penalty $f_n(\boldsymbol{\beta})$	**Directional derivative** $\rho_{\boldsymbol{\beta}_0}(\boldsymbol{b})$
$\frac{1}{2}\|\boldsymbol{\beta}\|_2^2$	$\langle \boldsymbol{b}, \boldsymbol{\beta}_0 \rangle$
$\|\boldsymbol{\beta}\|_1$	$\sum_j \left[b_j \operatorname{sign}(\beta_{0j}) I_{\{\beta_{0j} \neq 0\}} + \|b_j\| I_{\{\beta_{0j}=0\}} \right]$
$\sum_{j=1}^p \|\beta_j\|/\|\tilde{\beta}_{nj}\|$	$\sum_j \left[\frac{b_j}{\beta_{0j}} I_{\{\beta_{0j} \neq 0\}} + \iota_{\{0\}}(b_j) I_{\{\beta_{0j}=0\}} \right]$
$\sum_{k=1}^K \|\boldsymbol{\beta}_k\|_2$	$\sum_{k=1}^K \left[\frac{\langle \boldsymbol{b}_k, \boldsymbol{\beta}_{0k} \rangle}{\|\boldsymbol{\beta}_{0k}\|_2} I_{\{\boldsymbol{\beta}_{0k} \neq \mathbf{0}\}} + \|\boldsymbol{b}_k\|_2 I_{\{\boldsymbol{\beta}_{0k}=\mathbf{0}\}} \right]$
$\iota_C(\boldsymbol{\beta})$	$\sigma_{N_C(\boldsymbol{\beta}_0)}(\boldsymbol{b})$

Table 3.1: **Penalty functions and directional derivatives** $\rho_{\boldsymbol{\beta}_0}(\boldsymbol{b})$

the case of the Adaptive Lasso limit penalty in equation (3.3), which yields:

$$\operatorname{dom}(f_0) = \bigcap_{\{j:\beta_{0j}=0\}} \{\boldsymbol{\beta} : \beta_j = 0\} = \operatorname{span}\{\boldsymbol{e}_j : \beta_{0j} \neq 0\} , \tag{3.7}$$

and, using (Bauschke et al., 2016, Prop. 6.22):

$$N_{\operatorname{dom}(f_0)}(\beta_0) = \bigcap_{\{j:\beta_{0j}\neq 0\}} \{\boldsymbol{\theta} : \theta_j = 0\} .$$

The asymptotic distributions of the Adaptive Lasso and its conjugate estimator then directly follow in closed-form, under asymptotic regime (ii) in Proposition 4, since:[5]

$$\left(Id - P^{\boldsymbol{Q}_0}_{\boldsymbol{Q}_0^{-1}(N_{\operatorname{dom}(f_0)}(\beta_0))} \right)(\boldsymbol{W}) = \left(Id - P^{\boldsymbol{Q}_0}_{span\{\boldsymbol{e}_j:\beta_{0j}\neq 0\}^\perp} \right)(\boldsymbol{W}) = P^{\boldsymbol{Q}_0}_{span\{\boldsymbol{e}_j:\beta_{0j}\neq 0\}}(\boldsymbol{W}) .$$

A direct computation of projection $P^{\boldsymbol{Q}_0}_{\operatorname{span}\{\boldsymbol{e}_j:\beta_{0j}\neq 0\}}(\boldsymbol{W})$ finally yields the explicit asymptotic distribution below, denoting by $\mathcal{A} := \{j : \beta_{0j} \neq 0\}$ the active index set of parameter vector $\boldsymbol{\beta}_0$:

$$\sqrt{n}(\operatorname{prox}^{\boldsymbol{Q}_n}_{\lambda_n f_n}(\hat{\boldsymbol{\beta}}^{ls}_n) - \boldsymbol{\beta}_0)_{\mathcal{A}} \quad \to_d \quad [(\boldsymbol{Q}_0)_{\mathcal{A}}]^{-1}(\boldsymbol{Z})_{\mathcal{A}} , \tag{3.8}$$

$$\sqrt{n}(\operatorname{prox}^{\boldsymbol{Q}_n}_{\lambda_n f_n}(\hat{\boldsymbol{\beta}}^{ls}_n))_{\mathcal{A}^c} \quad \to_d \quad \mathbf{0} , \tag{3.9}$$

[5] Orthogonal complements are defined here under inner product $\langle \cdot, \cdot \rangle_{\boldsymbol{Q}_0}$.

Penalty $f_n(\boldsymbol{\beta})$	**Subgradient** $\partial f_0(\boldsymbol{\beta}_0)$
$\frac{1}{2}\|\boldsymbol{\beta}\|_2^2$	$\{\boldsymbol{\beta}_0\}$
$\|\boldsymbol{\beta}\|_1$	$\left(\bigcap_{\{j:\beta_{0j}\neq 0\}}\{\boldsymbol{t} \,:\, t_j = \operatorname{sign}(\beta_{0j})\}\right) \bigcap \left(\bigcap_{\{j:\beta_{0j}=0\}}\{\boldsymbol{t} \,:\, t_j \in [-1,1]\}\right)$
$\sum_{j=1}^p \lvert\beta_j\rvert/\lvert\tilde{\beta}_{nj}\rvert$	$\bigcap_{\{j:\beta_{0j}\neq 0\}}\{\boldsymbol{t} \,:\, t_i = \operatorname{sign}(\beta_{0j})\}$
$\sum_{k=1}^K \|\boldsymbol{\beta}_k\|_2$	$\left(\bigcap_{\{k:\boldsymbol{\beta}_{0k}\neq \mathbf{0}\}}\{\boldsymbol{t} \,:\, t_k^{(j)} = \operatorname{sign}(\beta_{0k}^{(j)})/\|\boldsymbol{b}_k\|_2^2\}\right) \bigcap \left(\bigcap_{\{k:\boldsymbol{\beta}_{0k}=\mathbf{0}\}}\{\boldsymbol{t} \,:\, t_k^{(j)} \in [-1,1]\}\right)$
$\iota_C(\boldsymbol{\beta})$	$N_C(\boldsymbol{\beta}_0)$

Table 3.2: **Penalty functions and subgradients** $\partial f_0(\boldsymbol{\beta}_0)$

where $(\boldsymbol{v})_{\mathcal{A}}$ ($\boldsymbol{M}_{\mathcal{A}}$) denotes the subvector (submatrix) of vector $\boldsymbol{v}$ (matrix $\boldsymbol{M}$) consisting of rows (rows and columns) with index in set $\mathcal{A}$. Analogously, for the Adaptive Lasso conjugate estimator, it follows:

$$\sqrt{n}(\operatorname{prox}^{\boldsymbol{Q}_n}_{(\lambda_n f_n)^*}(\hat{\boldsymbol{\beta}}_n^{ls}) - \boldsymbol{\beta}_0)_{\mathcal{A}} \quad \to_d \quad (\boldsymbol{W})_{\mathcal{A}} - [(\boldsymbol{Q}_0)_{\mathcal{A}}]^{-1}(\boldsymbol{Z})_{\mathcal{A}}\,, \tag{3.10}$$

$$\sqrt{n}(\operatorname{prox}^{\boldsymbol{Q}_n}_{(\lambda_n f_n)^*}(\hat{\boldsymbol{\beta}}_n^{ls}))_{\mathcal{A}^c} \quad \to_d \quad (\boldsymbol{W})_{\mathcal{A}^c}\,. \tag{3.11}$$

In limit (3.8), random vector $(\boldsymbol{Z})_{\mathcal{A}}$ is a zero-mean normally distributed random variable with covariance matrix $(\boldsymbol{\Omega}_0)_{\mathcal{A}}$. Therefore, asymptotic distribution (3.8) implies an asymptotic covariance matrix $[(\boldsymbol{Q}_0)_{\mathcal{A}}]^{-1}(\boldsymbol{\Omega}_0)_{\mathcal{A}}[(\boldsymbol{Q}_0)_{\mathcal{A}}]^{-1}$, which is optimal under the standard homoskedasticity assumption $\boldsymbol{\Omega}_0 = \sigma^2\boldsymbol{Q}_0$ for some $\sigma > 0$.[6] On the other hand, limit (3.9) implies a degenerate asymptotic distribution, which is not suited for testing parametric hypotheses depending on components of parameter vector $\boldsymbol{\beta}_0$ in the inactive set. Parametric hypotheses on inactive parameters are instead in principle better testable by means of conjugate Adaptive Lasso estimator and its asymptotic distribution (3.11). However, such a test has to rely on a PLSE that is additionally able to consistently select the inactive components in the given parameter of interest.

[6]The Adaptive Lasso proximal operator is naturally extendible to obtain an optimal PLSE under heteroskedasticity. To this end, let $\hat{\boldsymbol{\beta}}_n^{gls}$ be a GLSE such that $\sqrt{n}(\hat{\boldsymbol{\beta}}_n^{gls} - \boldsymbol{\beta}_0) \to_d \tilde{\boldsymbol{\Omega}}_0^{-1}\tilde{\boldsymbol{Z}}$, where $\tilde{\boldsymbol{Z}} \sim \mathcal{N}(\mathbf{0}, \tilde{\boldsymbol{\Omega}}_0)$, and $\operatorname{prox}^{\tilde{\boldsymbol{\Omega}}_n}_{\lambda_n f_n}(\hat{\boldsymbol{\beta}}_n^{gls})$ be an Adaptive Lasso proximal operator defined with a sequence of weighting matrices such that $\tilde{\boldsymbol{\Omega}}_n \to_{\Pr} \tilde{\boldsymbol{\Omega}}_0$. Optimality of this PGLSE then easily follows with simple modifications of the arguments used to obtain the asymptotic distribution of the Adaptive Lasso proximal operator $\operatorname{prox}^{\boldsymbol{Q}_n}_{(\lambda_n f_n)^*}(\hat{\boldsymbol{\beta}}_n^{ls})$.

3.3 Oracle properties of PLSEs

In our proximal operator approach, the Oracle property (Fan and Li (2001)) of a generic PLSE is naturally formulated as follows.

Definition 1. *A sequence of PLSEs* $\left\{\text{prox}_{\lambda_n f_n}^{\boldsymbol{Q}_n}(\hat{\boldsymbol{\beta}}_n^{ls}) : n \in \mathbb{N}\right\}$ *satisfies the Oracle property if:*

1. *It implies a consistent selection of the nonzero components in parameter vector* $\boldsymbol{\beta}_0$*:* $\lim_{n\to\infty} \mathbb{P}(\hat{\mathcal{A}}_n = \mathcal{A}) = 1$*, where:*

$$\hat{\mathcal{A}}_n := \left\{ j : \left(\text{prox}_{\lambda_n f_n}^{\boldsymbol{Q}_n}(\hat{\boldsymbol{\beta}}_n^{ls})\right)_j \neq 0 \right\} . \tag{3.12}$$

2. *It gives rise to the convergence in distribution:*

$$\sqrt{n}(\text{prox}_{\lambda_n f_n}^{\boldsymbol{Q}_n}(\hat{\boldsymbol{\beta}}_n^{ls}) - \boldsymbol{\beta}_0) \to_d P_{span\{\boldsymbol{e}_j : \beta_{0j} \neq 0\}}^{\boldsymbol{Q}_0}(\boldsymbol{W}) . \tag{3.13}$$

Oracle Property 2 gives rise to an asymptotic distribution for the estimator of the nonzero components of vector $\boldsymbol{\beta}_0$ that is equivalent to the one of a LSE for the same components. Therefore, asymptotic distribution (3.13) is optimal under the standard assumption of homoskedasticity ($\boldsymbol{\Omega}_0 = \sigma \boldsymbol{Q}_0$ for some $\sigma > 0$).[7] Under asymptotic regime (ii) in Proposition 4, Oracle Property 2 corresponds in a one-to-one way to a limit penalty such that:

$$\text{dom}(f_0) = \text{span}\{\boldsymbol{e}_j : \beta_{0j} \neq 0\} . \tag{3.14}$$

Hence, in such settings Oracle Property 2 holds if and only if a PLSE's limit penalty equals the limit penalty implied by an Adaptive Lasso estimator, which gives rise to the explicit asymptotic distributions (3.8)–(3.9) and (3.10)–(3.11) for an Oracle PLSE and its conjugate estimator. This last feature in turn also implies that the inner product between a conjugate Oracle PLSE and the underlying parameter of interest is of order $o(1/\sqrt{n})$: $\sqrt{n}\langle \text{prox}_{(\lambda_n f_n)^*}^{\boldsymbol{Q}_n}(\hat{\boldsymbol{\beta}}_n^{ls}) - \boldsymbol{\beta}_0, \boldsymbol{\beta}_0 \rangle_{\boldsymbol{Q}_n} \to_{\text{Pr}} 0$.

Oracle Property 1 in Definition 1 is determined by the thresholding properties of a PLSE. It is satisfied if and only if $\lim_{n\to\infty} \mathbb{P}(\mathcal{A} \subset \hat{\mathcal{A}}_n) = 1$ and $\lim_{n\to\infty} \mathbb{P}(\hat{\mathcal{A}}_n \cap \mathcal{A}^c = \emptyset) = 1$, where the first of these

[7] Clearly, a natural definition of Oracle Property 2 under heteroskedasticity can be based on proximal operators defined via a GLS estimator. For simplicity of exposition, we do not explicitly incorporate this aspect in the above definition of Oracle Property 2.

two conditions is satisfied by any consistent PLSE. The next proposition provides a first easily verifiable necessary condition for Property 1 in Definition 1 to hold, under the assumptions of Proposition 4.

Proposition 5. *Let the assumptions of Proposition 4 be satisfied and assume that* $\limsup_n \mathbb{P}(\hat{\mathcal{A}}_n \cap \mathcal{A}^c = \emptyset) = 1$. *It then follows:*

$$\mathbb{P}\left((\boldsymbol{W})_{\mathcal{A}^c} = (P^{\boldsymbol{Q}_0}_{\boldsymbol{Q}_0^{-1}(B_0)}(\boldsymbol{W}))_{\mathcal{A}^c}\right) = 1\,, \tag{3.15}$$

where either $B_0 := \lambda_0 \partial f_0(\boldsymbol{\beta}_0)$ *or* $B_0 := N_{\mathrm{dom}(f_0)}(\boldsymbol{\beta}_0)$, *under conditions (i) or (ii) in Proposition 4, respectively.*

Necessary condition (3.15) for Oracle Property 1 is inherently related to the subgradient of the limit penalty of a PLSE. It directly implies, e.g., that limit penalties with a bounded subgradient, such as the Lasso penalty, give rise to PLSEs that do not satisfy Oracle Property 1 in Definition 1. Sufficient conditions for Oracle Property 1 have to impose constraints on the sequence of penalties in Assumption 2, which enable an asymptotic identification of the zero components in vector $\boldsymbol{\beta}_0$. Such constraints restrict the final sample behaviour of the PLSE penalties' subgradients, as detailed by the next proposition.

Proposition 6. *(i) Let Oracle Property 2 in Definition 1 hold. Then, Oracle Property 1 applies if there exists* $\epsilon > 0$ *such that:*

$$\liminf_{n\to\infty} \mathbb{P}\left(\inf_{v_n \in \mathbb{R}^p} \left\{\left\|(v_n)_{\mathcal{A}}\right\| : v_n \in \sqrt{n}\lambda_n \partial f_n((\mathit{span}\{\boldsymbol{e}_j : j \in \mathcal{A}\})^c)\right\} > \epsilon\right) = 1\,. \tag{3.16}$$

(ii) Let asymptotic regime (ii) in Proposition 4 hold. Then, Oracle Property 1 applies if:

$$\inf_{v_n \in \mathbb{R}^p} \left\{\|v_n\| : v_n \in \sqrt{n}\lambda_n \partial f_n((\mathit{span}\{\boldsymbol{e}_j : j \in \mathcal{A}\})^c)\right\} \to_{\mathrm{Pr}} +\infty\,. \tag{3.17}$$

Condition (3.16) enables a correct asymptotic identification of the active components in vector $\boldsymbol{\beta}_0$. It relies on the fact that Oracle Property 2 implies a sequence of optimal penalty subgradient vectors $\boldsymbol{v}_n^{opt}$ such that $(\boldsymbol{v}_n^{opt})_{\mathcal{A}} = o_p(1)$. Therefore, Condition (3.16) requires that all subvectors $(\boldsymbol{v}_n)_{\mathcal{A}}$ in subgradient vectors $\boldsymbol{v}_n \in \lambda_n\sqrt{n}\partial f_n((\mathrm{span}\{\boldsymbol{e}_j : j \in \mathcal{A}\})^c)$, which correspond to incorrect selections of the nonzero elements of $\boldsymbol{\beta}_0$, are uniformly bounded away from the zero vector. In this way, the finite sample PLSE optimality conditions are asymptotically incompatible with an incorrect selection of the active components in vector $\boldsymbol{\beta}_0$. Similarly, under asymptotic regime (ii) in Proposition 4, optimal penalty subgradient vectors

$\boldsymbol{v}_n^{opt}$ are bounded in probability. Therefore, Condition (3.17) ensures that the finite sample PLSE optimality conditions are asymptotically incompatible with an incorrect selection of the active components of $\boldsymbol{\beta}_0$.

Finally, it is important to emphasize that both conditions (3.16) and (3.17) do not depend on the inner product defining a particular PLSE, i.e., they are an intrinsic property of a PLSE's penalties. Hence, they imply a consistent parameter selection for any associated PLSE that is compatible with Oracle Property 2 in Definition 1, or giving rise to an asympotic distribution in asymptotic regime (ii) of Proposition 4, respectively.

3.3.1 Oracle properties of benchmak PLSEs

A violation of Oracle Property 1 by benchmark PLSEs can be inferred from necessary condition (3.15), by inspecting the closed-form subgradients of the associated limit penalties, showing, e.g., that the Lasso estimator does not satisfy Property 1 in Definition 1 under the assumptions of Proposition 5, because it implies a penalty with bounded subgradient.[8] A validation of Oracle Property 1 for PLSEs based on established adaptive penalties, such as the Adaptive Lasso and Adaptive Elastic Net penalties, follows instead directly from Proposition 6. Indeed, the subgradient of the Adaptive Lasso penalty in Table 3.2 reads:

$$\partial f_n(\boldsymbol{\beta}) = \left(\bigcap_{\{j:\beta_j \neq 0\}} \left\{ \boldsymbol{t} \,:\, t_j = \frac{\operatorname{sign}(\beta_j)}{|\tilde{\beta}_{nj}|} \right\} \right) \bigcap \left(\bigcap_{\{j:\beta_j = 0\}} \left\{ \boldsymbol{t} \,:\, t_j \in \left[-\frac{1}{|\tilde{\beta}_{nj}|}, \frac{1}{|\tilde{\beta}_{nj}|} \right] \right\} \right) . \quad (3.18)$$

Therefore, under asymptotic regime (ii) in Proposition 4 direct calculations yield:

$$\inf_{v_n \in \mathbb{R}^p} \left\{ \left\| (v_n)_{\mathcal{A}} \right\| : v_n \in \lambda_n \sqrt{n} \partial f_n((\operatorname{span}\{\boldsymbol{e}_j : j \in \mathcal{A}\})^c) \right\} \geq \min_{j \in \mathcal{A}^c} \{ \lambda_n \sqrt{n} / |\tilde{\beta}_{nj}| \} \to_{\Pr} \infty \,, \quad (3.19)$$

which implies sufficient condition (3.16). Condition (3.16) is satisfied as well, under the same asymptotic regime, for Adaptive Elastic Net penalties, because the resulting subgradient is a linear combination of the subgradients of the Ridge and Adaptive Lasso penalties (Bauschke et al., 2016, Cor. 16.38).

[8] More detailed derivations are reported in the Online Appendix.

3.4 Local alternatives and Oracle inference

The asymptotic distribution in Proposition 4 can be naturally extended to incorporate local alternatives of the form:

$$\boldsymbol{\beta}_{0n} = \boldsymbol{\beta}_0 + \frac{1}{\sqrt{n}}\boldsymbol{Q}_0^{-1}\boldsymbol{\zeta}\,, \tag{3.20}$$

for a Pitman-drift parameter $\boldsymbol{\zeta}$ in a compact subset of $\mathbb{R}^p$. Indeed, also in such a setting Assumption 1 gives rise to a LSE's well-defined asymptotic distribution: $\sqrt{n}(\hat{\boldsymbol{\beta}}_n^{ls} - \boldsymbol{\beta}_0) \to_d \boldsymbol{W}_{\boldsymbol{\zeta}} := \boldsymbol{Q}_0^{-1}(\boldsymbol{Z} + \boldsymbol{\zeta})$. Therefore, in all cases where asymptotic regime (ii) in Proposition 4 holds uniformly over local alternatives (3.20), we obtain:

$$\sqrt{n}\begin{pmatrix} \mathrm{prox}^{\boldsymbol{Q}_n}_{\lambda_n f_n}(\hat{\boldsymbol{\beta}}_n^{ls}) - \boldsymbol{\beta}_0 \\ \mathrm{prox}^{\boldsymbol{Q}_n}_{(\lambda_n f_n)^*}(\hat{\boldsymbol{\beta}}_n^{ls}) - \boldsymbol{\beta}_0 \end{pmatrix} \to_d \begin{pmatrix} \mathrm{prox}^{\boldsymbol{Q}_0}_{\sigma_{N_{\mathrm{dom}(f_0)}(\boldsymbol{\beta}_0)}}(\boldsymbol{W}_{\boldsymbol{\zeta}}) \\ P^{\boldsymbol{Q}_0}_{\boldsymbol{Q}_0^{-1}(N_{\mathrm{dom}(f_0)}(\boldsymbol{\beta}_0))}(\boldsymbol{W}_{\boldsymbol{\zeta}}) \end{pmatrix}. \tag{3.21}$$

Since $\sqrt{n}(\hat{\boldsymbol{\beta}}_n^{ls})_{\mathcal{A}^c} = O_p(1)$ uniformly over local alternatives (3.20), one such situation arises for the Adaptive Lasso and Adaptive Elastic Net penalties constructed with $\sqrt{n}-$consistent estimator $\tilde{\boldsymbol{\beta}}_n = \hat{\boldsymbol{\beta}}_n^{ls}$, in which case:

$$\begin{pmatrix} \mathrm{prox}^{\boldsymbol{Q}_0}_{\sigma_{N_{\mathrm{dom}(f_0)}(\boldsymbol{\beta}_0)}}(\boldsymbol{W}_{\boldsymbol{\zeta}}) \\ P^{\boldsymbol{Q}_0}_{\boldsymbol{Q}_0^{-1}(N_{\mathrm{dom}(f_0)}(\boldsymbol{\beta}_0))}(\boldsymbol{W}_{\boldsymbol{\zeta}}) \end{pmatrix} = \begin{pmatrix} P^{\boldsymbol{Q}_0}_{span\{\boldsymbol{e}_j : j\in\mathcal{A}\}}(\boldsymbol{W}_{\boldsymbol{\zeta}}) \\ P^{\boldsymbol{Q}_0}_{(span\{\boldsymbol{e}_j : j\in\mathcal{A}\})^\perp}(\boldsymbol{W}_{\boldsymbol{\zeta}}) \end{pmatrix}. \tag{3.22}$$

Boundedness in probability of $\sqrt{n}(\hat{\boldsymbol{\beta}}_n^{ls})_{\mathcal{A}^c}$, uniformly over local alternatives (3.20), also implies that the Adaptive Lasso and Adaptive Elastic Net penalties satisfy sufficient condition (3.16) in Proposition 6, and hence Oracle Property 1 in Definition 1, uniformly over local alternatives (3.20). Altogether, this gives following convergence in distribution under local alternatives (3.20) using the Adaptive Lasso or Adaptive Elastic Net estimators:

$$\sqrt{n}\begin{pmatrix} (\mathrm{prox}^{\boldsymbol{Q}_n}_{\lambda_n f_n}(\hat{\boldsymbol{\beta}}_n^{ls}) - \boldsymbol{\beta}_0)_{\hat{\mathcal{A}}_n} \\ (\mathrm{prox}^{\boldsymbol{Q}_n}_{(\lambda_n f_n)^*}(\hat{\boldsymbol{\beta}}_n^{ls}) - \boldsymbol{\beta}_0)_{\hat{\mathcal{A}}_n^c} \end{pmatrix} \to_d \begin{pmatrix} [(\boldsymbol{Q}_0)_{\mathcal{A}}]^{-1}(\boldsymbol{Z} + \boldsymbol{\zeta})_{\mathcal{A}} \\ (\boldsymbol{Q}_0^{-1}(\boldsymbol{Z} + \boldsymbol{\zeta}))_{\mathcal{A}^c} \end{pmatrix}. \tag{3.23}$$

This Oracle distribution naturally gives rise to a set of corresponding Oracle asymptotic tests, which imply a non trivial power for testing parametric hypotheses depending on both the active and the inactive components

in parameter vector $\boldsymbol{\beta}_0$. The power of this asymptotic distribution for testing components in the active set corresponds to the one of an Oracle LSE for estimating these components. In contrast, the power for testing parameters in the inactive set corresponds to the one of a LSE jointly estimating all parameters, which is used to test hypotheses about components in the inactive set.

4 PLSEs and singular designs

In a number of relevant applications with singular designs, the assumption of a positive definite matrix $\boldsymbol{Q}_0$ is not satisfied. In such settings, the solution set of LS estimation problem (1.2) is not a singleton and our approach based on proximal operators needs to be adapted. The main idea is to first select a natural particular solution of LS problem (1.2) and then apply our proximal operator methodology to such solution. To this end, we introduce following weakened version of Assumption 1.

Assumption 3. $\{(\boldsymbol{X}_i, \varepsilon_i) : i \in \mathbb{N}\}$ *is a second-order stationary stochastic process such that* $\mathbb{E}[\boldsymbol{X}_1 \varepsilon_1] = \boldsymbol{0}$ *and satisfying following properties:*

(i) $\boldsymbol{Q}_n \to_{\mathrm{Pr}} \boldsymbol{Q}_0$*;*

(ii) $\boldsymbol{X}'\boldsymbol{\varepsilon}/\sqrt{n} \to_d \boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Omega}_0)$*, for some positive semi-definite matrix* $\boldsymbol{\Omega}_0$*.*

Assumption 3 weakens Assumption 1 by not requiring matrices $\boldsymbol{Q}_0$ and $\boldsymbol{\Omega}_0$ to be positive definite, which allows us to cover singular designs. Note that if matrix $\boldsymbol{Q}_0$ is positive definite then $\boldsymbol{Q}_n^{-1} \to_{\mathrm{Pr}} \boldsymbol{Q}_0^{-1}$ under Assumption 1. More generally, when $\boldsymbol{Q}_0$ and $\boldsymbol{Q}_n$ are singular we can work with Moore-Penrose generalized inverses $\boldsymbol{Q}_0^+$ and $\boldsymbol{Q}_n^+$. In such a setting, Assumption 3 (i) implies the two following properties, as $n \to \infty$ (see, e.g., the Appendix in Madan et al. (1984) and Stewart (1969)):

$$\boldsymbol{Q}_n^+ \to_{\mathrm{Pr}} \boldsymbol{Q}_0^+ \quad \text{and} \quad \mathbb{P}(\mathrm{Range}(\boldsymbol{Q}_n) = \mathrm{Range}(\boldsymbol{Q}_0)) \to 1 \,. \tag{4.1}$$

Assumption 3 also implicitly determines the set of admissible parameters of interest $\boldsymbol{\beta}_0$ in linear model (1.1):

$$\mathcal{B}_0 := \{\boldsymbol{\beta}_0 \in \mathbb{R}^p : \boldsymbol{Q}_0 \boldsymbol{\beta}_0 = \boldsymbol{\delta}_0\} = \{\boldsymbol{Q}_0^+ \boldsymbol{\delta}_0\} + \mathrm{Kernel}(\boldsymbol{Q}_0) \,.$$

This set is nonempty by assumption and $\boldsymbol{Q}_0^+ \boldsymbol{\delta}_0$ is its minimum Euclidean norm parameter. Equivalently, $\boldsymbol{Q}_0^+ \boldsymbol{\delta}_0$ is the unique element of $\mathcal{B}_0$ that belongs to the range of matrix $\boldsymbol{Q}_0$. Therefore, we obtain following

convenient definition of a parameter of interest in the singular design setting:

$$\boldsymbol{\beta}_0^+ := \boldsymbol{Q}_0^+ \boldsymbol{\delta}_0 = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}}\{\|\boldsymbol{\beta}\|_2 : \boldsymbol{Q}_0\boldsymbol{\beta} = \boldsymbol{\delta}_0\} \, .$$

Under Assumption 3, property (4.1) gives rise to an asymptotically normal LSE for parameter $\boldsymbol{\beta}_0^+$, which is given by:

$$\hat{\boldsymbol{\beta}}_n^{ls+} = \boldsymbol{Q}_n^+ \boldsymbol{X}'\boldsymbol{Y}/n = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}}\{\|\boldsymbol{\beta}\|_2 : \boldsymbol{Q}_n\boldsymbol{\beta} = P_{\mathrm{Range}(\boldsymbol{Q}_n)}(\boldsymbol{X}'\boldsymbol{Y}/n)\} \, , \tag{4.2}$$

where $P_{\mathrm{Range}(\boldsymbol{Q}_n)}(\boldsymbol{X}'\boldsymbol{Y}/n) = \boldsymbol{Q}_n\boldsymbol{Q}_n^+\boldsymbol{X}'\boldsymbol{Y}/n$ is the Euclidean projection of $\boldsymbol{X}'\boldsymbol{Y}/n$ on $\mathrm{Range}(\boldsymbol{Q}_n)$. Using definition (4.2), the challenge for our approach based on proximal operators is the singularity of matrix $\boldsymbol{Q}_n$, which does not induce a corresponding inner product in $\mathbb{R}^p$. Therefore, we introduce a different equivalent definition more appropriate for our purposes, which is based on a positive definite matrix $\bar{\boldsymbol{Q}}_n := \boldsymbol{I} + \boldsymbol{Q}_n(\boldsymbol{I} - \boldsymbol{Q}_n^+)$:

$$\hat{\boldsymbol{\beta}}_n^{ls+} := \bar{\boldsymbol{Q}}_n^{-1} P_{\mathrm{Range}(\boldsymbol{Q}_n)}(\boldsymbol{X}'\boldsymbol{Y}/n) = \bar{\boldsymbol{Q}}_n^{-1}(\boldsymbol{Q}_n\boldsymbol{Q}_n^+\boldsymbol{X}'\boldsymbol{Y}/n) = \boldsymbol{Q}_n^+\boldsymbol{X}'\boldsymbol{Y}/n \, . \tag{4.3}$$

In definition (4.3), LSE (4.2) is reproduced as the unique preimage under regular matrix $\bar{\boldsymbol{Q}}_n$ of the Euclidean projection of $\boldsymbol{X}'\boldsymbol{Y}/n$ on $\mathrm{Range}(\boldsymbol{Q}_n)$. Importantly, matrix $\bar{\boldsymbol{Q}}_n$ gives rise to a well-defined inner product $\langle\cdot,\cdot\rangle_{\bar{\boldsymbol{Q}}_n}$, under which the norm of a vector is given by the sum of norms $\|\cdot\|_{\boldsymbol{Q}_n}$ and $\|\cdot\|_{(\boldsymbol{I}-\boldsymbol{Q}_n\boldsymbol{Q}_n^+)}$. These norms are defined on $\mathrm{Range}(\boldsymbol{Q}_n)$ and $\mathrm{Kernel}(\boldsymbol{Q}_n)$, respectively.

4.1 PLSEs with singular designs and their asymptotic distribution

Given LSE definition (4.3) and penalty $f_n \in \Gamma(\mathbb{R}^p)$, following proximal operators are the direct analogues of PLSEs (2.1) and (2.3) under a singular design:

$$\mathrm{prox}_{\lambda_n f_n}^{\bar{\boldsymbol{Q}}_n}(\hat{\boldsymbol{\beta}}_n^{ls+}) := \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \left\{ \frac{1}{2}\left\|\hat{\boldsymbol{\beta}}_n^{ls+} - \boldsymbol{\beta}\right\|_{\bar{\boldsymbol{Q}}_n}^2 + \lambda_n f_n(\boldsymbol{\beta}) \right\} \, , \tag{4.4}$$

$$\mathrm{prox}_{(\lambda_n f_n)^*}^{\bar{\boldsymbol{Q}}_n}(\hat{\boldsymbol{\beta}}_n^{ls+}) := \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \left\{ \frac{1}{2}\left\|\hat{\boldsymbol{\beta}}_n^{ls+} - \boldsymbol{\theta}\right\|_{\bar{\boldsymbol{Q}}_n}^2 + \lambda_n f_n^*(\boldsymbol{\theta}/\lambda_n) \right\} \, , \tag{4.5}$$

where f_n^* is the convex conjugate of f_n under inner product $\langle\cdot,\cdot\rangle_{\bar{\boldsymbol{Q}}_n}$. Given these definitions of PLSEs for singular designs as proximal operators, a convenient feature of our approach is that it allows us to obtain

the asymptotic distribution of these PLSEs with fully analogous arguments as those used to obtain the asymptotic distribution of PLSEs under a regular design.

Proposition 7. *Consider positive definite matrix* $\bar{\boldsymbol{Q}}_0 := \boldsymbol{I} + \boldsymbol{Q}_0(\boldsymbol{I} - \boldsymbol{Q}_0^+)$*, let Assumption 3 be satisfied and Assumption 2 hold with* $\boldsymbol{\beta}_0$ *replaced by* $\boldsymbol{\beta}_0^+$*.*

(i) If $\lambda_n\sqrt{n} \to \lambda_0 > 0$*, it follows:*

$$\sqrt{n}\begin{pmatrix} \operatorname{prox}^{\bar{\boldsymbol{Q}}_n}_{\lambda_n f_n}(\hat{\boldsymbol{\beta}}_n^{ls+}) - \boldsymbol{\beta}_0^+ \\ \operatorname{prox}^{\bar{\boldsymbol{Q}}_n}_{(\lambda_n f_n)^*}(\hat{\boldsymbol{\beta}}_n^{ls+}) - \boldsymbol{\beta}_0^+ \end{pmatrix} \to_d \begin{pmatrix} \operatorname{prox}^{\bar{\boldsymbol{Q}}_0}_{\lambda_0\rho_{\boldsymbol{\beta}_0^+}}(\boldsymbol{W}^+) \\ P^{\bar{\boldsymbol{Q}}_0}_{\bar{\boldsymbol{Q}}_0^{-1}(\lambda_0\partial f_0(\boldsymbol{\beta}_0^+))}(\boldsymbol{W}^+) \end{pmatrix}, \tag{4.6}$$

with random vector $\boldsymbol{W}^+ := \boldsymbol{Q}_0^+\boldsymbol{Z}$*. Moreover,*

$$\operatorname{prox}^{\bar{\boldsymbol{Q}}_0}_{\lambda_0\rho_{\boldsymbol{\beta}_0^+}}(\boldsymbol{W}^+) = (Id - P^{\bar{\boldsymbol{Q}}_0}_{\bar{\boldsymbol{Q}}_0^{-1}(\lambda_0\partial f_0(\boldsymbol{\beta}_0^+))})(\boldsymbol{W}^+)\,. \tag{4.7}$$

(ii) If $\lambda_n\sqrt{n} \to \lambda_0 = 0$ *and* $\lambda_n\sqrt{n} f_n \to_{Pr} \iota_{\operatorname{dom}(f_0)}$ *in epigraph, the above limit in distribution holds with* $\lambda_0\rho_{\boldsymbol{\beta}_0^+}$ *replaced by* $\sigma_{N_{\operatorname{dom}(f_0)}(\boldsymbol{\beta}_0^+)}$ *and* $\lambda_0\partial f_0(\boldsymbol{\beta}_0^+)$ *by* $N_{\operatorname{dom}(f_0)}(\boldsymbol{\beta}_0^+)$*, respectively.*

Proposition 7 is a suitable modification for singular designs of Proposition 4, in which the asymptotic distribution of PLSEs (4.4) and (4.5) is reproduced in the first and the second row of limit (4.6), respectively. The asymptotic distribution of PLSE (4.4) is reproduced by a proximal operator with penalty $\lambda_0\rho_{\boldsymbol{\beta}_0^+}$ or $\sigma_{N_{\operatorname{dom}(f_0)}(\boldsymbol{\beta}_0^+)}$, under asymptotic regimes (i) and (ii), respectively. Conversely, the asymptotic distribution of PLSE (4.5) is reproduced by a projection on the subgradient $\lambda_0\partial f_0(\boldsymbol{\beta}_0^+)$ or $N_{\operatorname{dom}(f_0)}(\boldsymbol{\beta}_0)$, under asymptotic regimes (i) and (ii), respectively. All these proximal operators are applied to a zero mean normally distributed random vector:

$$\boldsymbol{W}^+ = \boldsymbol{Q}_0^+\boldsymbol{Z} = \boldsymbol{Q}_0^+ P_{\operatorname{Range}(\boldsymbol{Q}_0)}(\boldsymbol{Z})\,, \tag{4.8}$$

which reproduces the asymptotic distribution of LSE (4.3) and is given by the minimum Euclidean norm solution in linear system $\boldsymbol{Q}_0\boldsymbol{W} = P_{\operatorname{Range}(\boldsymbol{Q}_0)}(\boldsymbol{Z})$.[9]

Since $\bar{\boldsymbol{Q}}_0 = \boldsymbol{Q}_0$, $\boldsymbol{Q}_0^+ = \boldsymbol{Q}_0^{-1}$ and $\boldsymbol{W} = \boldsymbol{W}^+$ under a regular design, Proposition 7 is a coherent extension for singular designs of characterization (3.6). In analogy to that characterization, it shows that the

[9] Apparently, a direct consequence of the singular design setting is a random vector $\boldsymbol{W}^+$ with a singular covariance matrix $\boldsymbol{Q}_0^+\boldsymbol{\Omega}_0\boldsymbol{Q}_0^+$ in asymptotic distribution (4.6).

asymptotic distribution of PLSE (4.4) is equivalently given by the residual of a projection of normally distributed random variable $\boldsymbol{W}^+$ on the subgradient at $\boldsymbol{\beta}_0^+$ of either f_0 or $\iota_{\mathrm{dom}(f_0)}$, under asymptotic regimes (i) and (ii) of Proposition 7, respectively. In contrast to the regular design setting, both projection operators and convex conjugates in limit (4.6) are defined under inner product $\langle \cdot, \cdot \rangle_{\bar{\boldsymbol{Q}}_0}$, in order to embrace the singularity of matrix $\boldsymbol{Q}_0$. Finally, Proposition 7 also clearly implies following equivalent characterization of the asymptotic distribution of LSE (4.2) using matrix $\bar{\boldsymbol{Q}}_0$:

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n^{ls+} - \boldsymbol{\beta}_0^+) \to_d \bar{\boldsymbol{Q}}_0^{-1} P_{\mathrm{Range}(\boldsymbol{Q}_0)}(\boldsymbol{Z}) = \boldsymbol{Q}_0^+ \boldsymbol{Z} \, . \tag{4.9}$$

4.2 Oracle properties of PLSEs with singular designs

The transparent structure of the asymptotic distribution characterization in Proposition 7 suggests that it may be similarly feasible to study the asymptotic properties of PLSEs under a singular design as under a regular design, with, e.g., Oracle properties that may be analyzed with similar approaches to those introduced earlier for regular designs.

4.2.1 Adaptive Lasso with singular designs

Consider following Adaptive Lasso penalty for a singular design, with the corresponding convergence in epigraph holding under Assumption 3:

$$f_n^+(\boldsymbol{\beta}) := \sum_{j=1}^p \frac{|\beta_j|}{|\hat{\beta}_j^{ls+}|} \to_{\mathrm{Pr}} f_0^+(\boldsymbol{\beta}) := \sum_{j=1}^p \left[\frac{|\beta_j|}{|\beta_{0j}^+|} I(\beta_{0j}^+ \neq 0) + \iota_{\{0\}}(\beta_j) I(\beta_{0j}^+ = 0) \right] \, . \tag{4.10}$$

Under Assumption 3 and asymptotic regime (ii) in Proposition 7, following convergence in epigraph similarly holds:

$$\lambda_n \sqrt{n} f_n^+(\boldsymbol{\beta}) \to_{\mathrm{Pr}} \sum_{\{j : \beta_{0j}^+ = 0\}} \iota_{\{0\}}(\beta_j) = \iota_{\mathrm{dom}(f_0^+)}(\boldsymbol{\beta}) \, .$$

Therefore, from Proposition 7 we obtain following asymptotic distributions for the Adaptive Lasso under a singular design:

$$\sqrt{n}\begin{pmatrix} \text{prox}^{\bar{\boldsymbol{Q}}_n}_{\lambda_n f_n^+}(\hat{\boldsymbol{\beta}}_n^{ls+}) - \boldsymbol{\beta}_0^+ \\ \text{prox}^{\bar{\boldsymbol{Q}}_n}_{(\lambda_n f_n^+)^*}(\hat{\boldsymbol{\beta}}_n^{ls+}) - \boldsymbol{\beta}_0^+ \end{pmatrix} \to_d \begin{pmatrix} P^{\bar{\boldsymbol{Q}}_0}_{\text{span}\{\boldsymbol{e}_j:\beta_{0j}^+\neq 0\}}(\boldsymbol{W}^+) \\ P^{\bar{\boldsymbol{Q}}_0}_{(\text{span}\{\boldsymbol{e}_j:\beta_{0j}^+\neq 0\})^\perp}(\boldsymbol{W}^+) \end{pmatrix} . \tag{4.11}$$

Using the notation $\mathcal{A}^+ := \{j : \beta_{0j}^+ \neq 0\}$, the projection in the first row of the above limit yields following closed-form asymptotic distribution:

$$\sqrt{n}(\text{prox}^{\bar{\boldsymbol{Q}}_n}_{\lambda_n f_n^+}(\hat{\boldsymbol{\beta}}_n^{ls+}) - \boldsymbol{\beta}_0^+)_{\mathcal{A}^+} \to_d [(\bar{\boldsymbol{Q}}_0)_{\mathcal{A}^+}]^{-1}(P_{\text{Range}(\boldsymbol{Q}_0)}(\boldsymbol{Z}))_{\mathcal{A}^+} , \tag{4.12}$$

$$\sqrt{n}(\text{prox}^{\bar{\boldsymbol{Q}}_n}_{\lambda_n f_n^+}(\hat{\boldsymbol{\beta}}_n^{ls+}))_{\mathcal{A}^{+c}} \to_d \mathbf{0} . \tag{4.13}$$

Analogously, for the projection in the second row of the above limit:

$$\sqrt{n}(\text{prox}^{\bar{\boldsymbol{Q}}_n}_{(\lambda_n f_n^+)^*}(\hat{\boldsymbol{\beta}}_n^{ls+}) - \boldsymbol{\beta}_0^+)_{\mathcal{A}^+} \to_d (\boldsymbol{W}^+)_{\mathcal{A}^+} - [(\bar{\boldsymbol{Q}}_0)_{\mathcal{A}^+}]^{-1}(P_{\text{Range}(\boldsymbol{Q}_0)}(\boldsymbol{Z}))_{\mathcal{A}^+} , \tag{4.14}$$

$$\sqrt{n}(\text{prox}^{\bar{\boldsymbol{Q}}_n}_{(\lambda_n f_n^+)^*}(\hat{\boldsymbol{\beta}}_n^{ls+}))_{\mathcal{A}^{+c}} \to_d (\boldsymbol{W}^+)_{\mathcal{A}^{+c}} . \tag{4.15}$$

Compared to the LSE asymptotic distribution (4.9), the Adaptive Lasso for singular designs estimates consistently the zero components of vector $\boldsymbol{\beta}_0^+$ with a rate faster than $1/\sqrt{n}$. For the nonzero components of vector $\boldsymbol{\beta}_0^+$, it implies an asymptotically normal distribution equal to the one of an "Oracle" LSE with singular design, in which the coordinates corresponding to nonzero components of vector $\boldsymbol{\beta}_0^+$ in matrix $\bar{\boldsymbol{Q}}_0$ and projection $P_{\text{Range}(\boldsymbol{Q}_0)}(\boldsymbol{Z})$ of equation (4.9) have been preselected. Compared to the Adaptive Lasso asymptotic distribution for regular designs in equations (3.8)–(3.9), two further features appear. First, under a regular design, $P_{\text{Range}(\boldsymbol{Q}_0)}(\boldsymbol{Z}) = \boldsymbol{Z}$ by construction, i.e., this projection is trivially independent of matrix $\boldsymbol{Q}_0$. Second, $(\boldsymbol{Z})_{\mathcal{A}} \in \text{Range}\,((\boldsymbol{Q}_0)_{\mathcal{A}})$ by construction, a feature that may be violated under a singular design, which explains why matrix $(\bar{\boldsymbol{Q}}_0)_{\mathcal{A}^+}$ appears instead in limit (4.12).

Finally, note that along the same lines of the derivations in Section 3.4 for a regular design, under Assumption 3 asymptotic distribution (4.11) can be extended to incorporate local alternatives of the form:

$$\boldsymbol{\beta}_{0n}^+ = \boldsymbol{\beta}_0^+ + \bar{\boldsymbol{Q}}_0^{-1}\boldsymbol{\zeta} , \tag{4.16}$$

for a Pitman drift parameter $\boldsymbol{\zeta}$ in a compact subset of $\mathbb{R}^p$, simply by replacing $\boldsymbol{W}^+$ by $\boldsymbol{W}^+_{\boldsymbol{\zeta}} := \bar{\boldsymbol{Q}}_0^{-1}(\boldsymbol{Z}+\boldsymbol{\zeta})$ in limit (4.11).

4.2.2 Oracle properties

A natural question also in the singular design setting is whether the Adaptive Lasso satisfies Oracle Property 2 in Definition 1, i.e., whether asymptotic distribution (4.12) is as informative as the one of the infeasible Oracle LSE for the active components of parameter vector $\boldsymbol{\beta}_0$:

$$\hat{\boldsymbol{\beta}}_n^{ls+o} := [(\boldsymbol{Q}_n)_{\mathcal{A}^+}]^+(\boldsymbol{X}'\boldsymbol{Y}/n)_{\mathcal{A}^+} = \underset{\boldsymbol{\beta}^o\in\mathbb{R}^{|\mathcal{A}^+|}}{\operatorname{argmin}} \left\{\|\boldsymbol{\beta}\|_2 : (\boldsymbol{Q}_n)_{\mathcal{A}^+}\boldsymbol{\beta}^o = P_{\operatorname{Range}((\boldsymbol{Q}_n)_{\mathcal{A}^+})}\boldsymbol{X}'\boldsymbol{Y}/n\right\} , \quad (4.17)$$

which under Assumption 3 gives rise to following limit in distribution:

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n^{ls+o} - (\boldsymbol{\beta}_0^+)_{\mathcal{A}^+}) \to_d [(\boldsymbol{Q}_0)_{\mathcal{A}^+}]^+ P_{\operatorname{Range}((\boldsymbol{Q}_0)_{\mathcal{A}^+})}((\boldsymbol{Z})_{\mathcal{A}^+}) . \quad (4.18)$$

The next lemma provides the proper background to answer this question.

Lemma 1. *Following identities hold:*

$$(P_{\operatorname{Range}(\boldsymbol{\Omega}_0)}(\boldsymbol{Z}))_{\mathcal{A}^+} =_d (\boldsymbol{Z})_{\mathcal{A}^+} =_d P_{\operatorname{Range}((\boldsymbol{\Omega}_0)_{\mathcal{A}^+})}((\boldsymbol{Z})_{\mathcal{A}^+}) , \quad (4.19)$$

where $=_d$ denotes equality in distribution. Moreover,

$$[(\bar{\boldsymbol{Q}}_0)_{\mathcal{A}^+}]^{-1}(\boldsymbol{Q}_0)_{\mathcal{A}^+} = [(\boldsymbol{Q}_0)_{\mathcal{A}^+}]^+(\boldsymbol{Q}_0)_{\mathcal{A}^+} . \quad (4.20)$$

Lemma 1 provides a simple way to compare asymptotic distributions (4.12) and (4.18) and it implies identical distributions under the assumption of homoskedasticity ($\boldsymbol{\Omega}_0 = \sigma\boldsymbol{Q}_0$ for some $\sigma > 0$), i.e., Adaptive Lasso PLSE $\operatorname{prox}^{\bar{\boldsymbol{Q}}_n}_{\lambda_n f_n^+}(\hat{\boldsymbol{\beta}}_n^{ls+})$ satisfies Oracle Property 2 in Definition 1.[10] Moreover, for what concerns Oracle Property 2 in Definition 1, it it useful to recall that the sufficient conditions in Proposition 6 for Oracle Property 1 to hold are independent of the inner product used to define a PLSE. Therefore, the Adaptive Lasso PLSE under a singular design consistently selects the nonzero components of the underlying parameter of

[10] While in presence of heteroskedasticity asymptotic distributions (4.12) and (4.18) are generally suboptimal, an Oracle Adaptive Lasso PGLSE for singular designs is readily given by $\operatorname{prox}^{\tilde{\boldsymbol{\Omega}}_n}_{\lambda_n f_n^+}(\hat{\boldsymbol{\beta}}_n^{gls+})$, where matrix $\tilde{\boldsymbol{\Omega}}_n$ is obtained after applying a GLS weighting scheme to matrix $\boldsymbol{X}$ and $\hat{\boldsymbol{\beta}}_n^{gls+}$ is a corresponding GLSE for singular designs.

interest in a fully analogous way as under a regular design. Furthermore, with the same arguments adopted in Section 3.4, it also follows that such consistent selection holds uniformly over local alternatives of the form (4.16). Hence, we also obtain following convergence in distribution:

$$\sqrt{n}\begin{pmatrix} (\mathrm{prox}^{\bar{\boldsymbol{Q}}_n}_{\lambda_n f_n}(\hat{\boldsymbol{\beta}}^{ls+}_n) - \boldsymbol{\beta}^+_0)_{\hat{\mathcal{A}}^+_n} \\ (\mathrm{prox}^{\bar{\boldsymbol{Q}}_n}_{(\lambda_n f^+_n)^*}(\hat{\boldsymbol{\beta}}^{ls+}_n) - \boldsymbol{\beta}^+_0)_{\hat{\mathcal{A}}^{+c}_n} \end{pmatrix} \rightarrow_d \begin{pmatrix} [(\bar{\boldsymbol{Q}}_0)_{\mathcal{A}^+}]^{-1}(P_{\mathrm{Range}(\boldsymbol{Q}_0)}(\boldsymbol{Z}+\boldsymbol{\zeta}))_{\mathcal{A}^+} \\ (\boldsymbol{Q}^+_0(\boldsymbol{Z}+\boldsymbol{\zeta}))_{\mathcal{A}^{+c}} \end{pmatrix}. \quad (4.21)$$

Extending asymptotic distribution (3.23) under a regular design, this Oracle asymptotic distribution naturally gives rise to Oracle asymptotic tests for singular designs, which imply a non trivial power against alternative hypotheses that may depend on both the active and the inactive components in parameter vector $\boldsymbol{\beta}^+_0$.

4.2.3 Oracle properties and grouping effect

Under Assumption 1, Adaptive Lasso proximal operators $\mathrm{prox}^{\boldsymbol{Q}_n}_{\lambda_n f_n}(\hat{\boldsymbol{\beta}}^{ls}_n)$ and $\mathrm{prox}^{\bar{\boldsymbol{Q}}_n}_{\lambda_n f^+_n}(\hat{\boldsymbol{\beta}}^{ls+}_n)$ are numerically identical, since $\boldsymbol{Q}^+_n = \boldsymbol{Q}^{-1}_n$ and $\bar{\boldsymbol{Q}}_n = \boldsymbol{Q}_n$. Moreover, under asymptotic regime (ii) in Proposition 4, they are also asymptotically equivalent to the Adaptive Elastic Net proximal operator $\mathrm{prox}^{\check{\boldsymbol{Q}}_n(\lambda_{2n})}_{\lambda_{1n} f_n}(\check{\boldsymbol{\beta}}^{ls}_n(\lambda_{2n}))$ in equation (2.10), because $\check{\boldsymbol{Q}}_n(\lambda_{2n}) \rightarrow_{\mathrm{Pr}} \boldsymbol{Q}_0$ and $\sqrt{n}(\check{\boldsymbol{\beta}}^{ls}_n(\lambda_{2n}) - \hat{\boldsymbol{\beta}}^{ls}_n) = o_p(1)$.

A different situation emerges when Assumption 3 holds. In such a setting, while LSE $\hat{\boldsymbol{\beta}}^{ls}_n$ is not well-defined $\mathrm{prox}^{\check{\boldsymbol{Q}}_n(\lambda_{2n})}_{\lambda_{1n} f_n}(\check{\boldsymbol{\beta}}^{ls}_n(\lambda_{2n}))$ is still well-defined, because of the embedded Ridge regularizations in estimator $\check{\boldsymbol{\beta}}^{ls}_n(\lambda_{2n})$ and inner product $\langle \cdot, \cdot \rangle_{\boldsymbol{Q}_n(\lambda_{2n})}$. However, while $\mathrm{prox}^{\bar{\boldsymbol{Q}}_n}_{\lambda_n f^+_n}(\hat{\boldsymbol{\beta}}^{ls+}_n)$ satisfies the Oracle property in Definition 1, $\mathrm{prox}^{\check{\boldsymbol{Q}}_n(\lambda_{2n})}_{\lambda_{1n} f_n}(\check{\boldsymbol{\beta}}^{ls}_n(\lambda_{2n}))$ does not even give rise to a well-behaved asymptotic distribution under Assumption 3. This fact is the consequence of two related features. First, since $\check{\boldsymbol{Q}}_n(\lambda_{2n}) \rightarrow_{\mathrm{Pr}} \boldsymbol{Q}_0$ and $\boldsymbol{Q}_0$ is singular, $\langle \cdot, \cdot \rangle_{\check{\boldsymbol{Q}}_n(\lambda_{2n})}$ does not converge in probability to a limit inner product giving rise to a well-defined proximal operator. Second, singularity of $\boldsymbol{Q}_0$ also implies a Ridge estimator $\check{\boldsymbol{\beta}}^{ls}_n(\lambda_{2n})$ with a not well-behaved asymptotic distribution under asymptotic regime (ii) in Proposition 7.

Notwithstanding the above asymptotic distribution issues, a useful finite sample property of the Elastic Net penalty, which is not shared by the Lasso penalty, is the ability to group highly correlated predictors (Zou and Hastie, 2005, Lemma 2 and Thm. 1). Therefore, it is desirable to further modify Adaptive Lasso proximal operator $\mathrm{prox}^{\bar{\boldsymbol{Q}}_n}_{\lambda_n f^+_n}(\hat{\boldsymbol{\beta}}^{ls+}_n)$ for singular designs, in order to specify an Oracle PLSE with an improved bias-variance tradeoff in finite samples relative to the Adaptive Lasso for singular designs. We design such PLSE as a minimum norm Adaptive Elastic Net estimator, based on following proximal

operator:

$$\text{prox}_{\lambda_{1n} f_n^+}^{\check{\tilde{\boldsymbol{Q}}}_n(\lambda_{2n})}(\check{\boldsymbol{\beta}}_n^{ls+}(\lambda_{2n})) = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\text{argmin}} \left\{ \frac{1}{2} \left\| \check{\boldsymbol{\beta}}_n^{ls+}(\lambda_{2n}) - \boldsymbol{\beta} \right\|_{\check{\tilde{\boldsymbol{Q}}}_n(\lambda_{2n})}^2 + \lambda_{1n} f_n^+(\boldsymbol{\beta}) \right\} , \tag{4.22}$$

where

$$\check{\tilde{\boldsymbol{Q}}}_n(\lambda_{2n}) := \frac{1}{1+\lambda_{2n}}(\lambda_{2n}\boldsymbol{I} + \bar{\boldsymbol{Q}}_n) = \boldsymbol{I} + \frac{1}{1+\lambda_{2n}}\boldsymbol{Q}_n(\boldsymbol{I} - \boldsymbol{Q}_n^+) , \tag{4.23}$$

and

$$\check{\boldsymbol{\beta}}_n^{ls+}(\lambda_{2n}) := [\check{\tilde{\boldsymbol{Q}}}_n(\lambda_{2n})]^{-1}\boldsymbol{Q}_n\boldsymbol{Q}_n^+(\boldsymbol{X}'\boldsymbol{Y}/n) . \tag{4.24}$$

Indeed, since $\check{\tilde{\boldsymbol{Q}}}_n(\lambda_{2n}) \to_{\text{Pr}} \bar{\boldsymbol{Q}}_0$ and $\sqrt{n}(\hat{\boldsymbol{\beta}}_n^{ls+} - \check{\boldsymbol{\beta}}_n^{ls+}(\lambda_{2n})) = o_p(1)$ under asymptotic regime (ii) in Proposition 7, PLSE (4.22) is asymptotically equivalent to Adaptive Lasso estimator $\text{prox}_{\lambda_n f_n^+}^{\bar{\boldsymbol{Q}}_n}(\hat{\boldsymbol{\beta}}_n^{ls+})$, i.e., it satisfies Oracle Property 1 in Definition 1. It also does satisfy Oracle Property 2 in Definition 1, because penalty $\lambda_{1n} f_n^+$ satisfies the sufficient conditions in Proposition 6. Finally, it satisfies the grouping property as well, by (Zou and Hastie, 2005, Lemma 2 and Thm. 1) .

5 Bootstrap approximations for the asymptotic distributions of PLSEs

Our framework naturally provides a unifying bootstrap methodology for consistently estimating the asymptotic distribution of a PLSE under a singular design. To this end, we introduce high-level regularity conditions ensuring existence of (i) a valid bootstrap approximation for the asymptotic distribution of the underlying LSE and (ii) a consistent bootstrap estimator for the penalty of the limit proximal operators in the PLSE's asymptotic distribution from Proposition 7.[11]

[11] As under a regular design Proposition 7 coincides with Proposition 4, we implicitly cover with our unifying bootstrap methodology also settings with a regular design.

5.1 Bootstrap approximations for PLSEs with singular designs

Let $\{(\boldsymbol{X}_i^\star, Y_i^\star) : i = 1, \ldots, n\}$ be a bootstrap sample from $\{(\boldsymbol{X}_i, Y_i) : i = 1, \ldots, n\}$ and $\boldsymbol{Q}_n^\star := \boldsymbol{X}^{\star\prime}\boldsymbol{X}^\star/n$. The associated bootstrap LSE for singular designs is defined by:

$$\hat{\boldsymbol{\beta}}_n^{ls+\star} := \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \left\{ \|\boldsymbol{\beta}\|_2 : \boldsymbol{Q}_n^\star\boldsymbol{\beta} = P_{\mathrm{Range}(\boldsymbol{Q}_n^\star)}(\boldsymbol{X}^{\star\prime}\boldsymbol{Y}^\star/n) \right\} = \boldsymbol{Q}_n^{\star+}\boldsymbol{X}^{\star\prime}\boldsymbol{Y}^\star/n \,. \tag{5.1}$$

The next assumption is introduced to ensure (i) that bootstrap LSE (5.1) gives rise to a valid approximation for the asymptotic distribution of LSE (4.2) and (ii) that relevant inner product matrices $\bar{\boldsymbol{Q}}_n^\star = \boldsymbol{I} + \boldsymbol{Q}_n^\star(\boldsymbol{I} - \boldsymbol{Q}_n^{\star+})$ converge to limit matrix $\bar{\boldsymbol{Q}}_0$.

Assumption 4. *Stochastic process $\{(\boldsymbol{X}_i^\star, Y_i^\star) : i \in \mathbb{N}\}$ satisfies following properties:*

(i) $\boldsymbol{Q}_n^\star \to_{\mathrm{Pr}} \boldsymbol{Q}_0$ *and* $\boldsymbol{Q}_n^{\star+} \to_{\mathrm{Pr}} \boldsymbol{Q}_0^+$, *conditionally on* $\{(\boldsymbol{X}_i, Y_i) : i \in \mathbb{N}\}$;

(ii) $\boldsymbol{b}_n^{ls+\star} := \sqrt{n}(\hat{\boldsymbol{\beta}}_n^{ls+\star} - \hat{\boldsymbol{\beta}}_n^{ls+}) \to_d \boldsymbol{W}^+ = \boldsymbol{Q}_0^+\boldsymbol{Z}$, *conditionally on* $\{(\boldsymbol{X}_i, Y_i) : i \in \mathbb{N}\}$.

For a sequence of stochastic penalties $\phi_n^\star \in \Gamma(\mathbb{R}^p)$, we further define two conjugate bootstrap PLSEs by:

$$\mathrm{prox}^{\bar{\boldsymbol{Q}}_n^\star}_{n\lambda_n\phi_n^\star}(\boldsymbol{b}_n^{ls+\star}) = \underset{\boldsymbol{b}\in\mathbb{R}^p}{\operatorname{argmin}} \left\{ \frac{1}{2}\left\|\boldsymbol{b}_n^{ls+\star} - \boldsymbol{b}\right\|^2_{\bar{\boldsymbol{Q}}_n^\star} + n\lambda_n\phi_n^\star(\boldsymbol{b}) \right\} , \tag{5.2}$$

$$\mathrm{prox}^{\bar{\boldsymbol{Q}}_n^\star}_{(n\lambda_n\phi_n^\star)^*}(\boldsymbol{b}_n^{ls+\star}) = \underset{\boldsymbol{t}\in\mathbb{R}^p}{\operatorname{argmin}} \left\{ \frac{1}{2}\left\|\boldsymbol{b}_n^{ls+\star} - \boldsymbol{t}\right\|^2_{\bar{\boldsymbol{Q}}_n^\star} + n\lambda_n(\phi_n^\star)^*(\boldsymbol{t}/(n\lambda_n)) \right\} , \tag{5.3}$$

where convex conjugation is defined under inner product $\langle\cdot,\cdot\rangle_{\bar{\boldsymbol{Q}}_n^\star}$. Since Assumption 4 implies the convergence in distribution of random function $\left\|\boldsymbol{b}_n^{ls+\star} - \cdot\right\|^2_{\bar{\boldsymbol{Q}}_n^\star}$ to random function $\left\|\boldsymbol{W}^+ - \cdot\right\|^2_{\bar{\boldsymbol{Q}}_0}$, uniformly on compact sets and conditionally on $\{(\boldsymbol{X}_i, Y_i) : i \in \mathbb{N}\}$, these bootstrap estimators produces valid approximations to the asymptotic PLSE distributions in Proposition 7 whenever their penalties converge in epigraph to the limit penalties detailed in Proposition 7. Given the preservation of epigraph convergence under convex conjugation, the next assumption is all we need to obtain asymptotically valid bootstrap procedures for PLSEs with singular designs.[12]

Assumption 5. $\phi_n^\star \in \Gamma(\mathbb{R}^p)$, $\mathbb{P}-$*almost surely. Moreover, one of the following two conditions holds:*

(i) $\sqrt{n}\phi_n^\star \to_{\mathrm{Pr}} \rho_{\boldsymbol{\beta}_0^+}$ *in epigraph, conditionally on* $\{(\boldsymbol{X}_i, Y_i) : i = 1, \ldots, n\}$.

[12] See again (Mosco, 1971, Thm. 1).

(ii) $n\lambda_n\phi_n^\star \to_{\Pr} \sigma_{N_{\mathrm{dom}(f_0)}(\boldsymbol{\beta}_0^+)}$ *in epigraph, conditionally on* $\{(\boldsymbol{X}_i, Y_i) : i = 1, \ldots, n\}$.

Assumption 5 is a relatively weak one, since in many practical cases it can be satisfied using suitable consistent estimators for limit penalties $\rho_{\boldsymbol{\beta}_0^+}$ or $\sigma_{N_{\mathrm{dom}(f_0)}(\boldsymbol{\beta}_0^+)}$. While such consistent estimators may be in principle directly available from a consistent estimator of $\boldsymbol{\beta}_0^+$ when the limit penalties $\rho_\cdot$ or $\sigma_{N_{\mathrm{dom}(f_0)}(\cdot)}$ are continuous at $\boldsymbol{\beta}_0^+$, many highly relevant settings give rise to limit penalties that may not be continuous at $\boldsymbol{\beta}_0^+$.[13] Given Assumption 5, validity of the bootstrap approximation implied by bootstrap PLSEs (5.2) and (5.3) for the asymptotic distributions in Proposition 7 follows with the next proposition.

Proposition 8. *Let Assumption 4 be satisfied.*

(i) If $\sqrt{n}\lambda_n \to \lambda_0 > 0$ *and Assumption 5 (i) holds, then conditionally on* $\{(\boldsymbol{X}_i, Y_i) : i = 1, \ldots, n\}$*:*

$$\begin{pmatrix} \mathrm{prox}^{\bar{\boldsymbol{Q}}_n^\star}_{n\lambda_n\phi_n^\star}(\boldsymbol{b}_n^{ls+\star}) \\ \mathrm{prox}^{\bar{\boldsymbol{Q}}_n^\star}_{(n\lambda_n\phi_n^\star)^*}(\boldsymbol{b}_n^{ls+\star}) \end{pmatrix} \to_d \begin{pmatrix} \mathrm{prox}^{\bar{\boldsymbol{Q}}_0}_{\lambda_0\rho_{\boldsymbol{\beta}_0^+}}(\boldsymbol{W}^+) \\ P^{\bar{\boldsymbol{Q}}_0}_{\boldsymbol{Q}_0^{-1}(\lambda_0\partial f(\boldsymbol{\beta}_0^+))}(\boldsymbol{W}^+) \end{pmatrix} . \tag{5.4}$$

(ii) If $\sqrt{n}\lambda_n \to \lambda_0 = 0$ *and Assumption 5 (ii) holds, then the above limits hold with* $\lambda_0\rho_{\boldsymbol{\beta}_0^+}$ *and* $\lambda_0\partial f(\boldsymbol{\beta}_0^+)$ *replaced by* $\sigma_{N_{\mathrm{dom}(f_0)}(\boldsymbol{\beta}_0^+)}$ *and* $N_{\mathrm{dom}(f_0)}(\boldsymbol{\beta}_0^+)$*, respectively.*

The first row in the asymptotic distributions of Proposition 8 give rise to "direct" valid bootstrap approximations for the asymptotic distribution of PLSE $\mathrm{prox}^{\bar{\boldsymbol{Q}}_n}_{\lambda_n f_n}(\hat{\boldsymbol{\beta}}_n^{ls+})$. Analogously, the second row in the asymptotic distributions of Proposition 8 give rise to valid bootstrap approximations for the asymptotic distribution of conjugate PLSE $\mathrm{prox}^{\bar{\boldsymbol{Q}}_n}_{(\lambda_n f_n)^*}(\hat{\boldsymbol{\beta}}_n^{ls+})$. Therefore, recalling Moreau's decomposition, we can always equivalently characterize the asymptotic properties of bootstrap PLSE (5.2) using the conjugate bootstrap proximal operator. For instance, under asymptotic regime (i) from Proposition 8, we immediately obtain:

$$(Id - \mathrm{prox}^{\bar{\boldsymbol{Q}}_n^\star}_{(n\lambda_n\phi_n^\star)^*})(\boldsymbol{b}_n^{ls\star+}) \to_d \mathrm{prox}^{\bar{\boldsymbol{Q}}_0}_{\lambda_0\rho_{\boldsymbol{\beta}_0^+}}(\boldsymbol{W}^+) , \tag{5.5}$$

as well as a corresponding limit under asymptotic regime (ii).

[13] Similar to the regular design setting, consistent estimators of $\rho_{\boldsymbol{\beta}_0^+}$ or $\sigma_{N_{\mathrm{dom}(f_0)}(\boldsymbol{\beta}_0^+)}$ can be usually constructed by exploiting the specific penalties' functional forms. The Online Appendix collects various concrete examples of such bootstrap approaches for benchmark PLSEs. Section 5.2 details valid bootstrap approximations for the Oracle PLSEs with singular design introduced in Section 4.2.

5.2 Bootstrap approximations for Oracle PLSEs with singular designs

Recall the Adaptive Lasso limit penalty under asymptotic regime (ii) in Proposition 7:

$$\sigma_{N_{\mathrm{dom}(f_0^+)}(\boldsymbol{\beta}_0^+)}(\boldsymbol{b}) = \sum_j \iota_{\{0\}}(b_j) I_{\{\beta_{0j}^+=0\}} \ ,$$

and consider following penalty for bootstrap PLSE (5.2):

$$\phi_n^\star(\boldsymbol{b}) := \sum_{j=1}^p \frac{1}{\left|\hat{\beta}_{nj}^{ls+\star}\right|} \left[\left| \hat{\beta}_{nj}^{ls+} I_{\left\{|\hat{\beta}_{nj}|>0\right\}} + b_j/\sqrt{n} \right| - \left| \hat{\beta}_{nj}^{ls+} I_{\left\{|\hat{\beta}_{nj}|>0\right\}} \right| \right] \ , \tag{5.6}$$

where $\hat{\boldsymbol{\beta}}_n = \mathrm{prox}^{\bar{\boldsymbol{Q}}_n}_{\lambda_n f_n^+}(\hat{\boldsymbol{\beta}}_n^{ls+})$ is the Adaptive Lasso estimator for singular designs. Analogously to the regular design setting in Camponovo (2015), it then follows: $n\lambda_n \phi_n^\star \to_{\mathrm{Pr}} \sigma_{N_{\mathrm{dom}(f_0^+)}(\boldsymbol{\beta}_0^+)}$ in epigraph, conditionally on $\{(\boldsymbol{X}_i, Y_i) : i = 1, \ldots, n\}$. Hence, bootstrap PLSE (5.2) gives rise to a valid bootstrap approximation for the Oracle asymptotic distribution of the Adaptive Lasso under asymptotic regime (ii) in Proposition 7. The corresponding closed-form expression for dual bootstrap penalty $(n\lambda\phi_n^\star)^* = n\lambda_n(\phi_n^\star)^* \circ \frac{1}{n\lambda_n} Id$ is obtained by computing convex conjugate $(\phi_n^\star)^* \circ \frac{1}{n\lambda_n} Id$, which is given by following Fenchel-Young penalty:[14]

$$(\phi_n^\star)^*(\boldsymbol{t}/(n\lambda_n)) = \sum_{j=1}^p FYL\left(|\hat{\beta}_{nj}^{ls+}/\hat{\beta}_{nj}^{ls+\star}| I_{\left\{|\hat{\beta}_{nj}|>0\right\}}, t_j |\hat{\beta}_{nj}^{ls+\star}|/(\lambda_n\sqrt{n}) \right) \ , \tag{5.7}$$

with a closed-form Fenchel-Young loss $FYL : \mathbb{R}^2 \to \mathbb{R}_+$ defined by:

$$FYL\,(x, y) = |x| + \iota_{[-1,1]}(y) - xy \ . \tag{5.8}$$

With the above bootstrap penalty choices, we thus obtain that bootstrap PLSEs (5.2) and (5.3) imply a valid bootstrap approximation for the joint asymptotic distribution of conjugate Adaptive Lasso estimators $\mathrm{prox}^{\bar{\boldsymbol{Q}}_n}_{(\lambda_n f_n^+)}(\hat{\boldsymbol{\beta}}_n^{ls+})$ and $\mathrm{prox}^{\bar{\boldsymbol{Q}}_n}_{(\lambda_n f_n^+)^*}(\hat{\boldsymbol{\beta}}_n^{ls+})$, under asymptotic regime (ii) in Proposition 7:

$$\begin{pmatrix} \mathrm{prox}^{\bar{\boldsymbol{Q}}_n^\star}_{n\lambda_n\phi_n^\star}(\boldsymbol{b}_n^{ls+\star}) \\ \mathrm{prox}^{\bar{\boldsymbol{Q}}_n^\star}_{(n\lambda_n\phi_n^\star)^*}(\boldsymbol{b}_n^{ls+\star}) \end{pmatrix} \to_d \begin{pmatrix} \mathrm{prox}^{\bar{\boldsymbol{Q}}_0}_{\sigma_{N_{\mathrm{dom}(f_0^+)}(\boldsymbol{\beta}_0^+)}}(\boldsymbol{W}^+) \\ P^{\bar{\boldsymbol{Q}}_0}_{\bar{\boldsymbol{Q}}_0^{-1}(N_{\mathrm{dom}(f_0^+)}(\boldsymbol{\beta}_0^+))}(\boldsymbol{W}^+) \end{pmatrix} \ . \tag{5.9}$$

[14] See the Online Appendix for a derivation of this formula.

6 Conclusions

In this paper, we proposed a unified framework for the statistical analysis of Penalized Least Squares Estimators (PLSEs) with convex penalties, which is applicable both under a regular and a singular design. This framework is founded on a reinterpretation of PLSEs as proximity operators, under a corresponding inner product. It gives rise to an exhaustive characterization of PLSEs' statistical properties, which relies on suitable functional transformations of the PLSEs' defining penalties.

Under fairly general high-level assumptions, we have shown that PLSEs converge in distribution to limit operators evaluated at a Gaussian random vector. Such limit operators are fully determined by the directional derivative or, equivalently, the subgradient, of the PLSEs' limit penalties at the true parameter of interest. We also obtained necessary and sufficient conditions for a PLSE Oracle property to hold, which are formulated exclusively in terms of the subgradients of the PLSEs' defining penalties. Finally, we exploited Moreau's decomposition of proximity operators to uniquely link a PLSE to a (non penalized) Least Squares Estimator (LSE) via its conjugate PLSE. This link allowed us to make use of the joint asymptotic distribution of Oracle PLSEs and their conjugate PLSEs, in order to derive Oracle asymptotic distributions implying a nontrivial asymptotic power for testing hypotheses on the active and inactive components of the underlying parameter of interest.

While our theory naturally covers under a unifying framework existing statistical characterizations of PLSEs for regular designs, such as the Lasso, the Adaptive Lasso and the Adaptive Elastic Net, it naturally covers PLSEs with singular designs as well. The construction of well-behaved PLSEs for singular designs is an open problem in the literature and our theory clearly identifies the challenges that need to be overcome in such settings: the specification of a suitable well-behaved proximity operator and the definition of a well-behaved LSE. We bypassed these problems by designing convenient Oracle PLSEs for singular designs, which arise from proximal operators defined by appropriate modifications of existing adaptive penalties in the literature and applied to a minimum Euclidean norm LSE.

We additionally developed a systematic approach for constructing valid bootstrap approximations for PLSEs' asymptotic distributions, both under a regular and a singular design. This approach is based as well on transparent high-level assumptions requiring (i) existence of a valid bootstrap approximation for the asymptotic distribution of the underlying LSE and (ii) existence of a consistent bootstrap estimator for the directional derivative of the PLSEs' limit penalties at the parameter of interest.

Finally, while our theory in this paper covers penalized estimators induced by a quadratic loss function

under a fix dimension of the parameter space, interesting avenues for future research cover extensions to nonlinear penalized estimators and settings with a growing parameter dimension. For several penalized estimation settings induced by a non penalized loss function that is asymptotically equivalent to a quadratic loss function, our approach can be naturally extended with mild modifications. For settings with a growing parameter dimension, our proximal operator approach can be helpful, e.g., to construct nonasymptotic risk bounds for a broad class of Lipschitz continuous penalties, both under regular and singular designs.

References

Donald W. K. Andrews. Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica*, 68(2):399–405, 2000.

Hedy Attouch. *Variational convergence for functions and operators*, volume 1. Pitman Advanced Publishing Program, 1984.

Heinz H Bauschke, Patrick L Combettes, et al. *Convex analysis and monotone operator theory in Hilbert spaces*. Springer, 2nd edition, 2016.

Lorenzo Camponovo. On the validity of the pairs bootstrap for lasso estimators. *Biometrika*, 102(4):981–987, 2015.

Arindam Chatterjee and Soumendra Nath Lahiri. Bootstrapping lasso estimators. *Journal of the American Statistical Association*, 106(494):608–625, 2011.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

Charles J Geyer. On the asymptotics of constrained m-estimation. *The Annals of statistics*, pages 1993–2010, 1994.

Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of convex analysis*. Springer Science & Business Media, 2004.

Arthur E Hoerl and Robert W Kennard. Ridge regression: applications to nonorthogonal problems. *Technometrics*, 12(1):69–82, 1970.

Keith Knight. Epi-convergence in distribution and stochastic equi-semicontinuity. *Unpublished manuscript*, 37(7):14, 1999.

Chong Kiew Liew. Inequality constrained least-squares estimation. *Journal of the American Statistical Association*, 71(355):746–751, 1976.

L. Puri Madan, Carl T Russel, and Thomas Mathew. Convergence of generalized inverses with applications to asymptotic hypothesis testing. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 46(2): 277–286, 1984.

Jean Jacques Moreau. Fonctions convexes duales et points proximaux dans un espace hilbertien. *Comptes rendus hebdomadaires des séances de l'Académie des sciences*, 255:2897–2899, 1962.

Umberto Mosco. On the continuity of the young-fenchel transform. *Journal of Mathematical Analysis and Applications*, 35(3):518–535, 1971.

R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

Gabriella Salinetti and Roger J-B Wets. On the convergence of closed-valued measurable multifunctions. *Transactions of the American Mathematical Society*, 266(1):275–289, 1981.

Gabriella Salinetti and Roger J-B Wets. On the convergence in distribution of measurable multifunctions (random sets) normal integrands, stochastic processes and stochastic infima. *Mathematics of Operations Research*, 11(3):385–419, 1986.

G. Stewart. On the continuity of the generalized inverse. *SIAM Journal on Applied Mathematics*, 17(1): 33–45, 1969.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101 (476):1418–1429, 2006.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.

Hui Zou and Hao Helen Zhang. On the adaptive elastic-net with a diverging number of parameters. *Annals of statistics*, 37(4):1733, 2009.

Online Appendix

Online Appendix: Proofs

Proof of Proposition *1*. Since $\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\|_2 = \left\|\hat{\boldsymbol{\beta}}_n^{ls}-\boldsymbol{\beta}\right\|_{\boldsymbol{Q}_n} + K$, where term K is independent of $\boldsymbol{\beta}$, PLSE (1.3) equivalently reads:

$$\hat{\boldsymbol{\beta}}_n = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \left\{ \frac{1}{2}\left\|\hat{\boldsymbol{\beta}}_n^{ls}-\boldsymbol{\beta}\right\|_{\boldsymbol{Q}_n}^2 + \lambda_n f_n(\boldsymbol{\beta}) \right\} .$$

By assumption, $f_n \in \Gamma(\mathbb{R}^p)$ and $\langle\cdot,\cdot\rangle_{\boldsymbol{Q}_n}$ defines via symmetric positive definite matrix $\boldsymbol{Q}_n$ an inner product in $\mathbb{R}^p$, $\mathbb{P}-$almost surely, Therefore, $\hat{\boldsymbol{\beta}}_n$ is the proximity operator of $\lambda_n f_n$ at $\hat{\boldsymbol{\beta}}_n^{ls}$ under inner product $\langle\cdot,\cdot\rangle_{\boldsymbol{Q}_n}$ (Bauschke et al., 2016, Def. 12.23). Finally, the identity $\hat{\boldsymbol{\beta}}_n = \hat{\boldsymbol{\beta}}_n^{ls} - \operatorname{prox}^{\boldsymbol{Q}_n}_{(\lambda_n f_n)^*}(\hat{\boldsymbol{\beta}}_n^{ls})$, where $(\lambda_n f_n)^*$ is the convex conjugate of f_n under inner product $\langle\cdot,\cdot\rangle_{\boldsymbol{Q}_n}$, follows from Moreau's decomposition (Bauschke et al., 2016, Thm. 14.3). This concludes the proof. □

Proof of Proposition *2*. (i) Given Lasso penalty $\lambda_n f_n = \lambda_n\|\boldsymbol{\beta}\|_1$ and its convex conjugate $(f_n\lambda_n)^*(\boldsymbol{\theta}) = \iota_{\lambda_n\mathcal{B}_\infty}(\boldsymbol{Q}_n\boldsymbol{\theta})$ under inner product $\langle\cdot,\cdot\rangle_{\boldsymbol{Q}_n}$ (see Table 2.1), conjugate proximal operator (2.2) reads explicitly:

$$\operatorname{prox}^{\boldsymbol{Q}_n}_{(\lambda_n f_n)^*}(\hat{\boldsymbol{\beta}}_n^{ls}) = \underset{\boldsymbol{\theta}\in\mathbb{R}^p}{\operatorname{argmin}} \left\{ \frac{1}{2}\left\|\hat{\boldsymbol{\beta}}_n^{ls}-\boldsymbol{\theta}\right\|_{\boldsymbol{Q}_n}^2 \;:\; \boldsymbol{\theta}\in\boldsymbol{Q}_n^{-1}(\lambda_n\mathcal{B}_\infty) \right\} .$$

Since $\boldsymbol{Q}_n^{-1}(\lambda_n\mathcal{B}_\infty) = C_n$, with C_n the polyhedron defined in equation (2.5), the stated formula for the Lasso estimator follows from Moreau's decomposition. (ii) The proof of the formula for the Adaptive Lasso follows as in the case of the Lasso, using the corresponding convex conjugate $(\lambda_n f_n)^*(\boldsymbol{\theta}) = \iota_{B_n}(\boldsymbol{Q}_n\boldsymbol{\theta})$ under inner product $\langle\cdot,\cdot\rangle_{\boldsymbol{Q}_n}$ (see again Table 2.1). Finally, in order to prove the corresponding projection formula for the naïve Elastic Net estimator, we set $\lambda_{1n} := \lambda_n w$, $\lambda_{2n} := \lambda_n(1-w)/2$ and write down the optimization problem defining this estimator:

$$\begin{aligned}
\operatorname{prox}^{\boldsymbol{Q}_n}_{\lambda_n f_n}(\hat{\boldsymbol{\beta}}_n^{ls}) &= \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \left\{ \frac{1}{2}\left\|\hat{\boldsymbol{\beta}}_n^{ls}-\boldsymbol{\beta}\right\|_{\boldsymbol{Q}_n}^2 + \lambda_{1n}\|\boldsymbol{\beta}\|_1 + \lambda_{2n}\|\boldsymbol{\beta}\|_2^2 \right\} \\
&= \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \left\{ \frac{1}{2}\|\boldsymbol{\beta}\|^2_{\boldsymbol{Q}_n+\lambda_{2n}\boldsymbol{I}} - \boldsymbol{\beta}'(\boldsymbol{X}'\boldsymbol{Y}/n) + \lambda_{1n}\|\boldsymbol{\beta}\|_1 \right\} \\
&= \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \left\{ \frac{1}{2}\left\|[\boldsymbol{Q}_n(\lambda_{2n})]^{-1}(\boldsymbol{X}'\boldsymbol{Y}/n) - \boldsymbol{\beta}\right\|^2_{\boldsymbol{Q}_n(\lambda_{2n})} + \lambda_{1n}\|\boldsymbol{\beta}\|_1 \right\} ,
\end{aligned}$$

with matrix $\boldsymbol{Q}_n(\lambda_{2n}) := \boldsymbol{Q}_n + \lambda_{2n}\boldsymbol{I}_n$ and linear estimator $[\boldsymbol{Q}_n(\lambda_{2n})]^{-1}(\boldsymbol{X}'\boldsymbol{Y}/n) =: \hat{\boldsymbol{\beta}}_n^{ls}(\lambda_{2n})$. The naïve Elastic Net estimator is then given by a Lasso proximal operator with penalty parameter λ_{1n}, which is

applied to modified Least Squares estimator $\hat{\boldsymbol{\beta}}_n^{ls}(\lambda_{2n})$. Therefore, using our previous projection formula for the Lasso estimator, we obtain:

$$\mathrm{prox}_{\lambda_n f_n}^{\boldsymbol{Q}_n}(\hat{\boldsymbol{\beta}}_n^{ls}) = \left(Id - P_{C_n(\lambda_n,w)}^{\boldsymbol{Q}_n(\lambda_{2n})}\right)(\hat{\boldsymbol{\beta}}_n^{ls}(\lambda_{2n}))\,, \tag{6.1}$$

with the Polyhedron:

$$C_n(\lambda_n, w) := \bigcap_{j=1}^{p} \{\boldsymbol{\theta} : |\langle \boldsymbol{e}_j, \boldsymbol{\theta}\rangle_{\boldsymbol{Q}_n(\lambda_{2n})}| \leq \lambda_{1n}\}\,. \tag{6.2}$$

For the case of an orthonormal design, this projection formula reads:

$$\mathrm{prox}_{\lambda_n f_n}^{\boldsymbol{I}_n}(\hat{\boldsymbol{\beta}}_n^{ls}) = \left(Id - P_{C_n(\lambda_n,w)}^{(1+\lambda_{2n})\boldsymbol{I}_n}\right)(\hat{\boldsymbol{\beta}}_n^{ls}(\lambda_{2n}))\,, \tag{6.3}$$

with linear estimator $\hat{\boldsymbol{\beta}}_n^{ls}(\lambda_{2n}) = \frac{1}{1+\lambda_{2n}}\hat{\boldsymbol{\beta}}_n^{ls}$ and polyhedron:

$$C_n(\lambda_n, w) := \bigcap_{j=1}^{p} \{\boldsymbol{\theta} : |\langle \boldsymbol{e}_j, \boldsymbol{\theta}\rangle| \leq \lambda_{1n}/(1+\lambda_{2n})\}\,. \tag{6.4}$$

Together, this yields:

$$\mathrm{prox}_{\lambda_n f_n}^{\boldsymbol{I}_n}(\hat{\boldsymbol{\beta}}_n^{ls}) = \frac{1}{1+\lambda_{2n}}(Id - P_{C_n(\lambda_{1n})}^{\boldsymbol{I}_n})(\hat{\boldsymbol{\beta}}_n^{ls})\,, \tag{6.5}$$

with polyhedron:

$$C_n(\lambda_{1n}) = \bigcap_{j=1}^{p} \{\boldsymbol{\theta} : |\langle \boldsymbol{e}_j, \boldsymbol{\theta}\rangle| \leq \lambda_{1n}\}\,. \tag{6.6}$$

This concludes the proof. □

Proof of Proposition *3*. Let $g_n(\boldsymbol{\beta}) := \frac{1}{2}\left\|\hat{\boldsymbol{\beta}}_n^{ls} - \boldsymbol{\beta}\right\|_{\boldsymbol{Q}_n}^2$ and $g_0(\boldsymbol{\beta}) := \frac{1}{2}\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}\|_{\boldsymbol{Q}_0}^2$. Moreover, let $\mathcal{P}_n(\boldsymbol{\beta}) := g_n(\boldsymbol{\beta}) + \lambda_n f_n(\boldsymbol{\beta})$ and $\mathcal{P}_0(\boldsymbol{\beta}) := g_0(\boldsymbol{\beta}) + w_0(\boldsymbol{\beta})$, where $w_0(\boldsymbol{\beta}) := \lambda_0 f_0(\boldsymbol{\beta})$ and $w_0(\boldsymbol{\beta}) := \iota_{\mathrm{dom}(f_0)}(\boldsymbol{\beta})$, respectively, under asymptoptic regime (i) and (ii). Given the strict convexity and coercivity (Bauschke et al., 2016, Prop. 11.14) of functions $\mathcal{P}_n$ and $\mathcal{P}_0$, we only need to prove that $\mathcal{P}_n \to_{\mathrm{Pr}} \mathcal{P}_0$ in epigraph, in order to

imply the convergence:

$$\operatorname{prox}^{\boldsymbol{Q}_n}_{\lambda_n f_n}(\hat{\boldsymbol{\beta}}^{ls}_n) = \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \mathcal{P}_n(\boldsymbol{\beta}) \to_{\Pr} \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \mathcal{P}_0(\boldsymbol{\beta}) = \operatorname{prox}^{\boldsymbol{Q}_0}_{w_0}(\boldsymbol{\beta}_0) \; . \tag{6.7}$$

To show that $\mathcal{P}_n \to_{\Pr} \mathcal{P}_0$ in epigraph, note that $\lambda_n f_n \to_{\Pr} w_0$ in epigraph by assumption. Moreover, $g_n \to_{\Pr} g_0$, uniformly on compact subsets of $\mathbb{R}^p$, because g_n is continuous in parameters $\boldsymbol{Q}_n$ and $\hat{\boldsymbol{\beta}}^{ls}_n$, which have limit in probability $\boldsymbol{Q}_0$ and $\boldsymbol{\beta}_0$, respectively, under Assumption 1. Therefore, from (Attouch, 1984, Thm. 2.15) $\mathcal{P}_n \to_{\Pr} \mathcal{P}_0$ in epigraph, and convergence (6.7) holds. In addition, the identity $\operatorname{prox}^{\boldsymbol{Q}_0}_{\lambda_0 f_0} = Id - \operatorname{prox}^{\boldsymbol{Q}_0}_{w^*_0}$, where $w^*_0 = \lambda_0 f^*_0(\cdot/\lambda_0)$ under asymptotic regime (i) and $w^*_0 = \sigma_{\operatorname{dom}(f_0)}$ under asymptotic regime (ii), respectively, follows with Moreau's decomposition (Bauschke et al., 2016, Thm. 14.3). Finally, since $\boldsymbol{\beta}_0 \in \operatorname{dom}(f_0)$, we obtain $\iota_{\operatorname{dom}(f_0)}(\boldsymbol{\beta}_0) = 0$, therefore implying $\operatorname{prox}^{\boldsymbol{Q}_0}_{\iota_{\operatorname{dom}(f_0)}}(\boldsymbol{\beta}_0) = \boldsymbol{\beta}_0$ and $\operatorname{prox}^{\boldsymbol{Q}_0}_{\sigma_{\operatorname{dom}(f_0)}}(\boldsymbol{\beta}_0) = \mathbf{0}$. This concludes the proof. □

Proof of Proposition *4*. For any $\boldsymbol{b} \in \mathbb{R}^p$, let $\mathcal{P}_n(\boldsymbol{b}) := g_n(\boldsymbol{b}) + h_n(\boldsymbol{b})$, where:

$$g_n(\boldsymbol{b}) := \frac{1}{2}\left\|\sqrt{n}(\hat{\boldsymbol{\beta}}^{ls}_n - \boldsymbol{\beta}_0) - \boldsymbol{b}\right\|^2_{\boldsymbol{Q}_n} \; ; \; h_n(\boldsymbol{b}) := n\lambda_n \left[f_n(\boldsymbol{\beta}_0 + \boldsymbol{b}/\sqrt{n}) - f_n(\boldsymbol{\beta}_0)\right] \; .$$

Similarly, $\mathcal{P}_0(\boldsymbol{b}) := g_0(\boldsymbol{b}) + h_0(\boldsymbol{b})$ with $g_0(\boldsymbol{b}) := \frac{1}{2}\|\boldsymbol{W} - \boldsymbol{b}\|^2_{\boldsymbol{Q}_0}$ and $h_0(\boldsymbol{b}) := \lambda_0 \rho_{\boldsymbol{\beta}_0}(\boldsymbol{b})$, under asymptotic regime (i), or $h_0(\boldsymbol{b}) := \sigma_{N_{\operatorname{dom}(f_0)}(\boldsymbol{\beta}_0)}(\boldsymbol{b})$, under asymptotic regime (ii), respectively. By definition, $\mathcal{P}_n$ and $\mathcal{P}_0$ are strictly convex and coercive (Bauschke et al., 2016, Prop. 11.14). Moreover, the minimum of $\mathcal{P}_n$ is (almost surely) uniquely attained at $\hat{\boldsymbol{b}}_n = \sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)$. Analogously, the minimum of $\mathcal{P}_0$ is (almost surely) uniquely attained at $\operatorname{prox}^{\boldsymbol{Q}_0}_{\lambda_0 \rho_{\boldsymbol{\beta}_0}}(\boldsymbol{W})$, under asymptotic regime (i), or $\operatorname{prox}^{\boldsymbol{Q}_0}_{\sigma_{N_{\operatorname{dom}(f_0)}(\boldsymbol{\beta}_0)}}(\boldsymbol{W})$, under asymptotic regime (ii). Therefore, we only need to prove that $\mathcal{P}_n \to_d \mathcal{P}_0$ in epigraph, in order to imply the convergence:

$$\operatorname{prox}^{\boldsymbol{Q}_n}_{h_n}(\hat{\boldsymbol{b}}_n) = \operatorname*{argmin}_{\boldsymbol{b}\in\mathbb{R}^p} \mathcal{P}_n(\boldsymbol{b}) \to_d \operatorname*{argmin}_{\boldsymbol{b}\in\mathbb{R}^p} \mathcal{P}_0(\boldsymbol{b}) = \operatorname{prox}^{\boldsymbol{Q}_0}_{h_0}(\boldsymbol{W}) \; ; \tag{6.8}$$

see (Knight, 1999, Thm. 5). To prove that $\mathcal{P}_n \to_d \mathcal{P}_0$ in epigraph, note first that under Assumptions 1(i) and 1(iii), we have:

$$g_n(\boldsymbol{b}) \to_d g_0(\boldsymbol{b}) := \frac{1}{2}\|\boldsymbol{W} - \boldsymbol{b}\|^2_{\boldsymbol{Q}_0} \; , \tag{6.9}$$

uniformly on compact sets. Therefore, using (Attouch, 1984, Thm. 2.15) and Geyer (1994) we are left to prove that $h_n \to_{\Pr} h_0$ in epigraph. Consider first asymptotic regime (i). By Assumption 2 and (Attouch,

1984, Thm. 3.66), subgradient $\partial f_n(\boldsymbol{\beta}_0)$ converges in probability to subgradient $\partial f_0(\boldsymbol{\beta}_0)$ in the Painlevé-Kuratowski sense, and, equivalently, $\iota_{\partial f_n(\boldsymbol{\beta}_0)} \to_{\text{Pr}} \iota_{\partial f_0(\boldsymbol{\beta}_0)}$, in epigraph.[15] Moreover, by (Bauschke et al., 2016, Prop. 17.17), the convex conjugates of $\iota_{\partial f_n(\boldsymbol{\beta}_0)}$ and $\iota_{\partial f_0(\boldsymbol{\beta}_0)}$ equal the directional derivatives of f_n and f_0 at $\boldsymbol{\beta}_0$, respectively. Thus, under Assumption 2(ii), continuity of convex conjugation with respect to epigraph convergence implies that the directional derivative of f_n at $\boldsymbol{\beta}_0$ epiconverges in probability to $\rho_{\boldsymbol{\beta}_0}$, the directional derivative of f_0 at $\boldsymbol{\beta}_0$. From the definition of directional derivative, we therefore obtain $h_n \to_{\text{Pr}} h_0 = \lambda_0 \rho_{\boldsymbol{\beta}_0}$, in epigraph, as required. Consider next asymptotic regime (ii). Using the same arguments as for item (i), we obtain that h_n converges in epigraph to the directional derivative of $\iota_{\text{dom}(f_0)}$ at $\boldsymbol{\beta}_0$. By (Bauschke et al., 2016, Prop. 17.17), the directional derivative of $\iota_{\text{dom}(f_0)}$ at $\boldsymbol{\beta}_0$ is the convex conjugate of $\iota_{\partial \iota_{\text{dom}(f_0)}(\boldsymbol{\beta}_0)} = \iota_{N_{\text{dom}(f_0)}(\boldsymbol{\beta}_0)}$, which in turn is given by $\sigma_{N_{\text{dom}(f_0)}(\boldsymbol{\beta}_0)}$, the support function of $N_{\text{dom}(f_0)}(\boldsymbol{\beta}_0)$; see also (Bauschke et al., 2016, Example 13.3). Therefore, we obtain $h_n \to_{\text{Pr}} h_0 = \sigma_{N_{\text{dom}(f_0)}(\boldsymbol{\beta}_0)}$ in epigraph, as required. Finally, the conjugate asymptotic distribution characterization follows with Moreau's decomposition (Bauschke et al., 2016, Thm. 14.3):

$$\text{prox}_{h_0}^{\boldsymbol{Q}_0}(\boldsymbol{W}) = \left(Id - \text{prox}_{h_0^*}^{\boldsymbol{Q}_0}\right)(\boldsymbol{W}) \,,$$

where h_0^* is the convex conjugate of the directional derivative at $\boldsymbol{\beta}_0$ of function $\lambda_0 f_0$ and function $\iota_{\text{dom}(f_0)}$, under asymptotic regime (i) and (ii), respectively. Using again (Bauschke et al., 2016, Prop. 17.17), we thus obtain under asymptotic regime (i): $\text{prox}_{h_0^*}^{\boldsymbol{Q}_0} = P_{\boldsymbol{Q}_0^{-1}(\lambda_0 \partial f_0(\boldsymbol{\beta}_0))}^{\boldsymbol{Q}_0}$. Similarly, under asymptotic regimes (ii): $\text{prox}_{h_0^*}^{\boldsymbol{Q}_0} = P_{\boldsymbol{Q}_0^{-1}(N_{\text{dom}(f_0)}(\boldsymbol{\beta}_0))}^{\boldsymbol{Q}_0}$. This concludes the proof. □

Proof of Proposition 5. The definition of set $\hat{\mathcal{A}}_n$ directly gives:

$$\{\hat{\mathcal{A}}_n \cap \mathcal{A}^c \neq \emptyset\} = \bigcup_{j \in \mathcal{A}^c} \{\hat{\boldsymbol{\beta}}_{nj}^{ls} \neq (\text{prox}_{(\lambda_n f_n)^*}^{\boldsymbol{Q}_n}(\hat{\boldsymbol{\beta}}_n^{ls}))_j\} = \left(\bigcap_{j \in \mathcal{A}^c} \{\hat{\boldsymbol{\beta}}_{nj}^{ls} = (\text{prox}_{(\lambda_n f_n)^*}^{\boldsymbol{Q}_n}(\hat{\boldsymbol{\beta}}_n^{ls}))_j\}\right)^c .$$

Therefore, since $(\boldsymbol{\beta}_0)_{\mathcal{A}^c} = \mathbf{0}$:

$$\begin{aligned} \liminf_n \mathbb{P}(\hat{\mathcal{A}}_n \cap \mathcal{A}^c \neq \emptyset) &= 1 - \limsup_n \mathbb{P}\left(\sqrt{n}(\hat{\boldsymbol{\beta}}_n^{ls})_{\mathcal{A}^c} = \sqrt{n}(\text{prox}_{(\lambda_n f_n)^*}^{\boldsymbol{Q}_n}(\hat{\boldsymbol{\beta}}_n^{ls}))_{\mathcal{A}^c}\right) \\ &\geq 1 - \mathbb{P}\left((\boldsymbol{W})_{\mathcal{A}^c} = (P_{\boldsymbol{Q}_0^{-1}(B_0)}^{\boldsymbol{Q}_0}(\boldsymbol{W}))_{\mathcal{A}^c}\right) . \end{aligned}$$

[15] Mosco convergence, found in (Attouch, 1984, Thm. 3.66), is equivalent to epigraph convergence for functions in class $\Gamma(\mathbb{R}^p)$.

This concludes the proof. □

Proof of Proposition 6. Under Oracle Property 2 in Definition 1 we have $\mathrm{dom}(f_0) = \mathrm{span}\{\boldsymbol{e}_j : j \in \mathcal{A}^c\}$. To prove the first statement, consider the optimality condition defining PLSE $\hat{\boldsymbol{\beta}}_n = \mathrm{prox}^{\boldsymbol{Q}_n}_{\lambda_n f_n}(\hat{\boldsymbol{\beta}}^{ls}_n)$, which can be written as:

$$\sqrt{n}(\hat{\boldsymbol{\beta}}^{ls}_n - \hat{\boldsymbol{\beta}}_n) \in \boldsymbol{Q}_n^{-1}(\lambda_n \sqrt{n} \partial f_n(\hat{\boldsymbol{\beta}}_n)) \,. \tag{6.10}$$

It then follows, for any $j \in \mathcal{A}$:

$$\boldsymbol{e}_j' \boldsymbol{Q}_n \sqrt{n}(\hat{\boldsymbol{\beta}}^{ls}_n - \hat{\boldsymbol{\beta}}_n) \to_d \boldsymbol{e}_j' \boldsymbol{Q}_0 P^{\boldsymbol{Q}_0}_{\mathrm{span}\{\boldsymbol{e}_j : j \in \mathcal{A}\}^\perp}(\boldsymbol{W}) = 0 \,. \tag{6.11}$$

Consider now the sequence of events $\{\hat{\boldsymbol{\beta}}_n \in (\mathrm{span}\{\boldsymbol{e}_j : j \in \mathcal{A}\})^c\}$. Under condition (3.16) in Proposition 6 we then have:

$$\mathbb{P}\left(\left\{\hat{\boldsymbol{\beta}}_n \in (\mathrm{span}\{\boldsymbol{e}_j : j \in \mathcal{A}\})^c\right\} \bigcap \left\{\sqrt{n}(\hat{\boldsymbol{\beta}}^{ls}_n - \hat{\boldsymbol{\beta}}_n) \in \boldsymbol{Q}_n^{-1}(\lambda_n \sqrt{n} \partial f_n(\hat{\boldsymbol{\beta}}_n))\right\}\right) \to 0 \,. \tag{6.12}$$

Therefore, $\mathbb{P}(\hat{\boldsymbol{\beta}}_n \in \mathrm{span}\{\boldsymbol{e}_j : j \in \mathcal{A}\}) \to 1$, as $n \to \infty$. To prove the second statement, recall that under asymptotic regime (ii) in Proposition 4 the left hand side of optimality condition (6.10) is bounded in probability, while $\boldsymbol{Q}_n \to_{\mathrm{Pr}} \boldsymbol{Q}_0$. Therefore,

$$\mathbb{P}\left(\left\{\hat{\boldsymbol{\beta}}_n \in (\mathrm{span}\{\boldsymbol{e}_j : j \in \mathcal{A}\})^c\right\} \bigcap \left\{\sqrt{n}(\hat{\boldsymbol{\beta}}^{ls}_n - \hat{\boldsymbol{\beta}}_n) \in \boldsymbol{Q}_n^{-1}(\lambda_n \sqrt{n} \partial f_n(\hat{\boldsymbol{\beta}}_n))\right\}\right) \to 0 \,, \tag{6.13}$$

when Condition (3.17) holds. This concludes the proof. □

Proof of Proposition 7. The proof follows from Proposition 4, after noting that $\bar{\boldsymbol{Q}}_n \to_{\mathrm{Pr}} \bar{\boldsymbol{Q}}_0$ and $\sqrt{n}(\hat{\boldsymbol{\beta}}^{ls+}_n - \boldsymbol{\beta}^+_0) \to_d \boldsymbol{W}^+$, under the given assumptions. The latter convergence follows from the identity:

$$\sqrt{n}(\hat{\boldsymbol{\beta}}^{ls+}_n - \boldsymbol{\beta}^+_0) = -\sqrt{n}(\boldsymbol{I} - \boldsymbol{Q}_n^+ \boldsymbol{Q}_n)\boldsymbol{\beta}^+_0 + \boldsymbol{Q}_n^+(\boldsymbol{X}'\boldsymbol{\epsilon}/\sqrt{n}) \,, \tag{6.14}$$

after noting that $\boldsymbol{Q}_n^+(\boldsymbol{X}'\boldsymbol{\epsilon}/\sqrt{n}) \to_d \boldsymbol{Q}_0^+ \boldsymbol{Z} = \boldsymbol{W}^+$ and $\mathbb{E}[\boldsymbol{Q}_n] = \boldsymbol{Q}_0$. To see this, note that $\mathrm{Range}(\boldsymbol{Q}_n^+ \boldsymbol{Q}_n) = \mathrm{Range}(\boldsymbol{Q}_n \boldsymbol{Q}_n^+) = \mathrm{Range}(\boldsymbol{Q}_n)$ and $\mathrm{Range}(\boldsymbol{Q}_0^+ \boldsymbol{Q}_0) = \mathrm{Range}(\boldsymbol{Q}_0 \boldsymbol{Q}_0^+) = \mathrm{Range}(\boldsymbol{Q}_0)$, because the range of a matrix and its generalized inverse coincide. Therefore, using the results in the Appendix of Madan et al.

(1984), we obtain:

$$\mathbb{P}(\text{Range}(\boldsymbol{Q}_n^+\boldsymbol{Q}_n) = \text{Range}(\boldsymbol{Q}_0\boldsymbol{Q}_0^+)) = \mathbb{P}(\text{Range}(\boldsymbol{Q}_n) = \text{Range}(\boldsymbol{Q}_0)) \to 1 \ ,$$

as $n \to \infty$. Furthermore, recalling that $\boldsymbol{\beta}_0^+ \in \text{Range}(\boldsymbol{Q}_0)$ and that $\boldsymbol{I} - \boldsymbol{Q}_0\boldsymbol{Q}_0^+$ is the matrix of an orthogonal projection onto Kernel$(\boldsymbol{Q}_0)$:

$$\mathbb{P}((\boldsymbol{I} - \boldsymbol{Q}_n^+\boldsymbol{Q}_n)\boldsymbol{\beta}_0^+ = 0) \geq \mathbb{P}(\text{Range}(\boldsymbol{Q}_n^+\boldsymbol{Q}_n) = \text{Range}(\boldsymbol{Q}_0\boldsymbol{Q}_0^+)) \to 1 \ , \tag{6.15}$$

as $n \to \infty$. Hence, $\sqrt{n}(\boldsymbol{I} - \boldsymbol{Q}_n^+\boldsymbol{Q}_0)\boldsymbol{\beta}_0^+ = o_p(1)$ and the proof is concluded. □

Proof of Lemma 1 . Let $\tilde{\boldsymbol{Z}} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)$ and note that $\boldsymbol{Z} =_d \boldsymbol{\Omega}_0^{1/2}\tilde{\boldsymbol{Z}}$, which implies $\boldsymbol{Z} =_d P_{\text{Range}(\boldsymbol{\Omega}_0)}(\boldsymbol{Z})$. Similarly, $(\boldsymbol{Z})_{\mathcal{A}^+} =_d (\boldsymbol{\Omega}_0^{1/2}\tilde{\boldsymbol{Z}})_{\mathcal{A}^+} =_d (\boldsymbol{\Omega}_0)_{\mathcal{A}^+}^{1/2}(\tilde{\boldsymbol{Z}})_{\mathcal{A}_+}$, which implies $P_{\text{Range}((\boldsymbol{\Omega}_0)_{\mathcal{A}^+})}((\boldsymbol{Z})_{\mathcal{A}^+}) =_d (\boldsymbol{Z})_{\mathcal{A}^+}$. This proves identity (4.19). We next prove identity (4.20), which by the definition of $\bar{\boldsymbol{Q}}_0$ is satisfied if and only if:

$$(\boldsymbol{0})_{\mathcal{A}^+} = (\boldsymbol{I} - \boldsymbol{Q}_0\boldsymbol{Q}_0^+)_{\mathcal{A}^+}[(\boldsymbol{Q}_0)_{\mathcal{A}^+}]^+(\boldsymbol{Q}_0)_{\mathcal{A}^+} \ . \tag{6.16}$$

Consider next the $p \times p$ symmetric idempotent matrix $\boldsymbol{M}_{\mathcal{A}^+}$ representing the projection operator $P_{\{\boldsymbol{e}_j : j \in \mathcal{A}^+\}}$. This matrix is a diagonal matrix with ones at positions having column and row index $j \in \mathcal{A}^+$, and zeros else. Hence, $\text{Range}(\boldsymbol{M}_{\mathcal{A}^+}) = \text{span}\{\boldsymbol{e}_j : j \in \mathcal{A}^+\}$ and identity (6.16) holds if and only if:

$$\boldsymbol{0} = \boldsymbol{M}_{\mathcal{A}^+}(\boldsymbol{I} - \boldsymbol{Q}_0\boldsymbol{Q}_0^+)\boldsymbol{M}_{\mathcal{A}^+}[\boldsymbol{M}_{\mathcal{A}^+}\boldsymbol{Q}_0\boldsymbol{M}_{\mathcal{A}^+}]^+\boldsymbol{M}_{\mathcal{A}^+}\boldsymbol{Q}_0\boldsymbol{M}_{\mathcal{A}^+} \ . \tag{6.17}$$

Recalling that $\boldsymbol{I} - \boldsymbol{Q}_0\boldsymbol{Q}_0^+$ is the matrix of an orthogonal projection onto Kernel$(\boldsymbol{Q}_0)$, a sufficient condition for this last identity to hold is $\text{span}\{\boldsymbol{e}_j : j \in \mathcal{A}^+\} \subset \text{Range}(\boldsymbol{Q}_0)$. Therefore, we show that this last inclusion holds, which follows from the fact that $\boldsymbol{\beta}_0^+ \in \text{Range}(\boldsymbol{Q}_0)$, by the definition of generalized inverse $\boldsymbol{Q}_0^+$. Indeed, let $\boldsymbol{\beta}_0 = \sum_{j \in \mathcal{A}^+} \beta_{0j}\boldsymbol{e}_j$ and assume that there exists $\boldsymbol{e}_j' \notin \text{Range}(\boldsymbol{Q}_0)$ for some $j' \in \mathcal{A}^+$. Since $\beta_{0j'} \neq 0$ by the definition of set $\mathcal{A}^+$, we then obtain $\boldsymbol{\beta}_0 \notin \text{span}\{\boldsymbol{e}_j : j \neq j'\} \supset \text{Range}(\boldsymbol{Q}_0)$, i.e., a contradiction. Hence, identity (6.17) holds and the proof is concluded. □

Proof of Proposition 8. Given functions g_0 and h_0 defined as in the proof of Proposition 4, let:

$$g_n^\star(\boldsymbol{b}) := \frac{1}{2}\left\|\boldsymbol{b}_n^{ls\star} - \boldsymbol{b}\right\|_{\boldsymbol{Q}_n^\star}^2 , \tag{6.18}$$

for any $\boldsymbol{b} \in \mathbb{R}^p$. Under Assumption 4, $g_n^\star \to_d g_0$ uniformly over compact sets, conditionally on $\{(\boldsymbol{X}_i^\star, Y_i^\star) : i \in \mathbb{N}\}$. Moreover, $\phi_n^\star \in \Gamma(\mathbb{R}^p)$ and $n\lambda_n\phi_n^\star \to_{\mathrm{Pr}} h_0$ in epigraph, conditionally on $\{(\boldsymbol{X}_i^\star, Y_i^\star) : i \in \mathbb{N}\}$, because of Assumption 5. With analogous arguments as in the proof of Proposition 4, we thus obtain $\mathcal{P}_n^\star := g_n^\star + n\lambda_n\phi_n^\star \to_d g_0 + h_0 = \mathcal{P}_0$ in epigraph, and by (Knight, 1999, Thm. 5):

$$\mathrm{prox}_{n\lambda_n\phi_n^\star}^{\boldsymbol{Q}_n^\star}(\boldsymbol{b}_n^{ls\star}) = \operatorname*{argmin}_{\boldsymbol{b}\in\mathbb{R}^p} \mathcal{P}_n^\star(\boldsymbol{b}) \to_d \operatorname*{argmin}_{\boldsymbol{b}\in\mathbb{R}^p} \mathcal{P}_0(\boldsymbol{b}) = \mathrm{prox}_{h_0}^{\boldsymbol{Q}_0}(\boldsymbol{W}) ,$$

conditionally on $\{(\boldsymbol{X}_i^\star, Y_i^\star) : i \in \mathbb{N}\}$. This concludes the proof. □

Online Appendix: Additional Results

Asymptotic Adaptive Lasso functional

For the Adaptive Lasso penalty in Table 2.1, following epigraph convergence holds:

$$f_n(\boldsymbol{\beta}) = \sum_{j=1}^p \frac{|\beta_j|}{|\tilde{\beta}_{nj}|}[I(\beta_{0j} \neq 0) + I(\beta_{0j} = 0)] \to_{\mathrm{Pr}} f_0(\boldsymbol{\beta}) = \sum_{j=1}^p \left[\frac{|\beta_j|}{|\beta_{0j}|} I(\beta_{0j} \neq 0) + \iota_{\{0\}}(\beta_j) I(\beta_{0j} = 0)\right] .$$

The epigraph limit in probability of the corresponding convex conjugates directly follows from the convex conjugate $f_0^*(\boldsymbol{\theta}/\lambda_0) = \iota_{\boldsymbol{Q}_0^{-1}(B_0(\lambda_0))}(\boldsymbol{\theta})$, where:

$$B_0(\lambda_0) := \bigcap_{\{j:\beta_{0j}\neq 0\}} \left\{\boldsymbol{\theta} : |\theta_j| \leq \lambda_0/|\beta_{0j}|\right\} .$$

For any penalty parameter $\lambda_0 > 0$, the resulting bias functional reads explicitly:

$$-\mathrm{prox}_{(\lambda_0 f_0)^*}^{\boldsymbol{Q}_0}(\boldsymbol{\beta}_0) = -\operatorname*{argmin}_{\boldsymbol{\theta}\in\mathbb{R}^P} \left\{\frac{1}{2}\|\boldsymbol{\beta}_0 - \boldsymbol{\theta}\|_{\boldsymbol{Q}_0}^2 : \boldsymbol{Q}_0^{-1}(B(\lambda_0))\right\} ,$$

with the polyhedron:

$$\boldsymbol{Q}_0^{-1}(B(\lambda_0)) = \left(\bigcap_{\{j:\beta_{0j}\neq 0\}} \{\boldsymbol{\theta} : |\langle \boldsymbol{e}_j, \boldsymbol{\theta}\rangle_{\boldsymbol{Q}_0}| \leq \lambda_0/|\beta_{0j}|\} \right) .$$

Therefore, $\operatorname{prox}^{\boldsymbol{Q}_0}_{(\lambda_0 f_0)^*}$ equals a projection operator $P^{\boldsymbol{Q}_0}_{\boldsymbol{Q}_0^{-1}(B(\lambda_0))}$ and the asymptotic Adaptive Lasso functional is reproduced by a projection residual of the asymptotic LSE functional:

$$\operatorname{prox}^{\boldsymbol{Q}_0}_{\lambda_0 f_0}(\boldsymbol{\beta}_0) = \left(Id - P^{\boldsymbol{Q}_0}_{\boldsymbol{Q}_0^{-1}(B(\lambda_0))} \right)(\boldsymbol{\beta}_0) .$$

If additionally $\lambda_n \to \lambda_0 = 0$ and $\lambda_n n^\gamma \to +\infty$, following epigraph convergence holds:

$$\lambda_n f_n(\boldsymbol{\beta}) \to_{\Pr} \iota_{\operatorname{dom}(f_0)}(\boldsymbol{\beta}) = \sum_{j=1}^{p} \iota_{\{0\}}(\beta_j) I(\beta_{0j} = 0) ,$$

and $(\iota_{\operatorname{dom}(f_0)})^* = \sigma_{\operatorname{dom}(f_0)} = \iota_{B_0}$, with the subspace:

$$B_0 := B_0(0) = \bigcap_{\{j:\beta_{0j}\neq 0\}} \{\boldsymbol{\theta} : \theta_j = 0\} . \tag{6.19}$$

The corresponding bias functional is given by:

$$-\operatorname{prox}^{\boldsymbol{Q}_0}_{(\iota_{\operatorname{dom}(f_0)})^*}(\boldsymbol{\beta}_0) = -\operatorname*{argmin}_{\boldsymbol{\theta}\in\mathbb{R}^P} \left\{ \frac{1}{2}\|\boldsymbol{\beta}_0 - \boldsymbol{\theta}\|^2_{\boldsymbol{Q}_0} \ : \ \boldsymbol{\theta} \in \boldsymbol{Q}_0^{-1}(B_0) \right\} ,$$

with the subspace:[16]

$$\boldsymbol{Q}_0^{-1}(B_0) = \bigcap_{\{j:\beta_{0j}\neq 0\}} \{\boldsymbol{\theta} : \langle \boldsymbol{e}_j, \boldsymbol{\theta}\rangle_{\boldsymbol{Q}_0} = 0\} = \operatorname{span}\{\boldsymbol{e}_j : \beta_{0j} \neq 0\}^\perp .$$

In summary, we obtain:

$$\operatorname{prox}^{\boldsymbol{Q}_0}_{\iota_{\operatorname{dom}(f_0)}}(\boldsymbol{\beta}_0) = \left(Id - P^{\boldsymbol{Q}_0}_{\operatorname{span}\{\boldsymbol{e}_j:\beta_{0j}\neq 0\}^\perp} \right)(\boldsymbol{\beta}_0) = P^{\boldsymbol{Q}_0}_{\operatorname{span}\{\boldsymbol{e}_j:\beta_{0j}\neq 0\}}(\boldsymbol{\beta}_0) ,$$

[16] Orthogonal complements are computed here under inner product $\langle\cdot,\cdot\rangle_{\boldsymbol{Q}_0}$.

i.e., the asymptotic Adaptive Lasso functional is an orthogonal projection (under inner product $\langle \cdot, \cdot \rangle_{\boldsymbol{Q}_0}$) of the asymptotic LSE functional on the subspace generated by canonical basis vectors indexed by the nonzero components of the asymptotic LSE.

Asymptotic Distribution of benchmark PLSEs

Example 1. Consider the constrained LSE with penalty function $f_n = \iota_C$. The asymptotic distribution of this estimator follows in closed-form, with asymptotic regime (i) in Proposition 4, by setting without loss of generality $\lambda_n \sqrt{n} = 1$ and recalling that $\rho_{\boldsymbol{\beta}_0}(\boldsymbol{b}) = \sigma_{N_C(\boldsymbol{\beta}_0)}(\boldsymbol{b})$ (see again Table 3.1). Conversely, $\partial f(\boldsymbol{\beta}_0) = N_C(\boldsymbol{\beta}_0)$ and $\lambda_0 = 1$. Thus, asymptotic normality follows when $N_C(\boldsymbol{\beta}_0)$ is affine. This case arises when constraint set C itself is affine, since then $N_C(\boldsymbol{\beta}_0) = (C - C)^{\perp}$ by (Bauschke et al., 2016, Ex. 6.42), provided that $\boldsymbol{\beta}_0 \in C$, or when $\boldsymbol{\beta}_0$ belongs to the interior of C, by (Bauschke et al., 2016, Cor. 6.44), in which case PLSE (2.1) and LSE (1.2) are obviously asymptotically equivalent. Closed-form conjugate asymptotic distribution expressions also arise in settings where constraint set C is a convex cone, since then $N_C(\boldsymbol{\beta}_0) = C^{\ominus} \cap \{\boldsymbol{\beta}_0\}^{\perp}$, where $C^{\ominus}$ is the polar cone of C, by (Bauschke et al., 2016, Ex. 6.39).

Oracle properties of benchmark PLSEs

Example 2. Consider the Lasso estimator in Proposition 2. Recalling that $\mathrm{dom}(f_0) = \mathbb{R}^p$, we have $B_0 = N_{\mathrm{dom}(f_0)}(\boldsymbol{\beta}_0) = \{\mathbf{0}\}$ under asymptotic regime (ii) in Proposition 4. Under asymptotic (i) in Proposition 4, we instead have:

$$B_0 = \left(\bigcap_{\{j : \beta_{0j} \neq 0\}} \{\boldsymbol{t} \, : \, t_j = \lambda_0 \, \mathrm{sign}(\beta_{0j})\} \right) \bigcap \left(\bigcap_{\{j : \beta_{0j} = 0\}} \{\boldsymbol{t} \, : \, t_j \in [-\lambda_0, \lambda_0]\} \right) .$$

Overall, we thus have under asymptotic regime (ii):

$$\mathbb{P}\left((\boldsymbol{W})_{\mathcal{A}^c} = (P^{\boldsymbol{Q}_0}_{\boldsymbol{Q}_0^{-1}(B_0)}(\boldsymbol{W}))_{\mathcal{A}^c} \right) = \mathbb{P}\left((\boldsymbol{W})_{\mathcal{A}^c} = (P^{\boldsymbol{Q}_0}_{\{\mathbf{0}\}}(\boldsymbol{W}))_{\mathcal{A}^c} \right) = 0 \, .$$

Similarly, under asymptotic regime (i) the boundedness of set $\boldsymbol{Q}_0^{-1}(B_0)$ yields:

$$\mathbb{P}\left((\boldsymbol{W})_{\mathcal{A}^c} = (P^{\boldsymbol{Q}_0}_{\boldsymbol{Q}_0^{-1}(B_0)}(\boldsymbol{W}))_{\mathcal{A}^c} \right) < 1 \, .$$

Using Proposition 5, this shows that the Lasso estimator does not satisfy Property 1 in Definition 1, under both asymptotic regimes (i) and (ii) from Proposition 4.

Valid and invalid bootstrap approximations for benchmark PLSEs

This section borrows from existing bootstrap proposals for PLSEs with regular designs to cover benchmark examples of valid and invalid bootstrap approximations for PLSEs with singular designs.

Example 3. Consider the Lasso limit penalty under asymptotic regime (i) in Proposition 7:

$$\lambda_0 \rho_{\boldsymbol{\beta}_0^+}(\boldsymbol{b}) = \lambda_0 \sum_j \left[b_j \operatorname{sign}(\beta_{0j}^+) I_{\{\beta_{0j}^+ \neq 0\}} + |b_j| I_{\{\beta_{0j}^+ = 0\}} \right] , \tag{6.20}$$

and following penalty function for bootstrap PLSE (5.2):

$$\sqrt{n}\phi_n^\star(\boldsymbol{b}) := \sqrt{n} \sum_{j=1}^p \left[\left| \hat{\beta}_{nj}^{ls+} I_{\left\{|\hat{\beta}_{nj}^{ls+}| > a_n\right\}} + b_j/\sqrt{n} \right| - \left| \hat{\beta}_{nj}^{ls+} I_{\left\{|\hat{\beta}_{nj}^{ls+}| > a_n\right\}} \right| \right] , \tag{6.21}$$

for deterministic thresholds a_n such that $a_n + (n^{-1/2} \ln n) a_n^{-1} \to 0$ as $n \to \infty$; cf. Chatterjee and Lahiri (2011). It then follows: $\sqrt{n}\phi_n^\star \to_{\Pr} \rho_{\boldsymbol{\beta}_0^+}$ in epigraph. Hence, bootstrap PLSE (5.2) gives rise to asymptotically valid bootstrap approximations.[17]

Example 4. Recall the constrained Least Squares setting in Example 1, with a limit penalty $\lambda_0 \rho_{\boldsymbol{\beta}_0^+} = \sigma_{N_C(\boldsymbol{\beta}_0^+)}$. Here, one may try to estimate $\sigma_{N_C(\boldsymbol{\beta}_0^+)}$ using penalties $\sqrt{n}\phi_n^\star = \sigma_{N_C(\hat{\boldsymbol{\beta}}_n^{ls+})}$, since $\hat{\boldsymbol{\beta}}_n^{ls+}$ is a consistent estimator of $\boldsymbol{\beta}_0^+$. However, this approach fails because this sequence of penalties does not converge to the correct limit penalty when $\boldsymbol{\beta}_0^+$ is on the boundary of the admissible parameter set. Consider for instance a scalar parameter space with constraint set $C = \mathbb{R}_+$ as in Andrews (2000), with the closed-form normal cone (see again (Bauschke et al., 2016, Ex. 6.39)):

$$N_C(\beta_0^+) = \begin{cases} \{0\} & \text{if } \beta_0^+ > 0 \\ -\mathbb{R}_+ & \text{if } \beta_0^+ = 0 \end{cases} .$$

It then follows, $\sigma_{N_C(\beta_0^+)} = \iota_{\mathbb{R}_+} I_{\{\beta_0^+ = 0\}}$, while $\sigma_{N_C(\hat{\beta}_n^{ls+})} = \iota_{\mathbb{R}_+} I_{\{\hat{\beta}_n^{ls+} = 0\}} \to_{\Pr} 0$, in epigraph. On the other

[17] As stressed in Chatterjee and Lahiri (2011), the truncation implied by the indicator function in penalty (6.21) is key to obtain epigraph convergence in probability, as otherwise the second term in the sum within the square brackets of limit penalty (6.20) is not estimated consistently when some components of $\boldsymbol{\beta}_0^+$ are zero.

hand, following convergence holds, using the sequence of thresholds a_n from Example 3:

$$\sqrt{n}\phi_n^\star := \iota_{\mathbb{R}_+} I_{\{|\hat{\beta}_n^{ls+}|<a_n\}} \to_{\Pr} \iota_{\mathbb{R}_+} I_{\{\beta_0^+=0\}} \,, \tag{6.22}$$

in epigraph. Hence while penalty sequence (6.22) gives rise to asymptotically valid bootstrap approximations based on bootstrap PLSE (5.2), a penalty $\sqrt{n}\phi_n^\star = \sigma_{N_C(\hat{\beta}_n^{ls+})}$ gives rise to a bootstrap failure.

Fenchel-Young loss in equation (5.7) of the main text

The following Lemma provides the proof for closed-form convex conjugate expression (5.7) in the main text.

Lemma 2. *Consider the Adaptive lasso bootstrap penalty:*

$$\phi_n^\star(\boldsymbol{b}) := \sum_{j=1}^{p} \frac{1}{\left|\hat{\beta}_{nj}^{ls+\star}\right|} \left[\left| \hat{\beta}_{nj}^{ls+} I_{\left\{|\hat{\beta}_{nj}|>0\right\}} + b_j/\sqrt{n} \right| - \left| \hat{\beta}_{nj}^{ls+} I_{\left\{|\hat{\beta}_{nj}|>0\right\}} \right| \right] \,. \tag{6.23}$$

It then follows:

$$(\phi_n^\star)^*(\boldsymbol{t}/(n\lambda_n)) = \sum_{j=1}^{p} FYL\left(|\hat{\beta}_{nj}^{ls+}/\hat{\beta}_{nj}^{ls+\star}| I_{\left\{|\hat{\beta}_{nj}|>0\right\}}, t_j|\hat{\beta}_{nj}^{ls+\star}|/(\lambda_n\sqrt{n}) \right) \,, \tag{6.24}$$

with a Fenchel-Young loss $FYL : \mathbb{R}^2 \to \mathbb{R}_+$ given by:

$$FYL\,(b,t) = |b| + \iota_{[-1,1]}(t) - bt \,. \tag{6.25}$$

Proof. We first have $(\phi_n^\star)^*(\boldsymbol{t}) = \sum_{j=1}^{p} g_j^*(t_j)$, where:

$$g_j(b_j) = \frac{1}{\left|\hat{\beta}_{nj}^{ls+\star}\right|} \left[\left| \hat{\beta}_{nj}^{ls+} I_{\left\{|\hat{\beta}_{nj}|>0\right\}} + b_j/\sqrt{n} \right| - \left| \hat{\beta}_{nj}^{ls+} I_{\left\{|\hat{\beta}_{nj}|>0\right\}} \right| \right] \,. \tag{6.26}$$

Therefore,

$$\begin{aligned} g_j^*(t_j) &= \frac{1}{\left|\hat\beta_{nj}^{ls+\star}\right|}\left[\left|\hat\beta_{nj}^{ls+} I_{\left\{|\hat\beta_{nj}|>0\right\}}\right| + \left|\hat\beta_{nj}^{ls+} I_{\left\{|\hat\beta_{nj}|>0\right\}} + \cdot/\sqrt{n}\right|^* (t_j \left|\hat\beta_{nj}^{ls+\star}\right|)\right] \\ &= \frac{1}{\left|\hat\beta_{nj}^{ls+\star}\right|}\left[\left|\hat\beta_{nj}^{ls+} I_{\left\{|\hat\beta_{nj}|>0\right\}}\right| - \hat\beta_{nj}^{ls+} I_{\left\{|\hat\beta_{nj}|>0\right\}} t_j\sqrt{n}\left|\hat\beta_{nj}^{ls+\star}\right| + |\cdot|^* (t_j\sqrt{n}\left|\hat\beta_{nj}^{ls+\star}\right|)\right] \\ &= \frac{1}{\left|\hat\beta_{nj}^{ls+\star}\right|}\left[\left|\hat\beta_{nj}^{ls+} I_{\left\{|\hat\beta_{nj}|>0\right\}}\right| - \hat\beta_{nj}^{ls+} I_{\left\{|\hat\beta_{nj}|>0\right\}} t_j\sqrt{n}\left|\hat\beta_{nj}^{ls+\star}\right| + \iota_{[-1,1]}(t_j\sqrt{n}\left|\hat\beta_{nj}^{ls+\star}\right|)\right] \\ &= \frac{1}{\left|\hat\beta_{nj}^{ls+\star}\right|}\left|\hat\beta_{nj}^{ls+} I_{\left\{|\hat\beta_{nj}|>0\right\}}\right| - \hat\beta_{nj}^{ls+} I_{\left\{|\hat\beta_{nj}|>0\right\}} t_j\sqrt{n} + \iota_{\left[-1/\left|\hat\beta_{nj}^{ls+\star}\right|,1/\left|\hat\beta_{nj}^{ls+\star}\right|\right]}(t_j\sqrt{n}) \,. \end{aligned}$$

Hence:

$$g_j^*(t_j/(n\lambda_n)) = \frac{1}{\left|\hat\beta_{nj}^{ls+\star}\right|}\left|\hat\beta_{nj}^{ls+} I_{\left\{|\hat\beta_{nj}|>0\right\}}\right| - \hat\beta_{nj}^{ls+} I_{\left\{|\hat\beta_{nj}|>0\right\}} t_j/(\lambda_n\sqrt{n}) + \iota_{\left[-1/\left|\hat\beta_{nj}^{ls+\star}\right|,1/\left|\hat\beta_{nj}^{ls+\star}\right|\right]}(t_j/(\lambda_n\sqrt{n})) \,.$$

After collecting all terms $g_j^*(t_j/(n\lambda_n))$ into sum $\sum_{j=1}^p g_j^*(t_j/(n\lambda_n))$, formula (6.24) follows. □

Part 2: Stochastic Discount Factors

Smart Stochastic Discount Factors

SOFONIAS ALEMU KORSAYE, ALBERTO QUAINI and FABIO TROJANI*

First version: May 2019. This version: November 12, 2021

Abstract

We propose a novel no-arbitrage framework, which exploits convex asset pricing constraints to study investors' marginal utility of wealth or, more generally, Stochastic Discount Factors (SDFs). We establish a duality between minimum dispersion SDFs and penalized portfolio selection problems, building the foundation for characterizing the feasible tradeoffs between a SDF's pricing accuracy and its comovement with systematic risks. Empirically, a minimum variance CAPM–SDF produces a Pareto optimal tradeoff. This SDF only depends on two distinct risk factors: A traded market factor and a minimum variance excess return that bounds the mispricing of risks unspanned by market shocks.

Keywords: SDF, Convex Pricing Constraints, Minimum Dispersion SDF, Market Frictions, SDF regularization, Arbitrage Pricing Theory,

*Sofonias Alemu Korsaye (email: Sofonias.Korsaye@unige.ch) and Fabio Trojani (email: Fabio.Trojani@unige.ch) are with University of Geneva, Geneva Finance Research Institute & Swiss Finance Institute. Alberto Quaini (email: Alberto.Quaini@unige.ch) is with University of Geneva, Geneva Finance Research Institute. We thank Caio Almeida, Federico Bandi, Tony Berrada, Federico Carlini, Ines Chaieb, George Constantinides, Jerome Detemple, Patrick Gagliardini, Eric Ghysels, Lars Hansen, Oliver Linton, Stefan Nagel, Paolo Porchia, Olivier Scaillet, Paul Schneider, Andrea Vedolin, Michael Weber, Dacheng Xiu, conference participants at the 2019 SoFiE annual Meeting in Shanghai, the European Econometric Society Meeting in Manchester, the Workshop on Big Data and Economic Forecasting in Ispra, the Conference on Quantitative Finance and Financial Econometrics in Marseille, the Financial Econometrics Conference in Toulouse, the Vienna Congress on Mathematical Finance, the Swiss Finance Institute Research days, the ESSEC Workshop on Monte Carlo Methods and Approximate Dynamic Programming, the Paris December Finance Meeting, the 2020 Virtual World Congress of the Econometric Society in Milano, the Remote Seminar Series on Computational Economics and Finance, the Virtual Derivatives Workshop and seminar participants at BI Norwegian Business School, Bocconi University, Mc Gill University, University of Zurich, University of Lugano, University of Geneva, University of Lund, JRC in Ispra and Luiss University. All errors are ours.

The absence of arbitrage opportunities is equivalent to the existence of a Stochastic Discount Factor (SDF), which can be interpreted as a proxy for the marginal utility of wealth in asset markets. Since SDFs are key for understanding asset prices, they have been the object of a vast amount of research, which builds on various ex-ante hypotheses regarding the SDF pricing features or, equivalently, the underlying market structure. This paper focuses on the SDF properties that can be learned from asset prices, based exclusively on the information conveyed by a set of convex asset pricing constraints in an arbitrage-free market.

We first establish the arbitrage-free foundation of SDFs defined by general convex pricing constraints and provide a novel unifying framework for analysing and selecting these SDFs. Importantly, some assets in our approach may only satisfy a set of convex pricing constraints with respect to a given SDF, e.g., because they may involve transaction costs in partially liquid markets or because their payoffs may reflect ambiguous risk exposures with uncertain compensations. Other assets may instead be (or be not) priced exactly by the same SDF, e.g., assets involving negligible transaction costs and corresponding to unambiguous systematic risk exposures. We call such SDFs Smart SDFs (S–SDFs), because of their affinity to the motivation of smart beta portfolios in modern asset management.[1]

In arbitrage-free economies with market frictions, S–SDFs represent a positive linear pricing rule that bounds from below the nonlinear pricing rule of such markets. In frictionless arbitrage-free economies with ambiguity, S–SDFs represent the positive linear pricing rule of an unobserved marginal investor. This foundation provides a novel framework for several important asset pricing theories, such as Ross [1976]' Arbitrage Pricing Theory (APT), the good-deal bound SDFs in Cochrane and Saa-Requejo [2000], various approaches to SDF regularization such as Kozak et al. [2020], and robust approaches for the identification of investors' beliefs such as Chen et al. [2020].

[1] Assets that are often assumed exactly priced for empirical asset pricing purposes are tradeable portfolios replicating established systematic risk exposures in the literature, such as market risk or other exposures like size and value, among others.

Our analysis and selection framework for S–SDFs is based on minimum dispersion S–SDF problems subject to convex pricing constraints. We establish a duality between minimum dispersion S–SDF problems and penalized dual portfolio selection problems, in which the penalization choice stays in a one-to-one relation with the initial pricing constraints. Since minimum dispersion S–SDFs are given by simple transformations of dual optimal portfolio payoffs, we find that they are naturally interpretable as regularized SDFs in an arbitrage-free economy with frictions or ambiguous asset payoffs.

Our framework gives rise to a novel methodology for characterizing key S–SDF properties, based on the attainable tradeoffs between pricing accuracy and S–SDF comovement with systematic asset return risks. This framework relies on a broad family of minimum variance S–SDFs that satisfy convenient asset pricing bounds consistent with the APT. Using cross-sections of sorted portfolio returns of varying dimensions, we take our theoretical APT S–SDF framework to market data. We find that a minimum variance correction of a CAPM–SDF, in which market risk alone is priced exactly and the mispricing of risks orthogonal to market risk is bounded according to APT pricing constraints, consistently produces a Pareto optimal tradeoff. In contrast, S–SDFs forcing tradability of additional systematic risk exposures uniformly imply a suboptimal tradeoff.

Our minimum variance Pareto optimal S–SDF corrections are founded in arbitrage-free frictionless markets with ambiguity and feature a particularly simple structure, which only depends on two economically interpretable asset pricing factors. The first factor is naturally a traded excess return that strongly correlates with pervasive asset return risks. The second factor is a minimum variance portfolio excess return, which bounds the mispricing of excess returns unspanned by the traded factor. This evidence demonstrates that a convenient family of minimum variance S–SDFs, which depend on just two economically interpretable asset pricing factors and are founded in corresponding arbitrage-free markets, attains a robust Pareto optimal tradeoff between pricing accuracy and S–SDF comovement with systematic asset return risks. This optimal tradeoff provides a natural

benchmark for assessing also empirical S–SDF specifications not included in our family of minimum variance S–SDFs.

The rest of the paper proceeds as follows. After a review of the related literature, in Section 2, we define S–SDFs and prove their existence in arbitrage-free economies. We then introduce minimum dispersion S–SDF problems, derive their dual penalized portfolio problems and show how optimal S–SDFs arise as simple transformations of the optimal penalized portfolio payoff. In Section 3, we revisit various asset pricing theories and settings that can be founded and understood within our S–SDF framework, while Section 4 explicitly applies our theory to specify a family of model-free asset pricing relations consistent with the APT. The empirical evidence on APT S–SDFs regarding the tradeoffs between cross-sectional pricing accuracy and S–SDF comovement with systematic risks, together with the logic behind different S–SDFs specification choices, is collected in Section 5. Section 6 concludes, while an Online Appendix collects various auxiliary results, additional empirical evidence, and the proofs of the formal statements in the main text.

1 Related Literature

Our paper is related to various important strands of the literature. A first strand studies the no-arbitrage foundation of strictly positive SDFs with various versions of the fundamental theorem of asset pricing. Ross [1978], Harrison and Kreps [1979], and Jouini and Kallal [1995] prove early versions of the fundamental theorem for frictionless economies and for economies with sublinear transaction costs, respectively. In our arbitrage-free foundation of S–SDFs, we borrow from the market structure in Jouini and Kallal [1995] and build a class of suitable economies, in which a strictly positive S–SDF exists if and only if markets are arbitrage-free. This result provides a general foundation for S–SDFs consistent with convex pricing constraints, with distinct economic interpretations in corresponding arbitrage-free economies with frictions or ambiguity.

A second prominent direction builds on the no-arbitrage foundation in Harrison and

Kreps [1979] and studies minimum dispersion SDFs and SDF bounds under exact pricing constraints. First introduced by Hansen and Jagannathan [1991] as key objects to compute minimum variance bounds giving rise to model-free diagnostics for asset pricing models, they have been exploited in Snow [1991], Stutzer [1995], Bansal and Lehmann [1997], Alvarez and Jermann [2005], Backus et al. [2014], and Almeida and Garcia [2016], among others, to develop further model diagnostics relying on notions of SDF dispersion different from variance. Luttmer [1996] is the first to consider minimum variance SDFs with conic pricing constraints originating from market frictions such as short-sale constraints or bid-ask spreads. Our framework substantially extends this literature to general S–SDF settings with convex pricing constraints, which allows us to embed in a single unifying framework many asset pricing approaches that are not covered by this literature.

Our S–SDF framework provides a distinct arbitrage-free foundation for various directions in the literature. We show that the good-deal bound SDFs in Cochrane and Saa-Requejo [2000] and the robust SDFs in Kozak et al. [2020] are minimum variance S–SDFs corresponding to a particular choice of convex pricing constraints, while the robust investor belief identified in Chen et al. [2020] is interpretable as a S–SDF minimizing a particular notion of stochastic dispersion relative to a benchmark asset pricing model. We further embed the predictions of Ross [1976] APT and its version with misspecification in Raponi et al. [2018] and Uppal et al. [2019] in our framework using a corresponding family of APT S–SDFs, which give rise to useful model-free interpretations of the APT predictions and new ways to implement them empirically.[2] We show that such APT S–SDFs optimally capture the tradeoff between S–SDF pricing accuracy and the S–SDF ability to span systematic asset return risks.

Our approach is linked also to recent directions applying penalization techniques from

[2]The formalizations of the APT in Hubermann [1982], Chamberlain [1983], Chamberlain and Rothschild [1983], Ingersoll [1984], and Uppal et al. [2019] are based on an asymptotic no-arbitrage condition that gives rise to bounded convex pricing constraints under a quadratic APT metric. This feature allows us to construct a family of minimum variance APT S–SDFs replicating the APT asset pricing predictions, under varying assumptions on the set of tradeable systematic risk factors and the S–SDF sparsity features.

the machine learning literature in empirical asset pricing. Freyberger et al. [2020] explain expected returns with a small set of characteristics based on a nonparametric specification of dependencies between expected returns and characteristics. Gu et al. [2020b] find evidence that machine learning methods allowing for nonlinear predictive relations outperform in out-of-sample return prediction. Feng et al. [2020] propose a robust model-selection method for evaluating the contribution of new factors to an unknown SDF. Different from this literature, we propose a unifying framework that directly constructs minimum dispersion S–SDFs with convex pricing properties from a cross-section of asset returns, thereby coherently addressing also settings where expected excess returns may not be fully explained by a covariation with a SDF.

A more recent literature studies empirical SDFs with improved out-of-sample pricing properties. Kozak et al. [2020] introduce versions of such SDFs, using lasso, ridge, and elastic net penalizations. We show that the population versions of these SDFs are interpretable as minimum dispersion S–SDFs and obtain in closed-form their hidden pricing features. Given such pricing features, we show that a Pareto optimal tradeoff between pricing accuracy and SDF comovement with systematic return risks is attained by minimum variance S–SDFs that depend on the excess returns of just two economically interpretable portfolios. Our framework provides a useful starting point also for empirical SDFs aiming to incorporate conditional information from a large set of conditioning variables, such as Gu et al. [2020a]. A common assumption in these approaches is existence of a SDF exactly pricing all assets. Distinctions between these approaches arise in the way how they regularize the very high-dimensional unconditional pricing constraints that need to hold theoretically. Here, our framework can offer guidelines for interpreting a penalization choice in terms of the resulting pricing constraints of a corresponding S–SDF, which may exist in markets of growing dimension when a SDF may not.

2 Theory of Smart Stochastic Discount Factors

Denote by $L^p(\mathbb{P})$ $(p > 1)$ the payoff space of random variables defined on probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and having finite $p-$th moment, by $\mathcal{Z} \subset L^p(\mathbb{P})$ a market of traded payoffs, and by $\pi : \mathcal{Z} \to \mathbb{R}$ a pricing functional associating a price to each traded payoff.[3] We denote the resulting economy, which may or may not be completely observed, by the triplet $(\mathcal{Z}, \pi, \mathbb{P})$. In such economy, we finally let $\boldsymbol{X} = (X_n)_{n=1}^N$ be a finite observed subset of N traded payoffs, with associated price vector $\boldsymbol{P} \in \mathbb{R}^N$.

2.1 S–SDF definition and basic properties

For any set of convex pricing constraints on payoff vector $\boldsymbol{X}$, we study the arbitrage-free market structures $(\mathcal{Z}, \pi, \mathbb{P})$ and corresponding stochastic discount factors compatible with such constraints. Without loss of generality, we specify convex pricing constraints on payoff vector $\boldsymbol{X}$ with a closed convex set $C \subset \mathbb{R}^N$ containing $\mathbf{0}$. This gives following definition of a smart stochastic discount factor.

Definition 1 (Smart Stochastic Discount Factors). Given constraint set C, a Smart Stochastic Discount Factor (S–SDF) is an element of following family of random variables:

$$\mathcal{M}_+(C) := \{M \geq 0 \ : \ \mathbb{E}[M\boldsymbol{X}] - \boldsymbol{P} \in C\} \cap L^q(\mathbb{P}) \ , \tag{1}$$

where $q := p/(p-1)$.

In Definition 1, positive linear pricing functional $\mathbb{E}[M\cdot]$ gives rise to asset valuations constrained by set $\boldsymbol{P} + C := \{\boldsymbol{P} + \boldsymbol{\eta} : \boldsymbol{\eta} \in C\}$. Intuitively, set C fixes the geometry of the discrepancies between vector $\boldsymbol{P}$ and the vector of S–SDF asset valuations. It thus crucially determines key S–SDF properties, such as the structure of the resulting asset

[3] We consider power $p \notin \{1, \infty\}$ in order to obtain compact proofs based on $L^p - L^q$ duality, with $1/p + 1/q = 1$. In this setting, we also apply the standard convention that equalities, inequalities and strict inequalities are understood $\mathbb{P}-$almost surely.

excess returns. In general, following S–SDF expected excess return identity holds for any asset payoff:

$$\mathbb{E}[X_n] - \frac{P_n}{\mathbb{E}[M]} = -Cov\left[\frac{M}{\mathbb{E}[M]}, X_n\right] + \frac{\mathbb{E}[MX_n] - P_n}{\mathbb{E}[M]} \ . \tag{2}$$

Therefore, the choice of constraint set C directly determines the way how a S–SDF explains asset excess returns under a given reference probability $\mathbb{P}$. For instance, while S–SDFs completely explain the excess return of exactly valuated assets by their covariance with S–SDF risk, only a part of the expected excess return of assets that are not exactly valuated is explained by an exposure to S–SDF risk.

2.2 Existence of S–SDFs

Given price vector $\boldsymbol{P}$ and constraint set C, a key question is when does a S–SDF exist. We answer this question by explicitly constructing an arbitrage-free economy $(\mathcal{Z}_C, \pi_C, \mathbb{P})$ in which the S–SDF existence is granted. In such economy, investors can form portfolios $\boldsymbol{\theta} \in \mathbb{R}^N$ of payoffs $\boldsymbol{X}$, while incurring costs for including assets in a portfolio. These costs are described by a cost function defined by:

$$\sigma_C(\boldsymbol{\theta}) := \sup\{\boldsymbol{\theta}'\boldsymbol{\eta} : \boldsymbol{\eta} \in C\} \geq 0 \ . \tag{3}$$

Traded payoffs are those that can be replicated with finite costs:

$$\mathcal{Z}_C := \{Z = \boldsymbol{X}'\boldsymbol{\theta} \ : \ \sigma_C(\boldsymbol{\theta}) < +\infty\} \ . \tag{4}$$

Finally, the price of a traded payoff is the lowest total replication cost across payoff replicating portfolios:

$$\pi_C(Z) := \inf_{\boldsymbol{\theta} \in \mathbb{R}^N} \{\boldsymbol{P}'\boldsymbol{\theta} \ + \sigma_C(\boldsymbol{\theta}) \ : \ Z = \boldsymbol{X}'\boldsymbol{\theta}\} \ . \tag{5}$$

Definitions (4)–(5) give rise to an explicit market structure $(\mathcal{Z}_C, \pi_C, \mathbb{P})$ embedding traded payoff vector $\boldsymbol{X}$. As the next remark underscores, such a market structure is general and

uniquely determined by constraint set C.

Remark 1. By definition, cost function (3) is closed and sublinear.[4] Conversely, any closed sublinear cost function σ can be uniquely written in the form (3) for a corresponding closed convex set $C := C_\sigma$, given explicitly by (see [Hiriart-Urruty and Lemaréchal, 2012, Thm. 3.1.1]):

$$C_\sigma := \{\boldsymbol{\eta} \in \mathbb{R}^N \ : \ \boldsymbol{\eta}'\boldsymbol{\theta} \leq \sigma(\boldsymbol{\theta}) \text{ for all } \boldsymbol{\theta} \in \mathbb{R}^N\} \ . \tag{6}$$

The uniquely associated economy (4)–(5) to this cost function is $(\mathcal{Z}_{C_\sigma}, \pi_{C_\sigma}, \mathbb{P})$. In such economies, differences between $\pi_C(X_n)$ and P_n may arise when cost function σ_C does not vanish on the unit vectors in $\mathbb{R}^N$, and they are uniquely determined by the properties of constraint set C.

The next proposition characterizes the existence of a strictly positive S–SDF by the absence of arbitrage in market $(\mathcal{Z}_C, \pi_C, \mathbb{P})$, where arbitrage opportunities are defined as free lunches following Harrison and Kreps [1979].[5]

Proposition 1. *Market $(\mathcal{Z}_C, \pi_C, \mathbb{P})$ admits no free lunches if and only there exists $M \in \mathcal{M}_+(C)$ such that $M > 0$.*

From Proposition 1, a strictly positive S–SDF exists if and only if market $(\mathcal{Z}_C, \pi_C, \mathbb{P})$ is arbitrage-free. Together with Remark 1, this provides a general arbitrage-free foundation for S–SDFs.

Remark 2. Proposition 1 in Online Appendix B.4 proves a slightly more general version of Proposition 1 for a market structure (4)–(5) with closed convex (not necessarily sublinear) cost function σ. In this context, existence of a S–SDF is equivalent to the viability of market $(\mathcal{Z}_{C_\sigma}, \pi_{C_\sigma}, \mathbb{P})$, which is in general a stronger no arbitrage condition than the absence of free lunches.[6]

[4] A mapping $f : V \to \mathbb{R}$ is closed if sublevel set $\{x \in V : f(x) \leq \alpha\}$ is closed for any $\alpha \in \mathbb{R}$. It is sublinear if $f(\alpha x) = \alpha f(x)$ and $f(x + y) \leq f(x) + f(y)$ for any $\alpha \geq 0$ and $x, y \in V$.

[5] Formally, a free lunch in market $(\mathcal{Z}, \pi, \mathbb{P})$ consists of a sequence $\{Y_n\} \subset L^p(\mathbb{P})$ such that $Y_n \to Y \geq 0$ and a sequence $\{Z_n\} \subset \mathcal{Z}$ such that $\liminf_n \pi(Z_n) \leq 0$ and $Z_n \geq Y_n$ for every $n \in \mathbb{N}$.

[6] A market $(\mathcal{Z}, \pi, \mathbb{P})$ is viable if there exists an agent (represented by preference $\gtrsim$) and $Z^* \in \mathcal{Z}$ such

2.3 Foundation in markets with frictions or ambiguity

Our S–SDF framework embeds most existing asset pricing settings. An obvious one is the frictionless arbitrage-free framework of Hansen and Jagannathan [1991], in which $C = \{\mathbf{0}\}$, $\mathcal{Z} = \{\boldsymbol{\theta}'\boldsymbol{X} : \boldsymbol{\theta} \in \mathbb{R}^N\}$ and price vector $\boldsymbol{P}$ is matched exactly by the vector of asset valuations of a strictly positive SDF. Market settings with frictions corresponding to sublinear or convex transaction cost functions σ are naturally embedded as well. Here, Proposition 1 and Proposition 1 of Online Appendix B.4 provide the unique constraint set $C := C_\sigma$ of a strictly positive S–SDF associated to sublinear or convex specifications of transaction costs σ.[7]

It is important to note that S–SDFs can arise also in frictionless markets, when the unobserved marginal investor belief and the probability belief $\mathbb{P}$ used to describe asset payoffs are not equivalent, offering a convenient framework to formulate robust asset pricing predictions in a variety of settings where an SDF may not exist under belief $\mathbb{P}$.[8] To clarify this, denote by $\tilde{\mathbb{P}}$ the marginal investor belief and let frictionless market $(\tilde{Z}_{\{\mathbf{0}\}}, \tilde{\pi}_{\{\mathbf{0}\}}, \tilde{\mathbb{P}})$ admit no free lunches. From Proposition 1, there exists a SDF $\tilde{M}$ such that:

$$\tilde{\mathbb{E}}[\tilde{M}\boldsymbol{X}] = \boldsymbol{P} \ , \tag{7}$$

where $\tilde{\mathbb{E}}[\cdot]$ denotes expectations under probability $\tilde{\mathbb{P}}$. Importantly, an exact price representation of the form (7) can hold for a strictly positive SDF M with respect to a probability belief $\mathbb{P}$ if and only $\mathbb{P}$ and $\tilde{\mathbb{P}}$ are equivalent, in which case:

$$\boldsymbol{P} = \tilde{\mathbb{E}}[\tilde{M}\boldsymbol{X}] = \mathbb{E}[M\boldsymbol{X}] \ , \tag{8}$$

that $\pi(Z^*) \leq 0$ and $Z^* \gtrsim Z$ for all $Z \in \mathcal{Z}$ with $\pi(Z) \leq 0$. The preference relation $\gtrsim$ is assumed to be convex, continuous and strictly increasing; see again Harrison and Kreps [1979].

[7] For instance, economies with frictions given by short-sale constraints or bid-ask spreads arise for $C := -\mathbb{R}^N_+$ (the negative cone in $\mathbb{R}^N$) and a set $\mathcal{Z} = \{\boldsymbol{\theta}'\boldsymbol{X} : \boldsymbol{\theta} \in \mathbb{R}^N_+\}$ of traded payoffs with cost function $\sigma(\boldsymbol{\theta}) = 0$ ($\sigma(\boldsymbol{\theta}) = +\infty$) when $\boldsymbol{\theta} \geq \mathbf{0}$ (else); see again Luttmer [1996].

[8] For instance, in all continuous-time theoretical asset pricing settings, non equivalent beliefs $\tilde{\mathbb{P}}$ and $\mathbb{P}$ arise whenever the associated return volatility processes are not identical.

with the Radon-Nykodim derivative $N := \frac{d\tilde{\mathbb{P}}}{d\mathbb{P}}$ and $M := \tilde{M}N$. Conversely, no strictly positive SDF satisfying equation (8) under probability belief $\mathbb{P}$ can exist when some zero probability assessments under $\mathbb{P}$ and $\tilde{\mathbb{P}}$ differ. However, from Proposition 1 we also know more generally that for any closed convex set C containing $\{\mathbf{0}\}$ market $(\mathcal{Z}_C, \pi_C, \mathbb{P})$ admits no free lunches if and only if a strictly positive S–SDF M exists such that:

$$\mathbb{E}[M\boldsymbol{X}] - \tilde{\mathbb{E}}[\tilde{M}\boldsymbol{X}] \in C \ . \tag{9}$$

Here, set C can be interpreted as an implicit constraint on the degree of ambiguity between beliefs $\tilde{\mathbb{P}}$ and $\mathbb{P}$. Under such constraint, SDF $\tilde{M}$ and S–SDF M may imply distinct asset valuations under the corresponding beliefs, however with differences restricted to belong to set C. Such differences arise from the non-equivalence of $\mathbb{P}$ and $\tilde{\mathbb{P}}$, which renders the set of arbitrage opportunities perceived under each of these beliefs different.

Corollary 1 (S–SDFs and ambiguity). *Markets $(\tilde{\mathcal{Z}}_{\{\mathbf{0}\}}, \tilde{\pi}_{\{\mathbf{0}\}}, \tilde{\mathbb{P}})$ and $(\mathcal{Z}_C, \pi_C, \mathbb{P})$ admit no free lunches if an only if there exist strictly positive SDF $\tilde{M}$ and strictly positive S—SDF M such that constraints (9) hold, with pricing functional π_C explicitly given for any $Z \in \mathcal{Z}_C$ by:*

$$\pi_C(Z) = \inf_{\boldsymbol{\theta}\in\mathbb{R}^N} \left\{ \tilde{\mathbb{E}}[\tilde{M}\boldsymbol{X}'\boldsymbol{\theta}] + \sup_{\boldsymbol{\eta}\in C} \left\{ \boldsymbol{\eta}'\boldsymbol{\theta} \right\} \ : \ \mathbb{P}(Z = \boldsymbol{X}'\boldsymbol{\theta}) = 1 \right\} \ .$$

π_C in Corollary 1 is interpretable as the min-max replication cost of a payoff Z, after considering all ambiguous asset valuations under belief $\mathbb{P}$, which result from admissible price adjustments $\boldsymbol{\eta}$ relative to the marginal investor valuation $\boldsymbol{P} = \tilde{\mathbb{E}}[\tilde{M}\boldsymbol{X}]$. Since the no free lunches condition in market $(\mathcal{Z}_C, \pi_C, \mathbb{P})$ is weaker than the no free lunches condition in market $(\mathcal{Z}_{\{\mathbf{0}\}}, \pi_{\{\mathbf{0}\}}, \mathbb{P})$, because by definition $\mathcal{Z}_{\{\mathbf{0}\}} \supset \mathcal{Z}_C$ and $\pi_C(Z) \geq \pi_{\{\mathbf{0}\}}(Z)$ for all $Z \in \mathcal{Z}_C$, the pricing predictions of S–SDFs in ambiguous frictionless settings are also more robust than those of SDFs in unambiguous frictionless settings.

2.4 Minimum Dispersion S–SDFs

Having established their existence in arbitrage-free economies, we next provide an analysis and selection framework for S–SDFs, by minimizing established notions of stochastic dispersion in the literature. We measure dispersion via $\phi-$dispersions defined by integral functionals of the form $M \mapsto \mathbb{E}[\phi(M)]$, for a function $\phi : \mathbb{R} \to (-\infty, +\infty]$ such that ϕ_+, the restriction of ϕ to the nonnegative real line, is closed and strictly convex with $(0, +\infty) \subset \text{dom}(\phi_+)$.[9,10]

Definition 2 (Minimum dispersion S–SDFs). A minimum dispersion S–SDF M_0 is a S–SDF minimizing a particular $\phi-$dispersion, i.e., it solves the minimization problem:[11]

$$\Pi(C) := \inf_{M \in \mathcal{M}_+(C)} \mathbb{E}[\phi(M)] \ . \tag{10}$$

By definition, minimum dispersion S–SDF problems in Definition 2 give rise to an extended family of SDF dispersion bounds, relative to the bounds in Hansen and Jagannathan [1991], Luttmer [1996] and Almeida and Garcia [2016], among others, which can account for general convex pricing constraints. Since working directly with problem (10) is inconvenient, as it is an infinite-dimensional optimization problem, we provide in the sequel the dual characterization of minimum dispersion S–SDFs and S–SDF dispersion bounds via the solution of following penalized portfolio problem:

$$\Delta(C) := \min_{\boldsymbol{\theta} \in \mathbb{R}^N} \left\{ \mathbb{E}[\phi_+^*(-\boldsymbol{X}'\boldsymbol{\theta})] + \boldsymbol{P}'\boldsymbol{\theta} + \sigma_C(\boldsymbol{\theta}) \right\} = \min_{\boldsymbol{\theta} \in \mathbb{R}^N} \left\{ \mathbb{E}[\phi_+^*(-\boldsymbol{X}'\boldsymbol{\theta})] + \pi_C(\boldsymbol{X}'\boldsymbol{\theta}) \right\} \ , \tag{11}$$

with the penalization function σ_C in equation (3) and the convex conjugate of ϕ_+, defined for any $y \in \mathbb{R}$ by $\phi_+^*(y) := \sup_{x \in \mathbb{R}} \{yx - \phi_+(x)\}$. In this portfolio problem, the dual

[9] $\phi_+(x)$ is defined as $\phi_+(x) = \phi(x)$ when $x \geq 0$ and $\phi_+(x) = +\infty$ otherwise.

[10] Many well-known measures of SDF dispersion are $\phi-$dispersions, such as the variance, entropy-based dispersions and more generally dispersions in the Cressie and Read [1984] family. Online Appendix B.1 collects relevant explicit examples of $\phi-$dispersions in this family.

[11] Naturally, the $\phi-$dispersion in problem (10) is chosen so that $\phi(M) \in L_1$ for any $M \in \mathcal{M}_+(C) \subset L_q$, where $q = p/(p-1)$ and L_p ($p \in (1, \infty)$) is the underlying payoff space.

dispersion functional $Z \mapsto \mathbb{E}[\phi_+^*(Z)]$ measures the dispersion of portfolio payoffs Z, while $\sigma_C(\boldsymbol{\theta})$ penalizes a candidate assets' portfolio weight relative to the benchmark portfolio cost $\boldsymbol{P'\theta}$. Equivalently, in the second identity of equation (11), a candidate portfolio payoff is penalized by its total cost $\pi_C(\boldsymbol{X'\theta})$ in the economy $(\mathcal{Z}_C, \pi_C, \mathbb{P})$ of Section 2.2.

The strong duality between minimum dispersion S–SDFs and optimal portfolios solving penalized problem (11) is established next.

Proposition 2 (Duality). *For some convex closed set $\tilde{C} \subset \mathbb{R}^N$ containing $\mathbf{0}$ consider the corresponding economy $(\mathcal{Z}_{\tilde{C}}, \pi_{\tilde{C}}, \mathbb{P})$ from Section 2.2. If such economy admits no free lunches, then $\Pi(C) = -\Delta(C)$ for any closed convex set C with a relative interior containing $\tilde{C}$.*[12] *If additionally $\boldsymbol{\theta}_0$ is a solution of problem (11) such that $-\boldsymbol{X'\theta}_0 < \lim_{y\to\infty} \phi(y)/y$, $\mathbb{P}-$almost surely, then either minimum S–SDF dispersion problem (10) has no solution, or it has a unique solution given by:*[13]

$$M_0 = (\phi_+^*)'(-\boldsymbol{X'\theta}_0) \ . \tag{12}$$

Proposition 2 states that under appropriate conditions penalized portfolio problem (11) can be used to compute minimum dispersion bound (10) and minimum dispersion S–SDF M_0, as a simple transformation of the optimal dual portfolio payoff in equation (12).[14] Therefore, it provides a powerful device to compute minimum dispersion S–SDFs and S–SDF bounds in a wide variety of practically relevant asset pricing settings.

Remark 3. The duality result in Proposition 2 holds under a no-arbitrage condition of the form given in Proposition 1, for a set $\tilde{C}$ contained in the relative interior of constraint set C. This condition covers all practically relevant constraint sets C considered in this paper. For instance, is C is polyhedral as in, e.g., Luttmer [1996], then we can set $\tilde{C} = C$ without loss

[12] The relative interior of a set is the interior within the set's affine hull.

[13] With a slight abuse of notation, we denote by $(\phi_+^*)'(y)$ in equation (12) the first derivative of function ϕ_+^* in an interior point y of the domain of ϕ_+^*.

[14] Online Appendix B.1 collects closed-form expressions of ϕ_+, ϕ_+^* and the minimum dispersion S–SDF (12) for the family of Cressie and Read [1984] dispersions.

of generality. Analogously, if C is a closed norm ball then $\tilde{C}$ can be any closed convex set included in the interior of C. Furthermore, if C is polyhedral or σ_C real-valued, then the no free lunches condition in Proposition 2 is equivalent to the standard no-arbitrage condition in the literature.[15] Finally, the technical condition $-\boldsymbol{X}'\boldsymbol{\theta}_0 < \lim_{y\to\infty} \phi(y)/y$, $\mathbb{P}-$almost surely, is obviously satisfied by any $\phi-$dispersions function with superlinear growth, such as all Cressie-Read dispersions with power parameter $\gamma \geq 1$ in Online Appendix B.1.

Remark 4. When duality does not hold for a given closed convex set C ($\Pi(C) \neq \Delta(C)$), then for all closed convex sets $\tilde{C}$ included in the relative interior of C market $(\mathcal{Z}_{\tilde{C}}, \pi_{\tilde{C}}, \mathbb{P})$ is not arbitrage-free. In such situations, a solution to dual portfolio problem $\Delta(C)$ in equation (11) may still exist, but random variable M_0 in equation (12) may not be a minimum dispersion S–SDF. Emergence of such a duality failure can be assessed empirically, by verifying whether M_0 satisfies the pricing constraints in the primal minimum S–SDF dispersion problem (10). We borrow from this insight in our empirical analysis of Section 5, in order to distinguish asset pricing settings with solid S–SDF foundation from those with weak S–SDF foundation.

3 S–SDF Applications

Minimum dispersion S–SDFs and their dual portfolio problems are directly related to various important asset pricing settings corresponding to concrete equivalent specifications of constraint set C or penalty function σ_C in identity (3); see Table 1 for a list of closed-form (C, σ_C) pairs. These include, e.g., recent settings proposing SDFs that are regularized using techniques from the machine learning literature.

[15] See Theorem 6 in Clark [1993]. Formally, an arbitrage opportunity in a market $(\mathcal{Z}, \pi, \mathbb{P})$ is a nonzero traded payoff $Z \in \mathcal{Z}$ such that $X \geq 0$ and $\pi(Z) \leq 0$.

3.1 S–SDFs with bounded valuations

Minimum variance S–SDFs incorporating conic portfolio constraints, such as short sale constraints or bid-ask spreads, are studied in Luttmer [1996]. As detailed in Online Appendix B.2, such S–SDFs are obtained with Proposition 2 under a penalization implying zero (infinite) transaction costs for each portfolio satisfying (violating) the constraints. A specific property of these S–SDFs is that S–SDF asset valuations, while delimited by a convex cone, are not bounded. In the sequel, we address instead relevant asset pricing settings with bounded asset valuations.

A useful class of costs for the portfolio positions can be modeled using norms: $\sigma = \tau \|\cdot\|$, where $\|\cdot\|$ is some norm in $\mathbb{R}^N$ and $\tau > 0$ is a scaling parameter. With Proposition 1, S–SDFs in these markets give rise to bounded asset valuations based on the dual norm $\|\cdot\|_*$, since:[16]

$$C_\sigma = \{\boldsymbol{\eta} \in \mathbb{R}^N : \boldsymbol{\eta}'\boldsymbol{\theta} \leq \tau \|\boldsymbol{\theta}\| \text{ for all } \boldsymbol{\theta} \in \mathbb{R}^N\} = \{\boldsymbol{\eta} \in \mathbb{R}^N : \|\boldsymbol{\eta}\|_* \leq \tau\} . \tag{13}$$

For instance, a specification $\sigma = \tau \|\cdot\|_1$ of proportional transaction costs with an l_1-norm implies differences between S–SDF valuations and vector $\boldsymbol{P}$ of prices that are uniformly bounded in absolute value by τ.[17]

For asset pricing settings with no explicit role for transaction costs, S–SDFs implying norm bounded asset valuations can be founded in frictionless arbitrage-free economies with ambiguity. In such settings, a constraint set of the form (13) simply restricts the discrepancies between the marginal investors' arbitrage-free valuations of asset payoffs and the S–SDF valuations under belief $\mathbb{P}$.

Remark 5. Note that specification $\sigma_C = \tau \|\cdot\|_1$ with an l_1-norm reproduces the widely

[16] The dual norm $\|\cdot\|_*$ is defined by $\|\boldsymbol{\eta}\|_* := \max_{\boldsymbol{\theta} \in \mathbb{R}^N} \{\boldsymbol{\eta}'\boldsymbol{\theta} : \|\boldsymbol{\theta}\| \leq 1\}$.

[17] Given an l_p-norm such that $\|\boldsymbol{x}\|_p := \left(\sum_{i=1}^{N_D} |x_i|^p\right)^{1/p}$ when $p \in [1, +\infty)$ and $\|\boldsymbol{x}\|_p := \max_{i=1}^{N_D} |x_i|$ when $p = +\infty$, its dual norm is the l_q-norm with $1/p + 1/q = 1$. In particular, the $l_\infty-$norm is the dual norm of the l_1-norm and vice-versa, while the l_2-norm is self-dual.

used lasso penalty in the machine learning literature; see Tibshirani [1996]. This penalty is known to produce sparse optimal solutions, which in our setting implies sparse optimal portfolio weights in the solution of the dual portfolio problem of Proposition 2. Given the link (12) between optimal portfolios and minimum dispersion S–SDFs, this choice translates to S–SDFs that depend on a strict subset of asset returns. According to Proposition 2, these penalization techniques give rise to robust S–SDFs with pricing constraints defined by closed-form bounds of the form (13). Therefore, they also admit an explicit no-arbitrage foundation within a frictionless economy with ambiguity.

Remark 6. An important implication of our theory is that sparsity cannot be obtained together for the optimal portfolio weights of minimum dispersion S–SDFs and the vector $\boldsymbol{\eta}_0 = \mathbb{E}[M_0 \boldsymbol{X}] - \boldsymbol{P}$ of asset misvaluations under such SDF. This sparsity trade-off is general and arises also for penalizations implying a more flexible relation between shrinkage and sparsity than the lasso penalty. One such example is the Elastic Net penalty used in Kozak et al. [2020] to shrink the cross-section of returns. In our framework, a flexible relation between shrinkage and sparsity is obtained with norm-based penalizations combining the sparsity properties of the lasso and the shrinkage properties of l_2-penalties: $\sigma_C = \alpha \left\| \cdot \right\|_1 + \tau \left\| \cdot \right\|_2$. The associated closed-form constraint set C in Table 1 bounds with a threshold τ the Euclidean distance of vector $\boldsymbol{\eta}_0$ from the $l_\infty-$norm ball of radius α. Here, maximal differences in S–SDF asset valuations larger than α are penalized quadratically, which gives rise to a quadratic constraint above a threshold that implies no sparsity in vector $\boldsymbol{\eta}_0$.[18]

3.2 Minimum dispersion S–SDFs as pricing error minimizers

As formally shown in Proposition OA-1 of Online Appendix A, minimum dispersion S–SDFs are equivalently defined by means of a minimum pricing error problem with bounded

[18] The limit case $\alpha = 0$ corresponds to a standard ridge pricing metric with penalization $\sigma_C = \tau \left\| \cdot \right\|_2$, which gives rise to non sparse optimal dual portfolio weights and a non sparse vector $\boldsymbol{\eta}_0$. Many other pricing metrics and portfolio weight penalties can be implemented in our framework, covering all sublinear penalties corresponding to any convex and closed constraint set C. Online Appendix B.4 shows how convex, but not sublinear, penalizations are covered as well by our theory.

S–SDF dispersion, using a convex and closed pricing error metric $h := h_C$ and a threshold $\tau := \tau_C$ such that $C = \{\boldsymbol{\eta} \in \mathbb{R}^N : h(\boldsymbol{\eta}) \leq \tau\}$. Such equivalent minimum pricing error problem reads:

$$\tilde{\Pi}(C) := \inf_{M \in L^q_+} \{h(\mathbb{E}[M\boldsymbol{X} - \boldsymbol{P}]) \ : \ \mathbb{E}[\phi(M)] \leq \nu\} \ , \tag{14}$$

for a corresponding bound $\nu := \nu_C > 0$ on the S–SDF dispersion. This equivalence establishes a useful link between our framework and further directions in the literature, such as the good-deal bounds theory in Cochrane and Saa-Requejo [2000], the robust investor's belief identification approach in Chen et al. [2020] and the regularized SDFs proposed by Kozak et al. [2020].[19] Given a vector $\boldsymbol{R}^e$ of N_D excess returns, benchmark examples of the latter are parametric S–SDFs of the form $M(\boldsymbol{\theta}_0) := \theta_{0S} - \boldsymbol{\theta}'_{0D}(\boldsymbol{R}^e - \mathbb{E}[\boldsymbol{R}^e])$, with $\boldsymbol{\theta}_0 = (\theta_{0S}, \boldsymbol{\theta}'_{0D})'$, and such that:

$$\boldsymbol{\theta}_0 = \arg \min_{\boldsymbol{\theta} \in \mathbb{R}^{N_D+1}} \left\{ \|\mathbb{E}[M(\boldsymbol{\theta})\boldsymbol{R}^e]\|_2 \ : \ \mathbb{E}[M^2(\boldsymbol{\theta})] \leq \nu \text{ and } \mathbb{E}[M(\boldsymbol{\theta})] = 1 \right\} \ .$$

This problem is naturally embedded into our framework, using asset payoff vector $\boldsymbol{X} := (1, \boldsymbol{R}^{e\prime})'$, price vector $\boldsymbol{P} = (1, \boldsymbol{0}')'$ and constraint set $C = \{0\} \times \{\boldsymbol{\eta} \in \mathbb{R}^{N_D} : \|\boldsymbol{\eta}\|_2 \leq \tau\}$. Therefore, $M(\boldsymbol{\theta}_0)$ is a minimum variance S–SDF, with a norm-based pricing constraint of the form introduced in Section 3.1 and a dispersion function given explicitly by $\phi(x) = x^2$ for any $x \in \mathbb{R}$.

4 APT S–SDFs

The key constraint underlying the APT is a finite maximal Sharpe ratio in markets with no asymptotic arbitrage opportunities.[20] When returns satisfy a suitable factor structure, such a constraint gives rise to approximate linear valuation rules with asset pricing predictions that are reproducible by suitable S–SDFs with a norm bounded constraint set C. This

[19] See Online Appendix B.3 for a detailed derivation of the relation between our S–SDF methodology, good deal bounds theories and robust belief identification approaches.

[20] See, e.g., [Chamberlain and Rothschild, 1983, Corollary 1] for a formal statement.

section explains the properties of these S–SDFs.

4.1 APT setting

Consider a factor model for a $N_D \times 1$ vector of excess returns $\boldsymbol{R}^e_{N_D}$:

$$\boldsymbol{R}^e_{N_D} = \boldsymbol{\Lambda}_{N_D} \boldsymbol{F}^e + \boldsymbol{\zeta}_{N_D} \ , \tag{15}$$

where $\boldsymbol{F}^e$ is a $N_S \times 1$ vector of observed traded excess factor returns and $\boldsymbol{\Lambda}_{N_D}$ a $N_D \times N_S$ matrix of factor loadings. The $N_D \times 1$ vector of residuals $\boldsymbol{\zeta}_{N_D}$ is orthogonal to traded factor risk, but potentially cross-sectionally correlated, with variance covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\zeta}_{N_D}} = \boldsymbol{B}_{N_D} \boldsymbol{B}'_{N_D} + \boldsymbol{Q}_{N_D}$, where $\{\boldsymbol{B}_{N_D}\}$ is a sequence of $N_D \times N_K$ matrices and $\{\boldsymbol{Q}_{N_D}\}$ a sequence of symmetric positive definite $N_D \times N_D$ matrices with uniformly bounded eigenvalues.[21] In this setting, it follows from [Chamberlain and Rothschild, 1983, Corollary 2] that the absence of asymptotic arbitrage opportunities, under a sequence of factor models (15) with growing dimension N_D, yields existence of a constant $\tau \geq 0$ and a $N_K \times 1$ vector $\boldsymbol{\lambda}$ such that:[22]

$$\|\boldsymbol{\eta}_{N_D}\|_{2,\boldsymbol{\Sigma}^{-1/2}_{\boldsymbol{\zeta}_{N_D}}} := \sqrt{\boldsymbol{\eta}'_{N_D} \boldsymbol{\Sigma}^{-1}_{\boldsymbol{\zeta}_{N_D}} \boldsymbol{\eta}_{N_D}} \leq \tau \ , \tag{16}$$

where

$$\boldsymbol{\eta}_{N_D} := \mathbb{E}[\boldsymbol{\zeta}_{N_D}] - \boldsymbol{B}_{N_D} \boldsymbol{\lambda} \ . \tag{17}$$

In equation (17), $\boldsymbol{\eta}_{N_D}$ is an expected excess return component not explained by priced exposures to traded and unobservable factor risk in model (15).[23] Therefore, as stressed by Uppal et al. [2019], these predictions hold also in the case of a misspecified factor model

[21] Sufficient conditions for existence and uniqueness of such a factor representation are provided in [Chamberlain and Rothschild, 1983, Theorem 4].

[22] See also Uppal et al. [2019]. Constant τ in inequality (16) is explicitly given as $\delta^2 \lambda_{k+1}/\lambda_0$, where $\lambda_{k+1} < \infty$ ($\lambda_0 > 0$) is a uniform upper (lower) bound on the $k+1$ largest (the smallest) eigenvalue of $\boldsymbol{\Sigma}_{\boldsymbol{\zeta}_N}$ and δ^2 is the squared maximal Sharpe ratio in the underlying arbitrage-free economy.

[23] Since bound (16) is independent of N_D, we consider in the sequel an excess return vector $\boldsymbol{R}^e$ of fixed dimension and drop the N_D subscripts.

(15). On the other hand, $\mathbb{E}[\boldsymbol{F}^e]$ is the vector of risk premia of traded factors, which by construction have a price of zero.

4.2 Existence of APT–SDFs

The next corollary to Proposition 1 founds a family of S–SDFs compatible with APT pricing constraints (16), based on a founding economy that admits no arbitrage opportunities in the standard sense.[24]

Corollary 2 (APT S–SDF). *Given the above APT setting, consider constraint set* $C = \times\{\mathbf{0}_{N_S+1}\} \times \{\boldsymbol{\eta} \in \mathbb{R}^{N_D} : \|\boldsymbol{\eta}\|_{2,\boldsymbol{\Sigma}_\zeta^{-1/2}} \leq \tau\}$ *and the associated penalization* σ_C*, which for any* $\boldsymbol{\theta} = (\theta, \boldsymbol{\theta}_S', \boldsymbol{\theta}_D')'$ *is given in closed-form by:*

$$\sigma_C(\boldsymbol{\theta}) = \tau \, \|\boldsymbol{\theta}_D\|_{2,\boldsymbol{\Sigma}_\zeta^{1/2}} \ . \tag{18}$$

Then, market $(\mathcal{Z}_C, \pi_C, \mathbb{P})$ *in equations (4)–(5) is arbitrage-free if and only if there exists a strictly positive S–SDF* M_{APT}*, called APT S–SDF, with normalized expectation* $\mathbb{E}[M_{APT}] = 1$ *and such that:*

$$\mathbb{E}[M_{APT}\boldsymbol{F}^e] = \mathbf{0} \ \textit{and} \ \|\mathbb{E}[M_{APT}\boldsymbol{R}^e]\|_{2,\boldsymbol{\Sigma}_\zeta^{-1/2}} \leq \tau \ . \tag{19}$$

Corollary 2 provides a model-free foundation of APT asset pricing relations in a frictionless arbitrage-free economy with ambiguity, where the valuation of excess returns orthogonal to traded factor risks is constrained with the standardized APT metric $h = \|\cdot\|_{2,\boldsymbol{\Sigma}_\zeta^{-1/2}}$. Indeed, under the APT S–SDF in Corollary 2 traded factor risks obey the standard identity:

$$\mathbb{E}[\boldsymbol{F}^e] = -Cov(M_{APT}, \boldsymbol{F}^e) \ , \tag{20}$$

i.e., tradable factor risk premia are completely explained by the exposure of traded factor

[24] See again Footnote 15 for the formal definition of an arbitrage opportunity.

excess returns to APT S–SDF risk. In contrast, excess returns not fully spanned by traded factor risk satisfy the identity:

$$\mathbb{E}[\boldsymbol{R}^e] = -Cov(M_{APT}, \boldsymbol{R}^e) + \mathbb{E}[M_{APT}\boldsymbol{R}^e] \ . \tag{21}$$

These excess returns are the sum of a risk premium component, generated by their covariance with APT S–SDF risk, and a second component $\boldsymbol{\eta} = \mathbb{E}[M_{APT}\boldsymbol{R}^e]$ that satisfies the APT bound (16). In this sense, decomposition (21) is a model-free reproduction of the APT asset pricing predictions using APT S–SDFs. Such decomposition is naturally parametrized by threshold parameter τ. This parameter bounds the contribution of component $\boldsymbol{\eta}$ to expected excess returns in the arbitrage-free founding economy with ambiguity from Corollary 2.

Remark 7. Combining APT bound (19) and identity (21) with APT factor model (15) it follows:

$$\mathbb{E}[\boldsymbol{R}^e - \boldsymbol{\Lambda}\boldsymbol{F}^e]'\boldsymbol{\Sigma}_{\boldsymbol{\zeta}}^{-1}\mathbb{E}[\boldsymbol{R}^e - \boldsymbol{\Lambda}\boldsymbol{F}^e] \leq \tau \ . \tag{22}$$

Under the assumption of a Gaussian residuals distribution with common covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\zeta}}$, this bound is equivalent to a relative entropy bound, which constrains the ambiguity between residual distributions with non-zero mean and a zero-mean residual distribution in factor model (15); see, e.g., Hansen and Sargent [2007].

Remark 8. Corollary 2 involves a closed-form penalization $\sigma_C(\boldsymbol{\theta}) = \tau \left\|\boldsymbol{\theta}_D\right\|_{2,\boldsymbol{\Sigma}_{\boldsymbol{\zeta}}^{1/2}}$, which founds a strictly positive S–SDF satisfying the APT pricing constraints (19) under pricing metric $\|\cdot\|_{2,\boldsymbol{\Sigma}_{\boldsymbol{\zeta}}^{-1/2}}$. Clearly, these pricing constraints are satisfied also by any S–SDF M such that:

$$\mathbb{E}[M\boldsymbol{F}^e] = \boldsymbol{0} \ and \ \left\|\mathbb{E}[M\boldsymbol{R}^e]\right\|_{\boldsymbol{\Sigma}_{\boldsymbol{\zeta}}^{-1/2}} \leq \tau \ , \tag{23}$$

under a pricing metric such that $\|\cdot\| \geq \|\cdot\|_2$. Such a APT–SDF is founded in an arbitrage-

free economy with cost function $\sigma_C(\boldsymbol{\theta}) = \tau \left\|\boldsymbol{\theta}_D\right\|_{*,\boldsymbol{\Sigma}_{\boldsymbol{\zeta}}^{1/2}}$. However, since $\|\cdot\| \geq \|\cdot\|_2$ the no arbitrage condition in such economy is more restrictive that the one adopted in Corollary 2. Therefore, while model-free APT S–SDF predictions are reproducible using a whole family of S-SDFs with distinct valuation geometries, the foundation of the APT S–SDFs in Corollary 2 relies on the least restrictive no-arbitrage condition.

4.3 Minimum dispersion APT–SDFs

APT S–SDFs are naturally suited for an analysis and selection framework based on minimum dispersion S–SDFs, in which we can coherently isolate and study the effects of distinct assumptions underlying APT asset pricing predictions, such as assumptions regarding the set of traded factors or the tightness of APT pricing error bound (16). This framework is made explicit by the next corollary of Proposition 2, which provides the basis for our empirical analysis in Section 5.

Corollary 3 (**Minimum Dispersion APT S–SDF**). *Given the above APT setting, consider the minimum dispersion APT S–SDF, $M_{0,APT}$, solving minimum dispersion problem (10) for constraint set $C = \{\mathbf{0}_{N_S+1}\} \times \{\boldsymbol{\eta} \in \mathbb{R}^{N_D} : \|\boldsymbol{\eta}\|_{\boldsymbol{\Sigma}_{\boldsymbol{\zeta}}^{-1/2}} \leq \tau\}$ under a norm $\|\cdot\| \geq \|\cdot\|_2$. Then, dual portfolio problem (11) reads in closed-form:*

$$\Delta(C) = \min_{\boldsymbol{\theta} \in \mathbb{R}^N} \left\{ \mathbb{E}[\phi_+^*(-\boldsymbol{X}'\boldsymbol{\theta})] + \theta + \tau \left\|\boldsymbol{\theta}_D\right\|_{*,\boldsymbol{\Sigma}_{\boldsymbol{\zeta}}^{1/2}} \right\} . \tag{24}$$

Let further $\tilde{C} \subset \{\mathbf{0}_{N_S+1}\} \times \{\boldsymbol{\eta} \in \mathbb{R}^{N_D} : \|\boldsymbol{\eta}\|_{2,\Sigma_{\boldsymbol{\zeta}}^{-1/2}} < \tau\}$ be closed, convex and contain $\mathbf{0}$, and the corresponding economy $(\mathcal{Z}_{\tilde{C}}, \pi_{\tilde{C}}, \mathbb{P})$ from Section 2.2 be arbitrage-free. Then, $\Pi(C) = -\Delta(C)$. If additionally $\boldsymbol{\theta}_0$ is a solution of problem (24) such that $-\boldsymbol{X}'\boldsymbol{\theta}_0 < \lim_{y\to\infty} \phi(y)/y$, $\mathbb{P}-$almost surely, then either minimum S–SDF dispersion problem (10) has no solution, or it has a unique solution given by:

$$M_0 = (\phi_+^*)'(-\boldsymbol{X}'\boldsymbol{\theta}_0) . \tag{25}$$

In summary, under appropriate conditions, minimum dispersion APT S–SDFs exist and can be computed as closed-form transformations of optimal portfolio payoffs from closed-form penalized portfolio problem (24).[25] In such problem, different choices of norm $\|\cdot\|$ imply potentially very different S–SDF properties, in terms of, e.g., S–SDF sparsity, which are however all compatible with the model-free APT predictions (19).

Remark 9. By construction, minimum dispersion APT S–SDF problems give rise to an extended family of closed-form SDF dispersion bounds, relative to the bounds in Hansen and Jagannathan [1991], Luttmer [1996] and Almeida and Garcia [2016], among others, which are parametrized by the APT pricing constraints. Given the choice of a norm $\|\cdot\|$, this reduces to a simple parametrization with threshold τ in APT pricing constraint (23). Therefore, let M_0^S be the unique S–SDF solving problem $\Pi(\infty) := \sup_{\tau>0} \Pi(\tau)$, i.e., M_0^S is only required to have expectation 1 and to price the traded factor excess returns exactly, and define:

$$\tau^{max} := \left\| \mathbb{E}[M_0^S \boldsymbol{R}^e] \right\| \ . \tag{26}$$

τ^{max} defines a pricing constraint on excess returns satisfied by a normalized minimum dispersion APT SDF only required to price exactly the traded factors. The resulting generalized dispersion bound $\Pi(\tau)$ is strictly convex and strictly decreasing for any $\tau \leq \tau^{max}$, while it is constant at $\Pi(\tau) = E[\phi(M_0^S)]$ if $\tau \geq \tau^{max}$.[26]

5 Empirics of APT S–SDFs

Starting from our theoretical analysis and selection framework for minimum dispersion APT S–SDFs, we empirically investigate the in-sample and out-of-sample pricing properties of a convenient family of APT S–SDFs, which differ with respect to their cross-sectional

[25] See again Remark 3 for a more detailed discussion of the technical condition $-\boldsymbol{X}'\boldsymbol{\theta}_0 < \lim_{y\to\infty} \phi(y)/y$, $\mathbb{P}-$almost surely.

[26] This follows by [Luenberger, 1997, Prop. 1 and 2, p. 216-217], the strict convexity of $\phi-$dispersions and the convexity of norms.

pricing accuracy and their degree of comovement with systematic asset return risks. We show that a tradeoff between these two key S–SDF properties naturally emerges, thus providing a criterion to compare APT S–SDFs corresponding to different pricing metrics, dispersion measures and choices of traded factors.

5.1 Data

The results in this section are based on standard datasets of characteristics sorted portfolio returns of varying cross-sectional dimensions, which build on monthly returns of the risk-free asset, the S&P 500 index as a proxy for the market index, the SMB and the HML factors and on the following selections of the double sorted portfolios from Kenneth French's data library:

1. **Low dimensional dataset:** 25 portfolio returns sorted on size and book to market, 10 portfolio returns sorted on momentum, and 25 portfolio returns sorted on size and long term reversal.

2. **Intermediate dimensional dataset:** 100 portfolio returns sorted on size and book to market, 25 portfolio returns sorted on momentum, 25 portfolio returns sorted on size and long term reversal, 25 portfolio returns sorted on size and short term reversal, and 49 portfolio returns sorted on industry.

3. **High dimensional dataset:** The portfolios in the intermediate dimensional dataset augmented with the characteristics-based factors from the WRDS financial ratios (WFR) dataset, which consists of 69 ratios for 10 industries based on Fama-French industry classification. The construction of the factor returns follows Kozak et al. [2020].

We also consider the returns of various Fama-French factors for constructing simple benchmark linear SDFs and for incorporating well-known systematic risk exposures as traded factors in our APT S–SDFs. Our sample starts in January 1931 (January 1970 for the

characteristic-based factors in the high dimensional dataset) and ends in June 2018. Portfolios with missing time series observations are removed. After such removal, the low dimensional dataset consists of 57 assets and 1054 time series observations, the intermediate dimensional dataset consists of 188 assets and 1054 time series observations, while the high dimensional dataset has 260 asset returns with 561 time series observations.

5.2 Empirical setting

In order to study APT S–SDFs with varying asset pricing and sparsity features, we consider for $\lambda \in [0,1]$ the following APT norms:

$$h_{1,\boldsymbol{\Sigma}_{\boldsymbol{\zeta}}^{-1/2}}(\boldsymbol{\eta}_D) := (1-\lambda)\left\|\boldsymbol{\eta}_D\right\|_{1,\boldsymbol{\Sigma}_{\boldsymbol{\zeta}}^{-1/2}} + \lambda\left\|\boldsymbol{\eta}_D\right\|_{2,\boldsymbol{\Sigma}_{\boldsymbol{\zeta}}^{-1/2}} \ , \tag{27}$$

and

$$h_{\infty,\boldsymbol{\Sigma}_{\boldsymbol{\zeta}}^{-1/2}}(\boldsymbol{\eta}_D) := (1-\lambda)\sqrt{N_D}\left\|\boldsymbol{\eta}_D\right\|_{\infty,\boldsymbol{\Sigma}_{\boldsymbol{\zeta}}^{-1/2}} + \lambda\left\|\boldsymbol{\eta}_D\right\|_{2,\boldsymbol{\Sigma}_{\boldsymbol{\zeta}}^{-1/2}} \ . \tag{28}$$

Since $\|\cdot\|_2 \leq \|\cdot\|_1$ and $\|\cdot\|_2 \leq \sqrt{N_D}\,\|\cdot\|_\infty$, pricing constraints (23) based on these norms imply the benchmark APT pricing constraints (19). Moreover, while norm (27) induces varying degrees of sparsity on the resulting S–SDF valuations of excess returns, norm (28) implies sparsity on the portfolio weights supporting minimum dispersion APT S–SDFs.[27]

Estimators of S–SDF dispersion bounds and dual portfolio weights in stationary dynamic economies can be naturally developed. Let $\{\boldsymbol{X}_{t+1}\}_{t\in\mathbb{N}}$ be a time series of payoffs of the N basis assets, where $\boldsymbol{X}_{t+1} = (1, \boldsymbol{F}_{t+1}^{e\prime}, \boldsymbol{R}_{t+1}^{e\prime})'$, defined on probability space $(\Omega, \mathcal{F}, \mathbb{P})$. This induces the following convex pricing constraints under the stationary probability $\mathbb{P}$:

$$\mathbb{E}[M_{t+1}(1, \boldsymbol{F}_{t+1}^{e\prime})'] = (1, \mathbf{0}')' \text{ and } \mathbb{E}[M_{t+1}\boldsymbol{R}_{t+1}^{e}] \in C \ . \tag{29}$$

[27] These features follow from the presence of the l_1-norm in pricing norm (27) and in the dual penalization $\sigma_{C_\infty} = \tau h_{\infty*,\boldsymbol{\Sigma}_{\boldsymbol{\zeta}}^{1/2}}$ from Corollary 3, respectively. See again Table 1 and Lemma OA-1 in Online Appendix A, which reports the closed-form expressions for penalization σ_{C_i} $(i = 1, \infty)$ from Corollary 3.

By Corollary 3, the minimum S–SDF dispersion bound associated with convex pricing constraints of the form (29) equals $-\Delta(C)$. Given a sample of size $T > 0$, the estimator for population value $\Delta(C)$ is thus:

$$\Delta_T(C) := \min \{Q_T(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \mathbb{R}^N\} , \tag{30}$$

with the empirical objective function

$$Q_T(\boldsymbol{\theta}) := \frac{1}{T}\sum_{t=1}^{T}[\phi_+^*(-\boldsymbol{X}_{t+1}'\boldsymbol{\theta})] + \theta + \sigma_C(\boldsymbol{\theta}) .$$

Accordingly, $\boldsymbol{\theta}_T \in \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^N}\{Q_T(\boldsymbol{\theta})\}$ is the corresponding estimator of dual optimal portfolio weight vector $\boldsymbol{\theta}_0 := \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^N}\{\mathbb{E}[\phi_+^*(-\boldsymbol{X}_{t+1}'\boldsymbol{\theta})] + \theta + \sigma_C(\boldsymbol{\theta})\}$. We apply these estimators to compute minimum dispersion APT S–SDFs from Corollary 3 for pricing norms (27)–(28) and a set of measures of $\phi-$dispersion from the Cressie and Read [1984] family; see again Online Appendix B.1.[28]

5.3 Existence of empirical minimum dispersion S–SDFs

In Corollary 3, a dual APT S–SDF optimal portfolio representation (12) holds under a no arbitrage assumption for the underlying asset market $(\mathcal{Z}_{\tilde{C}}, \pi_{\tilde{C}}, \mathbb{P})$. Consequently, a duality failure ($\Pi(C) \neq -\Delta(C)$) is interpretable as a weak arbitrage-free foundation of a minimum dispersion APT S–SDF. Empirically, we find that a duality failure can arise for constraint sets C associated with low thresholds τ. Figure 1 documents this for the high dimensional dataset. Here, we estimate various minimum variance APT S–SDFs via their empirical dual portfolio problems. We then compute the empirical quantity $\left\|\frac{1}{T}\sum_{t=1}^{T}(M_{0t+1}\boldsymbol{X}_{t+1})\right\|_{2,\boldsymbol{\Sigma}^{-1/2}}$ and verify whether it is smaller than the imposed threshold τ.

[28] While a formal study of the asymptotic properties of estimator $\boldsymbol{\theta}_T$ is beyond the scope of this paper, note that $\boldsymbol{\theta}_T$ is an $M-$estimator with a strictly convex but possibly nonsmooth objective function, when penalty σ_C is not differentiable. Asymptotic properties of such estimators are established under a fix dimension N, e.g., with the techniques in Andrews [1994]. Negahban et al. [2012] offer a unified framework for establishing consistency and convergence rates of high-dimensional M-estimators with norm based penalizations.

We find that for thresholds smaller than discontinuity point $\hat{\tau}^{min}$ in Figure 1 an empirical duality failure emerges, i.e., no strictly positive minimum variance empirical APT SDF exists. By recognizing the possibility of non-existence of a strictly positive empirical APT S–SDF, we can discipline economically the choice of penalization parameter τ.[29] In our empirical analysis, penalization parameters are constrained to be not smaller than lower bound $\hat{\tau}^{min}$. Moreover, as threshold τ increases, the pricing constraint on minimum dispersion APT S–SDFs is progressively relaxed, to the point that a minimum dispersion APT S–SDF only required to exactly price the traded factors satisfies the constraints for a sufficiently large threshold $\hat{\tau}^{max}$; see again Equation (26). In our empirical analysis, we ensure a well-defined empirical tradeoff between APT S–SDF spanning features for systematic risks and their pricing accuracy by constraining admissible penalization parameters in the interval $[\hat{\tau}^{min}, \hat{\tau}^{max}]$.

5.4 In–sample analysis

In the in-sample analysis, we characterize in detail the properties of the attainable tradeoffs between pricing accuracy and spanning properties for systematic return risks generated by APT S–SDFs under varying assumptions on the latter.

5.4.1 Tradeoff between cross-sectional and time series S–SDF explanatory power

We empirically explore the tradeoffs between pricing accuracy and spanning properties for systematic return risks of minimum variance APT S–SDFs.[30] In this section, we take

[29] In principle, an empirical duality failure may arise because of the ill-posedness of the empirical minimum S–SDF dispersion problem, also when when no duality failure may appear at the population level. We thank an anonymous referee for pointing this out.

[30] We focus on minimum variance APT S–SDFs, because as explained below and documented empirically in Online Appendix C.1 they always outperform by construction other minimum dispersion APT S–SDFs in terms of such tradeoffs. This makes variance the natural measure of dispersion to optimize these tradeoffs.

the market excess return as the single traded factor. Therefore, these APT S–SDFs are interpretable as minimum variance corrections of an empirical SDF under the CAPM, which allow for bounded misvaluations of excess returns orthogonal to market risk. While in Section 5.4.3 we also explore the implications of the choice of traded factors for the resulting S–SDF tradeoffs, one may argue that tradability of market returns is the most robust economically motivated assumption one may rely on when assuming tradability of some test assets. Moreover, since all theoretical asset pricing predictions in Section 4 hold independent of whether further traded factors beyond the assumed ones exist, the resulting APT S–SDF pricing predictions are similarly robust to the presence of further traded factors. Figure 2 quantifies the tradeoffs in the low dimensional dataset, based on different norms (27) and (28) such that $\lambda \in \{0, 0.5, 1\}$. Note that when $\lambda = 1$ norm $h_1 = \|\cdot\|_1$ ($h_\infty = \sqrt{N_D}\,\|\cdot\|_\infty$) implies extremely sparse valuation errors (S–SDFs dual portfolios), while for $\lambda = 0$ no sparsity in valuations errors and S–SDFs arises, because $h_1 = h_\infty = \|\cdot\|_2$. Parameter value $\lambda = 0.5$ implies instead a less extreme degree of sparsity in misvaluations (norm h_1) or S–SDFs (norm h_∞).[31]

Panel (A) of Figure 2 summarizes the tradeoff between minimum S–SDF volatility and pricing threshold τ, in an extended version of the minimum S–SDF volatility bounds in Hansen and Jagannathan [1991] and Luttmer [1996]. Since different norms give rise to a different maximal threshold $\hat{\tau}^{max}$ implied by the S–SDF only required to price exactly market excess returns, we collect minimum S–SDF volatilities parametrized by relative thresholds $\kappa := \tau/\hat{\tau}^{max} \in [\hat{\tau}^{min}/\hat{\tau}^{max}, 1]$.

In the right plot of Panel (B) of Figure 2, the minimum S–SDF volatilities induced by pricing constraints under the APT l_2-norm ($l_\infty-$norm) correspond to the highest (lowest) S–SDF cross-sectional explanatory power, when the explanatory power is measured by the

[31] Section 5.5 documents these sparsity properties in more detail.

standard cross-sectional GLS R^2 metric:[32]

$$R^2_{GLS} := 1 - \frac{\left\|\mathbb{E}[M_0(\boldsymbol{X} - \bar{\boldsymbol{X}}\mathbf{1})]\right\|^2_{2,\boldsymbol{\Sigma}^{-1/2}}}{\left\|\mathbb{E}[\boldsymbol{X} - \bar{\boldsymbol{X}}\mathbf{1}]\right\|^2_{2,\boldsymbol{\Sigma}^{-1/2}}} , \tag{31}$$

where $\mathbf{1}$ is a $N \times 1$ vector of ones, $\bar{\boldsymbol{X}} := (\mathbf{1}'\boldsymbol{X})/N$ the average of the components of vector $\boldsymbol{X}$ and $\boldsymbol{\Sigma}$ the $N \times N$ covariance matrix of excess return vector $\boldsymbol{X}$. This correspondence offers a second useful interpretation of penalization parameter τ in Corollary 3, which can be understood as a parametrization of the relative loss $1 - R^2_{GLS}$ in cross-sectional explanatory power that an APT S–SDF suffers with respect to the minimum dispersion SDF exactly pricing all assets. Because of the scale independence of R^2_{GLS}, this interpretation simplifies the comparison of minimum dispersion APT S–SDFs induced by different pricing constraints.[33]

Given a set of traded systematic factor excess returns, the minimum S–SDF volatilities induced by APT pricing constraints are also directly related to the S–SDF ability to comove with systematic excess return risks. Indeed, since traded excess factor returns are priced exactly, lower S–SDF pricing errors are linked to both larger S–SDF volatilities and lower absolute correlations with traded systematic risks. This constraint impacts the time series S–SDF comovement with any excess return X, as measured by a standard time series OLS R^2:

$$R^2_{OLS} := \frac{(Cov(M_0, X))^2}{Var(M_0)Var(X)} = \frac{1}{Var(M_0)} \cdot \frac{(\mathbb{E}[X] - \mathbb{E}[M_0 X])^2}{Var(X)} . \tag{32}$$

Given a cross-section of asset (squared) Sharpe ratios, identity (32) constraints the S–SDF average time series explanatory power in the left plot of Panel (B) from Figure 2, as a function of the varying relative threshold κ. A higher cross-sectional explanatory power in the bottom right plot corresponds to a lower time series explanatory power in the bottom

[32] By definition, R^2_{GLS} equals one minus the ratio of the cross-sectional variance of pricing errors and the cross-sectional variance of expected returns, after a standardization by the return volatility matrix $\boldsymbol{\Sigma}^{1/2}$.

[33] As shown in Lewellen et al. [2010], R^2_{GLS} is a more appropriate measure of cross-sectional fit when pricing cross-sections of sorted portfolio returns.

left plot. However, what matters for an economic S–SDF comparison is the generated relative tradeoff between time series and cross-sectional explanatory powers. Panel (B) of Figure 2 shows that minimum variance S–SDFs under the APT l_2-norm always imply the most favourable tradeoff in-sample. For instance, while under this norm a cross-sectional GLS R^2 of 50% can be obtained together with an average time series R^2 of about 28%, under an APT l_1- and $l_\infty-$norm the resulting average time series R^2 is only about 20% and 14%, respectively. This evidence naturally follows from the fact that the l_2-pricing constraint is the only one producing for any given target value of cross-sectional GLS R^2 metric (31) the corresponding minimum variance S–SDF. Given the moment structure of excess returns, minimizing the S–SDF variance is equivalent to maximizing the time-series R^2 metric (32) with respect to the traded systematic risk factors. Therefore, the minimum variance S–SDF with l_2-pricing constraint produces a Pareto improved tradeoff under the standard GLS R^2 metric (32) of cross-sectional explanatory power.[34]

5.4.2 Metrics of pricing accuracy and robustness of S–SDF tradeoff

Two further important aspects for understanding the S–SDF tradeoff between pricing accuracy and time series explanatory power cover: (i) the role of the metric of pricing accuracy and (ii) the robustness of the tradeoff in presence of deviations from the in-sample assumptions.[35] Regarding the former aspect, recall that APT S–SDFs satisfy APT pricing constraints (23), under a norm $\|\cdot\|$ that may be different from the l_2-norm. Therefore, the natural equivalent of GLS R^2 metric (31) for measuring the S–SDF cross-sectional

[34] Figure OA-3 in the Online Appendix shows that absolute S–SDF betas monotonically decrease with threshold τ, as intuitively expected. Lewellen et al. [2010] emphasize the challenges arising in weakly identified empirical asset pricing settings relying on test assets with a hidden factor structure. In this context, incorporating well-chosen traded systematic risks in our framework can be interpreted as an informal S–SDF regularization against weak factor features.

[35] Intuitively, the second aspect relates to the fact that empirical S–SDFs with better resistance to varying in-sample data features are likely to perform better out-of-sample.

explanatory power is:

$$R^2_{GLS,\|\cdot\|} := 1 - \frac{\left\|\mathbb{E}[M_0(\boldsymbol{X} - \bar{\boldsymbol{X}}\mathbf{1})]\right\|^2_{\boldsymbol{\Sigma}^{-1/2}}}{\left\|\mathbb{E}[\boldsymbol{X} - \bar{\boldsymbol{X}}\mathbf{1}]\right\|^2_{\boldsymbol{\Sigma}^{-1/2}}} \,. \tag{33}$$

A natural question is how the tradeoff in Section 5.4.1 between cross-sectional and time series explanatory power is impacted by the choice of the metric of pricing accuracy. For each minimum variance APT S–SDF in Panel (B) of Figure 2, Figure 3 plots this tradeoff using cross-sectional GLS R^2 metric (33) with norms $\|\cdot\|_1$, $\|\cdot\|_2$, and $\sqrt{N_D}\,\|\cdot\|_\infty$. The three left plots show that minimum variance APT S–SDFs produce a Pareto dominating tradeoff when metric (33) is defined by the same norm as the one in APT bound (23). At the same time, it appears that the tradeoff induced by APT S–SDFs with l_2-pricing constraint is less sensitive to the choice of the GLS R^2 metric than the tradeoffs induced by l_1- and (scaled) $l_\infty-$pricing constraints.

Concerning the second aspect above, the three right plots of Figure 3 illustrate the effects of data perturbations, creating a discrepancy of about 10 years of monthly observations between S–SDF estimation and evaluation samples, for the S–SDF tradeoff between time series and cross-sectional explanatory powers. We find that APT S–SDFs satisfying an l_2-pricing constraint tend to produce a more consistent tradeoff on perturbed data, essentially outperforming the other S–SDFs with respect to all GLS R^2 metrics.

5.4.3 Traded factors and systematic S–SDF risks

Intuitively, traded factors in our framework correspond to tradeable systematic excess returns with a clearly understood risk compensation. As mentioned, the exact pricing condition on traded factor excess returns forces via equation (32) a maximal squared correlation with minimum variance S–SDFs, i.e., these S–SDFs maximize the comovement with the systematic risks spanned by the traded factors. Hence, the choice of the traded factor returns crucially influences the S–SDF tradeoff between cross-sectional and time

series explanatory power.

Figure 4 explores these tradeoffs for various minimum variance S–SDFs with a l_2-APT pricing constraint (16), under different choices of the traded factors: (i) no traded factor, (ii) one single traded factor given by the market excess return or by the first principal component of excess returns, and (iii) three traded factors given by the three Fama-French factor excess returns. We find that all S–SDFs imply a virtually identical cross-sectional GLS R^2 curve in-sample.[36] However, these S–SDFs in part differ in the way how they span systematic excess return risks. The S–SDFs including the market excess return or the first principal component of excess returns as a single traded factor produce similar and highest average time series R^2 curves. In contrast, S–SDFs including no traded factor uniformly produce a very low time series explanatory power that is essentially independent of the chosen degree of penalization. Therefore, including tradeable risk exposures that reflect well-understood systematic risks allows to balance the S–SDF trade-off between time series and cross-sectional explanatory power of APT S–SDFs under metrics (31) and (32). Extending the set of traded factors to include the three Fama-French factor returns produces a minimum variance correction of an empirical SDF under the three-factor Fama-French model. Interestingly, we find that this extension gives rise to a tradeoff that is Pareto dominated by the one of the simpler minimum variance correction of a CAPM SDF.

Overall, we conclude that minimum variance APT S–SDFs constraining valuations of asset excess returns with APT bound (19) and exactly pricing market excess returns or the first principal component of sorted excess returns consistently offer the best in-sample tradeoff between pricing accuracy and time series explanatory power.

[36] With the only distinction that in the intermediate dimensional dataset the empirical minimum variance S–SDF including the Fama-French factors as traded factors does not exist for a sufficiently low threshold τ.

5.5 The role of sparsity in APT S–SDFs

As emphasized in Section 3.1, sparsity cannot be obtained at the same time for the misvaluations of excess returns and the dual optimal portfolios of a minimum dispersion S–SDFs. Therefore, in an APT setting one can either impose sparsity on scaled vector $\boldsymbol{\Sigma}_{\boldsymbol{\zeta}}^{-1/2}\mathbb{E}[M_0\boldsymbol{X}_D]$ or on scaled portfolio weights $\boldsymbol{\Sigma}_{\boldsymbol{\zeta}}^{1/2}\boldsymbol{\theta}_{0D}$. Figure 5 documents these sparsity properties for the three minimum variance S–SDFs with 50% target GLS R^2 in Panel (B) of Figure 2.

In Panel (A) of Figure 5, the l_1-pricing norm gives rise to sparse excess return misvaluations. In contrast, the (scaled) $l_\infty-$norm produces extreme shrinkage and no sparsity. The l_2-norm induces shrinkage and no sparsity as well, but with maximal excess return misvaluations larger than under the (scaled) $l_\infty-$norm. The properties of S–SDF dual portfolio weights follow from the corresponding penalizations $\sigma_C = \tau\left\|\cdot\right\|_{*,\boldsymbol{\Sigma}_{\boldsymbol{\zeta}}}$ in Corollary 3. Since the dual norm of the l_1-norm is the $l_\infty-$norm, the l_1-pricing constraint produces extreme shrinkage in dual portfolio weights and no sparsity. Analogously, the (scaled) $l_\infty-$norm gives rise with a l_1-penalization to extreme sparsity in dual portfolio weights. Finally, the self-duality of the l_2-norm induces shrinkage but no sparsity in dual portfolio weights.

While linked to S–SDFs with different sparsity properties, the S–SDF time series in Panel (B) of Figure 5 are superficially quite similar in a number of dimensions.[37] However, they give rise to economically important differences in time series explanatory power for asset returns. As a consequence, the choice of a norm in the S–SDF pricing constraints and the corresponding S–SDF sparsity features are in the end strongly linked to a particular tradeoff between S–SDF cross-sectional and time series explanatory power. Our minimum dispersion APT S–SDF theory in Corollary 3 makes this tradeoff empirically measurable and indicates that S–SDF sparsity is quite costly with respect to the implied tradeoff between cross-sectional and time series explanatory powers under metrics (31) and (32).

[37] All S–SDF time series exhibit quite similar volatilities between about 1.75 and 1.85, as well as a quite substantial co-movement, with time series correlations above 0.8 between S–SDFs induced by the l_2-norm and the other S–SDFs.

5.6 APT S–SDFs, Non APT S–SDFs and Relation to Kozak et al. [2020]

When pricing accuracy is measured by the standard, not self-standardized, cross-sectional R^2 metric:

$$R^2_{OLS} := 1 - \frac{\left\|\mathbb{E}[M_0(\boldsymbol{X} - \bar{\boldsymbol{X}}\mathbf{1})]\right\|_2^2}{\left\|\mathbb{E}[\boldsymbol{X} - \bar{\boldsymbol{X}}\mathbf{1}]\right\|_2^2} \,, \tag{34}$$

the corresponding optimal minimum variance S–SDFs simply need to restrict the pricing with a constrained form of the l_2-pricing constraint (19), in which $\boldsymbol{\Sigma} := \boldsymbol{I}_{N\times N}$, i.e., such optimal S–SDFs are by construction non APT S–SDFs.

The two top panels of Figure 6 illustrate for the high dimensional dataset the optimal tradeoffs implied by minimum variance APT and non APT S–SDFs exactly pricing the first principal component of excess returns, under metrics of pricing accuracy (31) and metric (34), respectively.[38] Importantly, both such S–SDFs are sparse in two corresponding portfolios of excess returns: the first principal component of excess returns itself and a second optimal portfolio of returns orthogonal to traded systematic risk, which bounds the overall S–SDF mispricing with an APT bound (31) and with its constrained non APT version such that $\boldsymbol{\Sigma}_{\boldsymbol{\zeta}} := \boldsymbol{I}_{N_D\times N_D}$, respectively.

Kozak et al. [2020] study the sparsity implications of minimum variance S–SDFs in large asset markets. As elaborated in Section 3, their S–SDFs can be obtained with an Elastic Net norm penalization $\sigma_C = \alpha \left\|\cdot\right\|_1 + \tau \left\|\cdot\right\|_2$; see again Table 1. For $\alpha = 0$ these S–SDFs are not sparse and optimize the cross-sectional pricing accuracy under metric (34). In contrast, for $\alpha \neq 0$ they are sparse and give rise to a suboptimal tradeoff between the S–SDF time series explanatory power and pricing accuracy under metric (34). However, Kozak et al. [2020] show that this sparsity-induced suboptimality is less serious when it is imposed on the set of all principal components of returns, rather than on the original returns.

The bottom right panel of Figure 6 illustrates the degree of suboptimality of the tradeoff

[38] Analogous evidence for the low dimensional dataset is reported in Figure OA-2 of the Online Appendix.

between S–SDF time series explanatory power and pricing accuracy under metric (34), when sparsity is induced on the endogenously selected principal components in Kozak et al. [2020] S–SDFs. It additionally shows that the tradeoff of a non APT S–SDF exactly pricing the first principal component of excess returns alone is essentially identical to the one of the optimal non sparse Kozak et al. [2020] S–SDF, which depends on all principal components of excess returns. Finally, the bottom left panel of Figure 6 shows that principal component APT S–SDFs not forcing exact pricing of the first principal component are on average essentially uncorrelated with excess returns, i.e., they are unable to generate a tradeoff between S–SDF time series explanatory power and pricing accuracy.

In summary, our minimum variance APT and non APT S–SDFs exactly pricing the first principal component of excess returns provide an optimal tradeoff between pricing accuracy and time series explanatory power under metrics (31) and (34) of cross-sectional pricing accuracy. These S–SDFs are by construction sparse in the returns of two distinct portfolios, given by the first principal component of returns and a second optimal portfolio that bounds the mispricing across assets under l_2-APT pricing contraint (19) and under l_2-non APT pricing constraint (19) with $\mathbf{\Sigma}_{\boldsymbol{\zeta}} := \boldsymbol{I}_{N_D \times N_D}$, respectively.

5.7 Out-of-sample analysis

We next characterize the out-of-sample tradeoff between pricing accuracy and time series explanatory power generated by minimum variance APT S–SDFs.[39] To this end, we estimate data-driven out-of-sample minimum variance APT S–SDFs in a non forward-looking manner.[40] We then evaluate the out-of-sample S–SDF tradeoff between pricing accuracy and time series explanatory power according to metrics (31) and (32). Given our earlier results, we focus on minimum variance APT S–SDFs implying varying sparsity proper-

[39] We adopt a simple approach not requiring an explicit modelling of conditioning information. While the modelling of conditioning information is an important dimension we plan to address in future research, it would require a more extended treatment formally dealing with a possibly time varying constraint set C, a potentially large number of information variables, and a growing cross-sectional asset dimension.

[40] A forward-looking out-of-sample analysis, further supporting the evidence presented in this section, can be found in Online Appendix C.2.

ties under pricing constraints with norms $\|\cdot\|_1$, $\|\cdot\|_2$, and $\sqrt{N_D}\,\|\cdot\|_\infty$, and consider various choices for the set of traded factor returns, such as the market excess return alone.

To implement optimal minimum variance APT S–SDFs in a non forward-looking way, we propose a data-driven approach, which incorporates varying views on the desired in-sample S–SDF tradeoff between times series and cross-sectional explanatory power. These views may implicitly incorporate, e.g., the importance that a researcher/model builder assigns to these two tradeoff dimensions. For this purpose, we consider a sequence of estimation windows of 30 years updated at a semi-annual frequency, where in each 30-year window the optimal S–SDF dual portfolio weights are estimated and used to construct the out of sample S–SDFs for the 6 months following the estimation window. Specifically, given an estimation window of 30 years with last observation in month y, the vector of optimal portfolio weights, $\hat{\boldsymbol{\theta}}^\star_y$, is estimated by the solution of empirical optimization problem (30) for a corresponding optimal choice of the pricing threshold in the estimation window. The out-of-sample S–SDFs are then computed as $\hat{M}^\star_{y+m} = \max\{-\boldsymbol{X}'_{y+m}\hat{\boldsymbol{\theta}}^\star_y, 0\}$ for $m = 1, \ldots, 6$.[41] By rolling the 30-year window at a semi-annual frequency we obtain a monthly time series of data-driven APT S–SDFs for which we can evaluate the out-of-sample time series and cross-sectional explanatory power.

In order to select the optimal pricing threshold, we split each 30-year estimation window in a training window of 25 years and a separate window of 5 years. For a range of admissible relative pricing thresholds $\hat{\kappa}_y \in [\hat{\kappa}^{min}_y, 1]$ corresponding to the training window, we estimate various minimum variance APT S–SDFs using only the training data.[42] Based on the entire window of 30 years of data, we then compute the resulting curves of S–SDF metrics $R^2_{GLS}(\hat{\kappa}_y)$ and $R^2(\hat{\kappa}_y)$ defined in equations (31) and (32). Finally, we select an optimal pricing threshold $\hat{\kappa}^\star_y := \hat{\kappa}^\star_y(p)$, by maximizing cross-sectional metric $R^2_{GLS}(\hat{\kappa}_y)$ under a

[41] The specific shape of $\hat{M}^\star_{y+m}$ emerges from the closed-form link between minimum variance S–SDF and optimal dual portfolio payoff found in Corollary OA-2 of Online Appendix B.1.

[42] Recall that the minimum admissible relative threshold is given by $\hat{\kappa}^{min}_y = \hat{\tau}^{min}_y / \hat{\tau}^{max}_y$, where $\hat{\kappa}^{min}_y$ is the smallest threshold value for which in the training window no empirical duality failure arises and $\hat{\tau}^{max}_y$ is given by the empirical version of equation (26).

constraint on the average time series metric $\overline{R^2}(\hat{\kappa}_y)$ across assets. This constraint requires $\overline{R^2}(\hat{\kappa}_y)$ to be not lower than a fraction $p \in [0,1]$ of the average time series metric $\overline{R^2}(1)$ under an empirical SDF only pricing exactly the traded factor excess returns:

$$\hat{\kappa}_y^\star = \arg \max_{\hat{\kappa}_y \in [\hat{\kappa}_y^{min},1]} \{R^2_{GLS}(\hat{\kappa}_y) : \overline{R^2}(\hat{\kappa}_y) \geq p\overline{R^2}(1)\} \ . \tag{35}$$

We compute optimal threshold $\hat{\kappa}_y^\star$ for the choices $p \in \{0, 0.25, 0.5, 0.75, 1\}$, where the first choice corresponds to an optimization (35) with no constraint on the S–SDF time series explanatory power.

Figure 7 collects the out-of-sample evidence on data-driven minimum variance S–SDFs for various intuitive choices of the set of traded factor returns. Given our earlier results, we focus on l_2-APT pricing constraints. In the top two panels of Figure 7, we obtain an out-of-sample tradeoff between S–SDF time series and cross-sectional explanatory power that is consistent with the in-sample selection of the sequence of optimal thresholds $\{\hat{\kappa}_y^\star\}$. For instance, for the S–SDFs including the market excess return as the single traded factor, the out-of-sample GLS R^2 metric (31) in the low (intermediate) dimensional data set decreases monotonically with the tightness of the constraint on the in-sample time series explanatory power, yielding GLS R^2s between a maximum of 68% (38%) and a minimum of -2% (-2%) for parameters $p \in \{0, 0.25, 0.5, 0.75, 1\}$. Moreover, in all cases the maximum cross-sectional GLS R^2 is attained by APT S–SDFs with tuning parameter $\hat{\kappa}_y^\star > 0$, i.e., standard minimum variance SDFs produce a suboptimal cross-sectional fit. On the other hand, the out-of-sample average time series R^2s are increasing with the tightness of the in-sample constraint from a minimum of 2.8% (1%) to a maximum of 68% (62%). An analogous pattern arises for other choices of traded factors, given by the first principal component of returns and the three Fama-French factors, respectively.

The parameter choice $p = 1$ corresponds to minimum variance S–SDFs only required to price exactly the traded factors in-sample. Given the choice of the market excess return as the single traded factor, the resulting S–SDF produces a mildly negative cross-sectional

GLS R^2, together with a large average time series R^2 of 62% (68%) in the intermediate (low) dimensional dataset. In comparison, a minimum variance S–SDF only required to price exactly the three Fama-French factors produces an even more negative GLS R^2 and an average time series R^2 of only 18% (22%) in the intermediate (low) dimensional dataset. This explains why the minimum variance S–SDFs incorporating the three Fama-French factors as traded factors produce a Pareto dominated out-of-sample tradeoff between time series and cross-sectional explanatory power relative to the tradeoffs of S–SDFs incorporating the market alone as a traded factor.

A further Pareto improved tradeoff is achieved by minimum variance S–SDFs incorporating the first principal components of returns as a single traded factor in-sample. Interestingly, while these S–SDFs tend to produce larger average time series R^2s, due to the more consistent out-of-sample time series explanatory power of the first principal component of returns relative to market returns, the resulting tradeoff between time series and cross-sectional explanatory power is quite comparable to the one induced by the S–SDF pricing exactly market returns alone. Notably, this S–SDF is a simple minimum variance correction of an empirical SDF under the CAPM, which exactly prices market risk but otherwise bounds the amount of mispricing across assets with a standard APT pricing constraint.

Following the intuition underlying the APT, model-free APT S–SDFs able to partly span systematic return shocks need to correlate with potential common risk factors in asset returns. The bottom two panels of Figure 7 illustrate this intuition more precisely, by reporting the fraction of out-of-sample variation in model-free APT S–SDFs that is explained by variations of the first principal component of sorted portfolios excess returns. Consistent with our previous findings, APT S–SDFs maximizing the in-sample pricing accuracy produce the highest out-of-sample pricing accuracy, but they are also almost uncorrelated with the first principal component of excess returns out-of-sample. Conversely, APT S–SDFs only required to exactly price systematic risk exposures in-sample produce the lowest

out-of-sample pricing accuracy and the largest comovement with the first principal component of excess returns. Such comovement is naturally highest, and rather similar, for S–SDFs exactly pricing either market returns or the first principal component of sorted portfolios excess returns in-sample.

Finally, we also find that the APT S–SDFs in Figure 7 clearly outperform linear SDF specifications in the traded factor excess returns that are computed by globally minimizing the in-sample pricing error size according to a standard GMM criterion. As for the APT S–SDFs of Figure 7, we estimate such SDFs on rolling windows of 30 years of monthly data, while updating the resulting GMM estimates at a semi-annual frequency. We then evaluate their out-of-sample cross-sectional and time series explanatory power for different choices of the set of factors. In all cases, we obtain pairs of cross-sectional and time series explanatory power that are Pareto dominated by corresponding pairs induced by some minimum variance S–SDF in Figure 7. For instance, a CAPM SDF specification with a single market factor implies a cross-sectional GLS R^2 of -0.2% (0.5%) and an average time series R^2 of 51.8% (52%) in the low (intermediate) dimensional dataset. Similarly a three-factor Fama-French SDF specification implies a cross-sectional GLS R^2 of 2.8% (0.4%) and an average time series R^2 of 26.9% (26.3%) in the low (intermediate) dimensional dataset.

6 Conclusions

This paper introduces a unifying framework for analyzing and selecting Stochastic Discount Factors (SDFs) in arbitrage-free markets with convex pricing constraints. These SDFs have an arbitrage-free foundation in economies with frictions or economies with ambiguity. In economies with frictions, they represent a positive linear pricing rule that bounds from below the nonlinear pricing rule in the market with frictions. In frictionless economies with ambiguity, they represent the positive linear pricing rule of an unobserved marginal investor, but under a different probability belief than the marginal investor's belief. This arbitrage-free framework provides an independent foundation for several important asset

pricing theories, such as Ross [1976]' Arbitrage Pricing Theory (APT), the good-deal bound SDF theory in Cochrane and Saa-Requejo [2000], various approaches to SDF regularization such as Kozak et al. [2020], and robust approaches for the identification of investors' beliefs such as Chen et al. [2020].

Our SDF analysis and selection framework is based on minimum dispersion SDF problems subject to convex pricing constraints, which naturally extend the classical minimum variance SDF methodology in Hansen and Jagannathan [1991]. We establish a duality between minimum dispersion SDF problems and penalized dual portfolio selection problems, in which the portfolio penalization stays in a one-to-one relation with the SDF convex pricing constraints. Here, minimum dispersion SDFs are identified as simple transformation of the dual optimal payoff of a penalized portfolio, which gives rise to a SDF regularization that is economically interpretable in terms of an unobserved structure of frictions or ambiguity in a corresponding arbitrage-free market.

We use market data to extract minimum variance APT SDFs that summarize with a model-free approach the admissible tradeoffs between cross-sectional pricing accuracy and SDF co-movement with systematic return risks. We find that a simple minimum variance correction of a CAPM–SDF, in which the mispricing of risks orthogonal to market risk is bounded according to APT pricing constraints, produces a Pareto optimal tradeoff. Such Pareto optimal SDF corrections are founded in arbitrage-free frictionless markets with ambiguity and they depend on two economically distinct SDF factors alone. The first factor is a traded excess return that maximally correlates with systematic return risks. The second factor consists of a minimum variance portfolio excess return, which optimally bounds the mispricing of risks unspanned by the traded SDF factor. Therefore, the Pareto optimal tradeoffs attained by our minimum variance SDF corrections correspond to sparse SDFs dependent on the excess returns of just two economically interpretable portfolios.

Appendix A - Tables and Figures

Table 1: Convex pricing constraints and supporting cost functions.

Pricing constraint and supporting cost function	h	σ_C
Conic	δ_K	δ_{K^o}
Norm	$\|\cdot\|$	$\tau \|\cdot\|_*$
Self-standardized norm	$\|\cdot\|_{\boldsymbol{W}^{-1}}$	$\tau \|\cdot\|_{*,\boldsymbol{W}}$
Convex combination of norms	$(1-\lambda)\|\cdot\|^{(1)} + \lambda\|\cdot\|^{(2)}$	$\tau\left(\min_{\boldsymbol{z}\in\mathbb{R}^N}\left\{\max\left\{\frac{\|\cdot\|_*^{(1)}}{1-\lambda},\ \frac{\|\cdot-\boldsymbol{z}\|_*^{(2)}}{\lambda}\right\}\right\}\right)$
Euclidean distance from norm ball	$\mathrm{dist}_{\alpha B(\|\cdot\|)}$	$\alpha\|\cdot\|_* + \tau\|\cdot\|_2$
Short-sale constraint	$\delta_{\mathbb{R}_-^{N_D}}$	$\delta_{\mathbb{R}_+^N}$
l_2-norm	$\|\cdot\|_2$	$\tau\|\cdot\|_2$
l_1-norm	$\|\cdot\|_1$	$\tau\|\cdot\|_\infty$
Self-standardized $l_\infty-$norm	$\|\cdot\|_{\infty,\boldsymbol{W}^{-1}}$	$\tau\|\cdot\|_{1,\boldsymbol{W}}$
Convex combination of l_1- and l_2-norms	$(1-\lambda)\|\cdot\|_1 + \lambda\|\cdot\|_2$	$\tau\left(\min_{\boldsymbol{z}\in\mathbb{R}^N}\left\{\max\left\{\frac{\|\cdot\|_\infty}{1-\lambda},\ \frac{\|\cdot-\boldsymbol{z}\|_2}{\lambda}\right\}\right\}\right)$
Distance from $l_\infty-$ball	$\mathrm{dist}_{\alpha B_\infty}$	$\alpha\|\cdot\|_1 + \tau\|\cdot\|_2$

The table collects various specifications of function h parametrizing constraint set C, i.e., $C = \{\boldsymbol{\eta} \in \mathbb{R}^N \ : \ h(\boldsymbol{\eta}) \leq \tau\}$, with $\tau \geq 0$, and associated supporting cost functions σ_C, along with closed-form examples. From top to bottom, we consider: (i) conic pricing constraints obtained via characteristic functions of convex cone K and its polar cone K^o; for instance, short-sale constraints feature $K = \mathbb{R}_-^N$ and $K^o = \mathbb{R}_+^N$ while bid-ask spreads feature $K = \mathbb{R}_-^N \times \mathbb{R}_+^N$ and $K^o = \mathbb{R}_+^N \times \mathbb{R}_-^N$. (ii) norm constraints obtained via $h = \|\cdot\|$ with $\sigma_C = \tau\|\cdot\|_*$, where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$; for instance, the l_2-norm is self-dual, while the l_1- and $l_\infty-$norms are dual to each other. (iii) self-standardized norms defined via $\|\cdot\|_{\boldsymbol{W}^{-1}} := \left\|\boldsymbol{W}^{-1}\cdot\right\|$. (iv) Euclidean distance from a ball of radius α under norm $\|\cdot\|$ (i.e., $\alpha B(\|\cdot\|) := \{\boldsymbol{\eta} \in \mathbb{R}^N \ : \ \|\boldsymbol{\eta}\| \leq \alpha\}$) defined via $\mathrm{dist}_{\alpha B(\|\cdot\|)} := \inf_{\boldsymbol{\eta}\in\mathbb{R}^N}\{\|\cdot - \boldsymbol{\eta}\|_2 \ : \ \|\boldsymbol{\eta}\| \leq \alpha\}$. In the table, $\tau \geq 0$, $\lambda \in [0,1]$ and $\boldsymbol{W}$ denotes a generic symmetric positive definite weighting matrix.

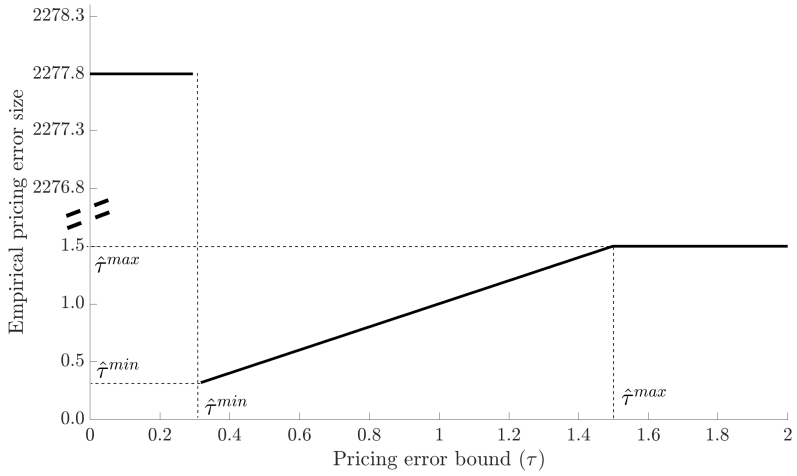
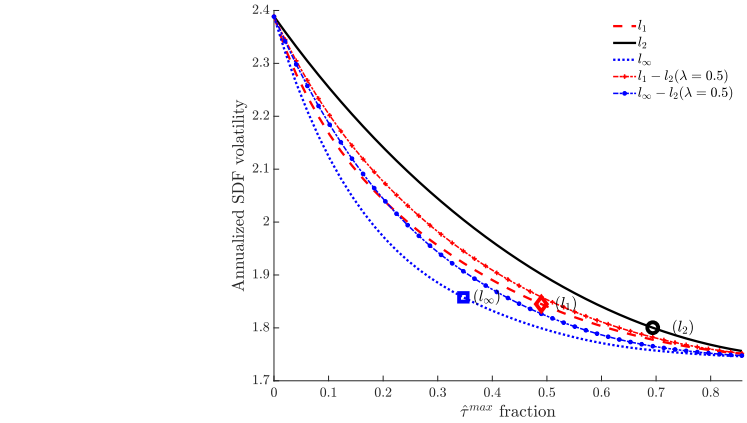
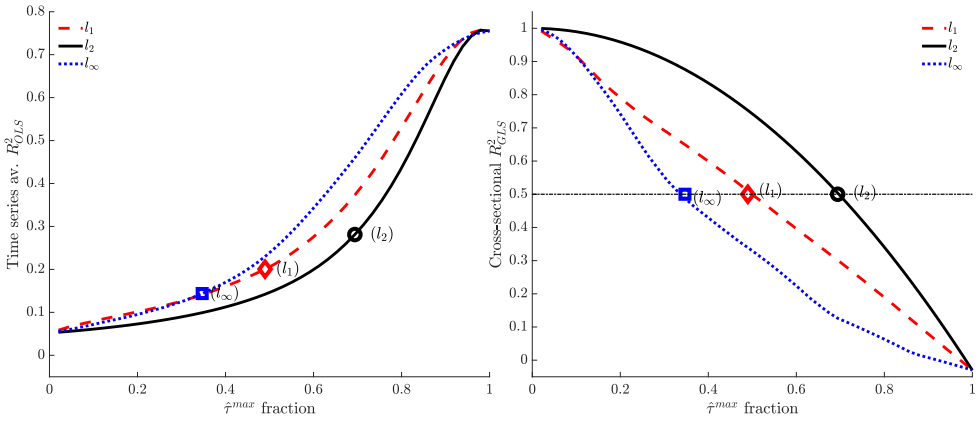


Figure 1: **Empirical duality failure.** We compute minimum variance S–SDFs satisfying standard APT pricing constraints. The figure reports the estimate of $\|\mathbb{E}[M_0\boldsymbol{X} - \boldsymbol{P}]\|_{2,\boldsymbol{\Sigma}^{-1/2}}$ for varying pricing thresholds τ. The point of discontinuity in the plot identifies the smallest threshold τ, for which a solution of the empirical primal S–SDF problem exists. The largest threshold $\hat{\tau}^{max}$ in the plot is computed as the sample version of the maximal threshold τ^{max} in equation (26). Sorted portfolio returns are used to construct the S–SDFs without assuming any observable traded factor. All calculations are based on the high dimensional dataset from January 1970 to June 2018.

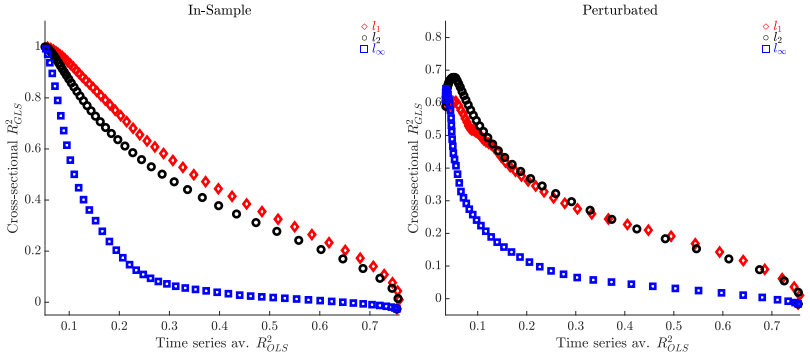


(1) **Panel (A)**

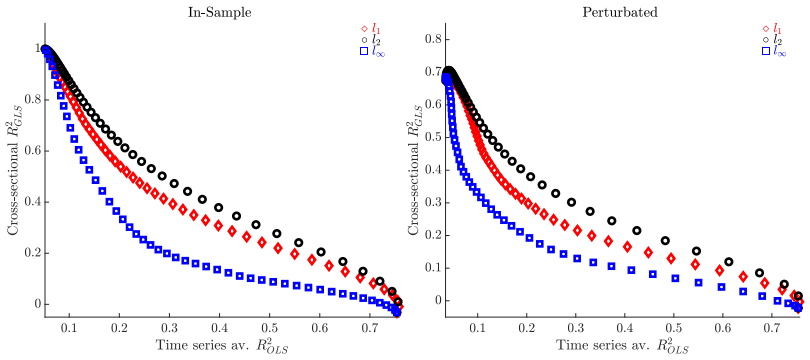


(2) **Panel (B)**

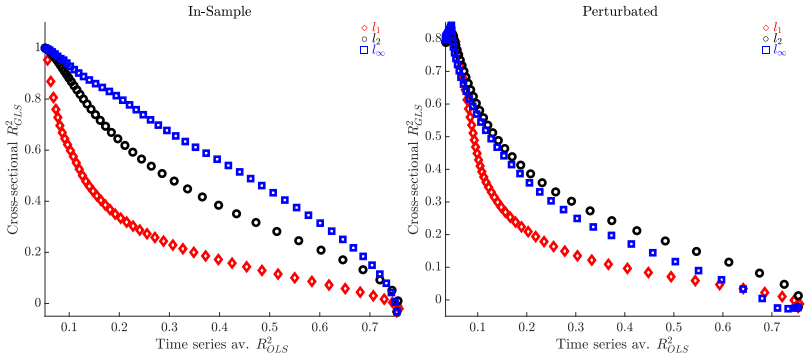
Figure 2: **Minimum APT S–SDF volatility curve and tradeoff between time-series and cross-sectional S–SDF explanatory power. Panel (A)** reports the annualized volatility of minimum variance APT S–SDFs as a function of the relative threshold $\kappa \in (\tau^{min}/\tau^{max}, 1]$, with threshold τ^{max} defined in equation (26). We consider the pricing metrics $h = \|\cdot\|_1$ (dashed line), $h = \|\cdot\|_2$ (solid line) and $h = \sqrt{N_D}\,\|\cdot\|_\infty$ (dotted line). The two additional curves report the minimum S–SDF volatility curve for pricing metrics $h = \lambda\,\|\cdot\|_1 + (1-\lambda)\,\|\cdot\|_2$ (dashed-crossed line) and $h = \sqrt{N_D}\lambda\,\|\cdot\|_\infty + (1-\lambda)\,\|\cdot\|_2$ (dashed dotted line), where $\lambda = 0.5$. **Panel (B)** reports the average time series R^2 metric (32) (left panel) and the cross-sectional GLS R^2 metric (31) (right panel) for the first three minimum variance S–SDFs from Panel (A). Finally, the diamonds, the circles, and the squares in Panel (B) identify the corresponding S–SDFs attaining a cross-sectional GLS R^2 of 50%. Sorted portfolio returns are used to construct the S–SDFs with traded risk factor given by the market return. All calculations are based on the low dimensional dataset from July 1963 to June 2018.



(1) **Panel (A)**

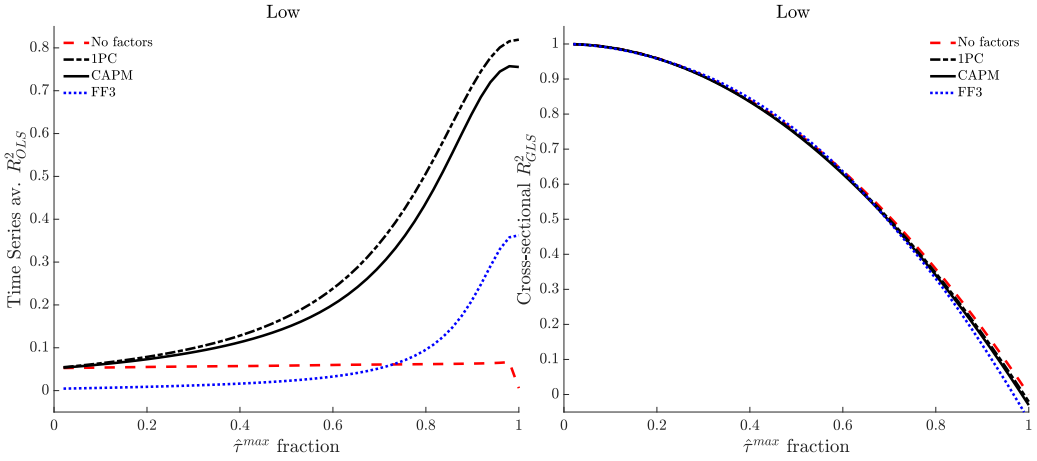


(2) **Panel (B)**

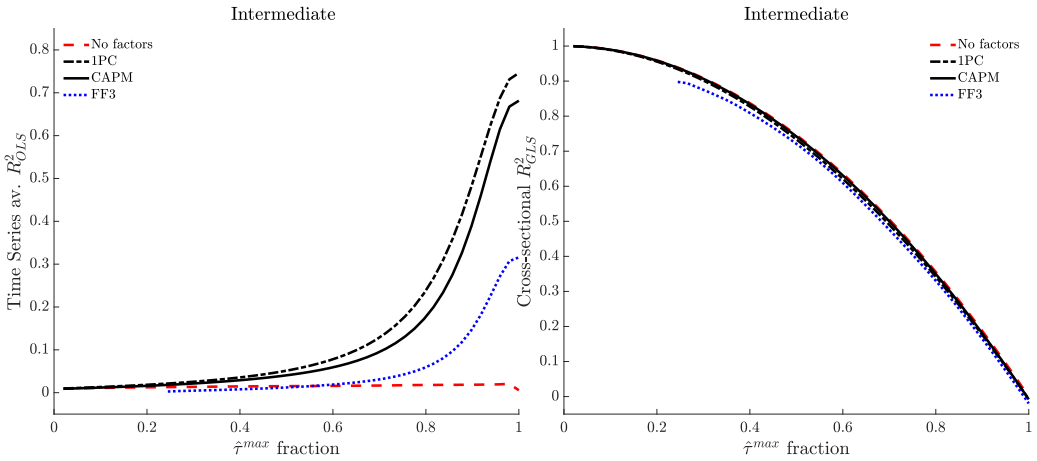


(3) **Panel (C)**

Figure 3: **Tradeoff between time series and cross-sectional APT S–SDF explanatory power.** The figure reports the average time-series R^2 metric (32) and cross-sectional GLS R^2 metric (33), using the l_1-norm in **Panel (A)**, the l_2-norm in **Panel (B)**, and the (scaled) $l_\infty-$norm in **Panel (C)**, for minimum variance APT S–SDFs based on a l_1-, l_2-, and (scaled) $l_\infty-$pricing metric, while taking the market excess return as the single traded risk factor. Left panels report in-sample results with S–SDFs estimated and evaluated on the same 40 year window 1965-2005, while right panels report results with S–SDFs estimated on 40 year window 1965-2005 and evaluated on 40 year window 1975-2015. All computations are based on the low dimensional dataset.

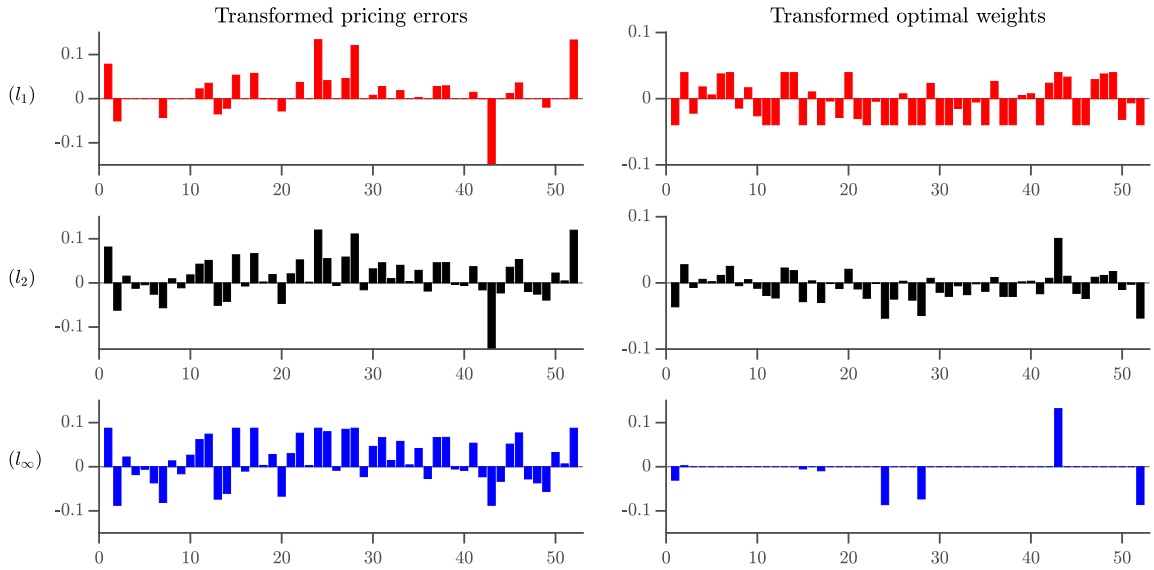


(1) **Panel (A)**

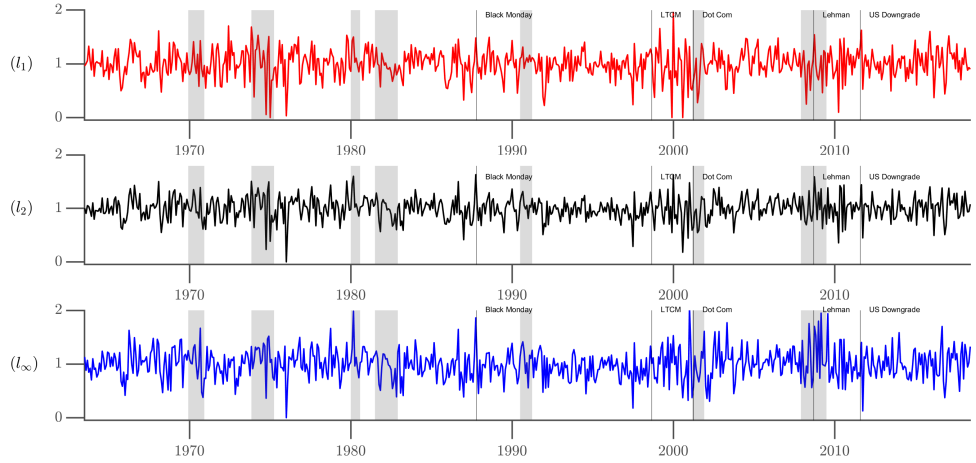


(2) **Panel (B)**

Figure 4: **Role of traded factors in minimum dispersion APT S–SDFs.** The figure reports the average time-series R^2 metric (32) (left panel) and cross-sectional GLS R^2 metric (31) for various minimum variance APT S–SDFs, corresponding to different choices for the set of traded risk factors, in the low dimensional (**Panel (A)**) and intermediate dimensional (**Panel (B)**) datasets. We consider following choices for the traded factors: (1) No factors; (2) CAPM, corresponding to the market return as the single traded factor; (3) FF3, corresponding to the three Fama-French factor returns and traded factors; (4) 1PC, corresponding to the first principal component of excess returns as the single traded factor. All calculations are for APT S–SDFs based on the standard l_2-pricing metric. Both the low and intermediate dimensional datasets run from July 1963 to June 2018.

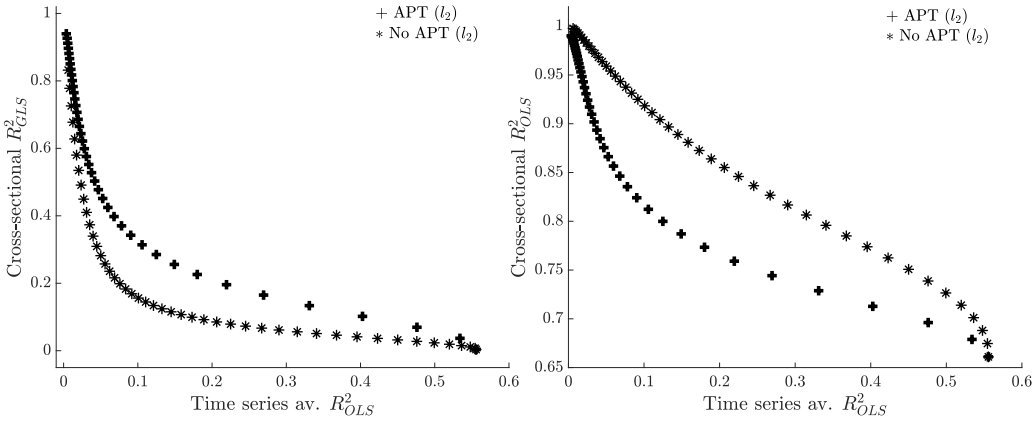


(1) **Panel (A)**

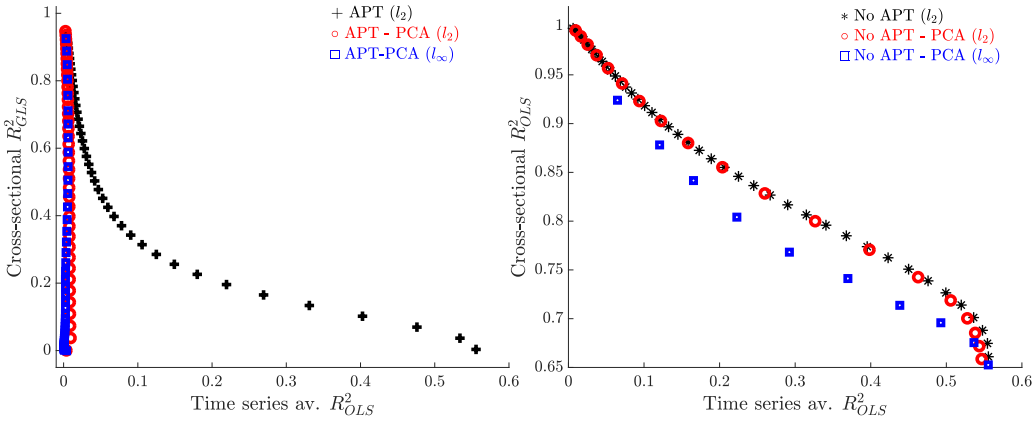


(2) **Panel (B)**

Figure 5: **Properties of APT S–SDFs. Panel (A)** reports, from top to bottom, $\mathbf{\Sigma}_{\boldsymbol{\zeta}}^{-1/2}-$transformed pricing errors (left) and $\mathbf{\Sigma}_{\boldsymbol{\zeta}}^{1/2}-$transformed optimal portfolio weights (right) of the three minimun variance APT S–SDFs highlighted with a diamond, a circle and a square in Panel (B) of Figure 2, which correspond to a l_1- (first row), l_2- (second row), and scaled $l_\infty-$ (third row) APT pricing metric, respectively. **Panel (B)** reports the time-series of these three minimum variance APT S–SDFs. All calculations are based on the low dimensional dataset from July 1963 to June 2018, taking the market excess return as the single traded factor. Grey shaded areas in Panel (B) highlight NBER recession periods.

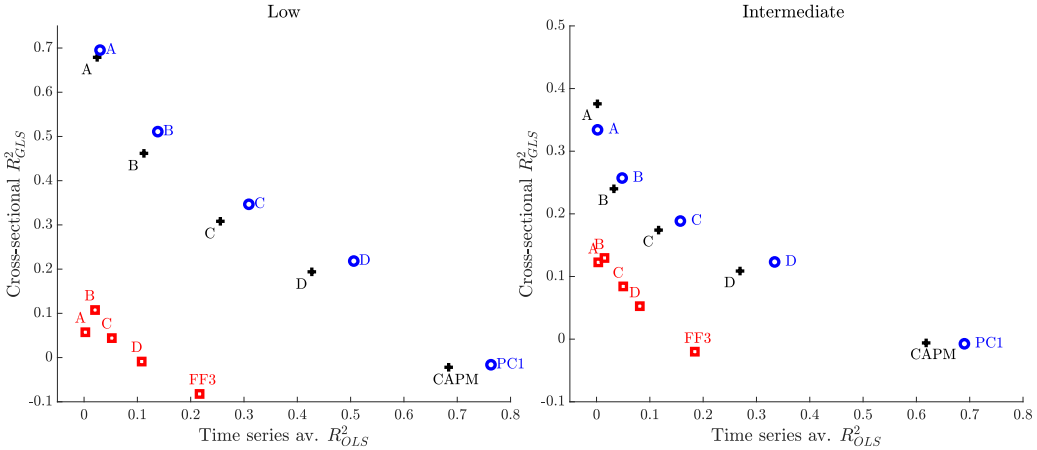


(1) **Panel (A)**

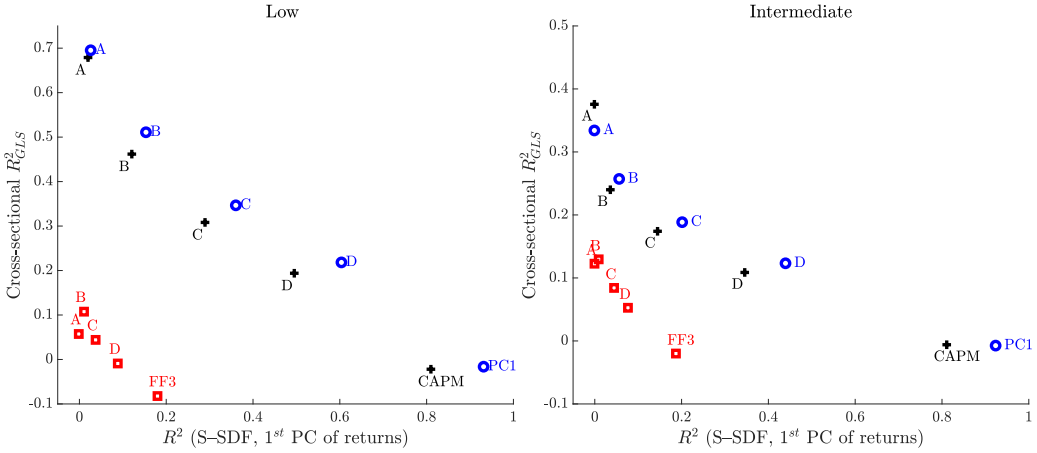


(2) **Panel (B)**

Figure 6: **Tradeoff between time-series and cross-sectional S–SDF explanatory power of APT S–SDFs, non APT S–SDFs, and principal components based S–SDFs.** For various minimum variance S–SDFs, the figure reports the tradeoff between average time series R^2 metric (32) and cross-sectional GLS R^2 metric (31) (denoted by R^2_{GLS}, left panels) or cross-sectional R^2 metric (34) (denoted by R^2_{OLS}, right panels). The following S–SDFs are compared in Panel (A). First, APT S–SDFs based on the l_2-pricing metric. Second, non APT S–SDFs based on the l_2-pricing metric, but with a diagonal weighting matrix $\boldsymbol{\Sigma}_{\boldsymbol{\eta}} = \boldsymbol{I}_{N_D \times N_D}$. Both these S–SDFs take the first principal component of excess returns as the single traded factor. In each figure of Panel (B), results for two additional principal components based S–SDFs are reported, which impose the corresponding pricing constraint on the principal components of sorted portfolio excess returns, instead of the original excess returns, without incorporating any traded factor. The first and second of these S-SDFs impose in the left (right) figure an APT l_2- and $l_\infty-$ pricing constraint (a non APT l_2- and $l_\infty-$ pricing constraint based on a diagonal weighting matrix $\boldsymbol{\Sigma}_{\boldsymbol{\eta}} = \boldsymbol{I}_{N_D \times N_D}$), respectively. All results are based on the high dimensional dataset running from 1970 to 2018.



(1) **Panel (A)**



(2) **Panel (B)**

Figure 7: **Out-of-sample tradeoff between time-series and cross-sectional explanatory power of optimal data-driven APT S–SDFs.** The figure reports in **Panel (A)** (**Panel (B)**) out-of-sample average time-series R^2 metric (32) (out-of-sample time-series R^2 metric (32) with respect to the first principal component of excess returns) and out-of-sample cross-sectional GLS R^2 metric (31), for the optimal data-driven minimum variance APT S–SDFs with l_2-pricing metric from Section 5.7. These S–SDFs are estimated on rolling windows of 30 years by maximizing the in-sample cross-sectional metric $R^2_{GLS}(\hat{\kappa}_y)$ in constrained optimization problem (35), for parameters $p = 0$ (label A), $p = 0.25$ (label B), $p = 0.5$ (label C), $p = 0.75$ (label D), and $p = 1$. APT S–SDFs are based on three choices for the traded factor returns: (1) the market excess return, (2) the three Fama-French factor excess returns, and (3) the first principal component of sorted portfolio excess returns. Label CAPM corresponds to case (1) for $p = 1$, label FF3 to case (2) for $p = 1$ and label PC1 to case (3) for $p = 1$. Results are based on the low dimensional (left panels) and the intermediate dimensional (right panels) datasets.

References

Caio Almeida and René Garcia. Economic implications of nonlinear pricing kernels. *Management Science*, 63(10):3361–3380, 2016.

Fernando Alvarez and Urban J Jermann. Using asset prices to measure the persistence of the marginal utility of wealth. *Econometrica*, 73(6):1977–2016, 2005.

Donald WK Andrews. Empirical process methods in econometrics. *Handbook of econometrics*, 4:2247–2294, 1994.

David Backus, Mikhail Chernov, and Stanley Zin. Sources of entropy in representative agent models. *Journal of Finance*, 64(1):51–99, 2014.

Ravi Bansal and Bruce N Lehmann. Growth-optimal portfolio restrictions on asset pricing models. *Macroeconomic Dynamics*, 108(1):333–354, 1997.

Gary Chamberlain. Funds, factors and diversification in arbitrage pricing models. *Econometrica*, 51(5):1305–1324, 1983.

Gary Chamberlain and Michael Rothschild. Arbitrage, factor structure and mean-variance analysis on large asset markets. *Econometrica*, 51(5):1281–1304, 1983.

Xiaohong Chen, Lars Peter Hansen, and Peter G Hansen. Robust identification of investor beliefs. Technical report, 2020.

Stephen A Clark. The valuation problem in arbitrage price theory. *Journal of Mathematical Economics*, 22(5):463–478, 1993.

John H Cochrane and Jesus Saa-Requejo. Beyond arbitrage: Good-deal asset price bounds in incomplete markets. *Journal of Political Economy*, 108(1):79–119, 2000.

Noel Cressie and Timothy RC Read. Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 46(3):440–464, 1984.

Guanhao Feng, Stefano Giglio, and Dacheng Xiu. Taming the factor zoo: A test of new factors. *Journal of Finance*, 75(3):1327–1370, 2020.

Joachim Freyberger, Andreas Neuhierl, and Michael Weber. Dissecting characteristics nonparametrically. *The Review of Financial Studies*, 33(5):2326–2377, 2020.

Shihao Gu, Brian Kelly, and Dacheng Xiu. Autoencoder asset pricing models. *Journal of Econometrics, forthcoming*, 2020a.

Shihao Gu, Bryan T Kelly, and Dacheng Xiu. Empirical asset pricing via machine learning. *Review of Financial Studies*, 33(5):2223–2273, 2020b.

Lars Peter Hansen and Ravi Jagannathan. Implications of security market data for models of dynamic economies. *Journal of Political Economy*, 99(2):225–262, 1991.

Lars Peter Hansen and Thomas J Sargent. *Robustness*. Princeton University Press, 2007.

J Michael Harrison and David M Kreps. Martingales and arbitrage in multiperiod securities markets. *Journal of Economic theory*, 20(3):381–408, 1979.

Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of convex analysis*. Springer Science & Business Media, 2012.

Gur Hubermann. A simple approach to the arbitrage pricing theory. *Journal of Economic Theory*, 28(1):183–191, 1982.

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

Elyes Jouini and Hédi Kallal. Martingales and arbitrage in securities markets with transaction costs. *Journal of Economic Theory*, 66(1):178–197, 1995.

Serhiy Kozak, Stefan Nagel, and Shrihari Santosh. Shrinking the cross-section. *Journal of Financial Economics*, 135(2):271–292, 2020.

Jonathan Lewellen, Stefan Nagel, and Jay Shanken. A skeptical appraisal of asset pricing tests. *Journal of Financial Economics*, 96(2):175–194, 2010.

David G Luenberger. *Optimization by vector space methods*. John Wiley & Sons, 1997.

Erzo GJ Luttmer. Asset pricing in economies with frictions. *Econometrica*, 64(6):1439–1467, 1996.

Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557–119, 2012.

Valentina Raponi, Raman Uppal, and Paolo Zaffaroni. Portfolio choice with model misspecification: a foundation for alpha and beta portfolios. 2018.

Stephen Ross. The arbitrage theory of capital asset pricing. *Journal of Economic Theory*, 13(3):341–360, 1976.

Stephen Ross. Mutual fund separation in financial theory – the separating distributions. *Journal of Economic Theory*, 17(2):254–286, 1978.

Karl Snow. Diagnosing asset pricing models using the distribution of asset returns. *Journal of Finance*, 46(3):955–983, 1991.

Michael Stutzer. A bayesian approach to diagnosis of asset pricing models. *Journal of Econometrics*, 68(2):367–397, 1995.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 58(1):267–288, 1996.

Raman Uppal, Paolo Zaffaroni, and Irina Zviadadze. Correcting misspecified stochastic discount factors. Technical report, EDHEC Business School, HEC Paris and Imperial College, 2019.

Smart Stochastic Discount Factors

ONLINE APPENDIX

SOFONIAS ALEMU KORSAYE, ALBERTO QUAINI and FABIO TROJANI*

First version: May 2019. This version: July 15, 2021

Keywords: SDF, Convex Pricing Constraints, Minimum Dispersion SDF, Market Frictions, SDF regularization, Arbitrage Pricing Theory,

*Sofonias Alemu Korsaye (email: Sofonias.Korsaye@unige.ch) and Fabio Trojani (email: Fabio.Trojani@unige.ch) are with University of Geneva, Geneva Finance Research Institute & Swiss Finance Institute. Alberto Quaini (email: Alberto.Quaini@unige.ch) is with University of Geneva, Geneva Finance Research Institute. We thank Caio Almeida, Federico Bandi, Tony Berrada, Federico Carlini, Ines Chaieb, George Constantinides, Jerome Detemple, Patrick Gagliardini, Eric Ghysels, Lars Hansen, Oliver Linton, Stefan Nagel, Paolo Porchia, Olivier Scaillet, Paul Schneider, Andrea Vedolin, Michael Weber, Dacheng Xiu, conference participants at the 2019 SoFiE annual Meeting in Shanghai, the European Econometric Society Meeting in Manchester, the Workshop on Big Data and Economic Forecasting in Ispra, the Conference on Quantitative Finance and Financial Econometrics in Marseille, the Financial Econometrics Conference in Toulouse, the Vienna Congress on Mathematical Finance, the Swiss Finance Institute Research days, the ESSEC Workshop on Monte Carlo Methods and Approximate Dynamic Programming, the Paris December Finance Meeting, the 2020 Virtual World Congress of the Econometric Society in Milano, the Remote Seminar Series on Computational Economics and Finance, the Virtual Derivatives Workshop and seminar participants at BI Norwegian Business School, Bocconi University, Mc Gill University, University of Zurich, University of Lugano, University of Geneva, University of Lund, JRC in Ispra and Luiss University. All errors are ours.

In this Online Appendix, we collect the proofs of the theoretical statements in the main text, additional theoretical findings and useful auxiliary results to the theory in the main text, and further empirical evidence on S–SDFs complementing the one presented in the main text. Section A first reports proofs of propositions and corollaries in the main text. It then collects auxiliary results not fully discussed in the main text, together with the corresponding proofs. Section B summarizes the properties of $\phi-$dispersions in the Cressie-Read family, details further areas of application of our S–SDF framework, and provides the foundations of S–SDFs in viable markets with convex cost function σ. Section C collects complementary empirical evidence on minimum dispersion APT S–SDFs. Section D collects figures illustrating the additional empirical evidence in Section C.

A ONLINE APPENDIX – Proofs

Proof of Proposition 1. By [Rockafellar, 1970, Thm. 13.2] σ_C given in equation (3) of the main text is proper, closed and sublinear.[1] This implies that the pricing functional π is sublinear and that the set of payoffs, $\mathcal{Z}$ is a convex cone. Hence, by [Chen, 2001, Thm. 1, 5 and Cor. 1], absence of free lunches is equivalent to the existence of an almost surely strictly positive SDF, M, such that $\mathbb{E}[MZ] \leq \pi(Z)$ for any $Z \in \mathcal{Z}$.[2]

By definition of cost functional π and payoff space $\mathcal{Z}$, it equivalently follows for any $\boldsymbol{\theta} \in \mathbb{R}^N$ that $\boldsymbol{\theta}'\boldsymbol{P} + \sigma_C(\boldsymbol{\theta}) \geq \mathbb{E}[M\boldsymbol{\theta}'\boldsymbol{X}]$, i.e., $\boldsymbol{\theta}'\left(\mathbb{E}[M\boldsymbol{X}] - \boldsymbol{P}\right) \leq \sigma_C(\boldsymbol{\theta})$. By [Bauschke and Combettes, 2011, Prop. 13.10 (i)] this inequality holds if and only if

$$(\sigma_C)^*(\mathbb{E}[M\boldsymbol{X}] - \boldsymbol{P}) \leq 0 \ . \tag{OA-1}$$

By [Rockafellar, 1970, Thm. 13.2], the convex conjugate σ_C^* is given by the character-

[1] Properness follows since constrained set C is not empty.

[2] Technically, in order to use [Chen, 2001, Thm. 5], we require that there exists a traded payoff that is strictly positive almost surely. Such assumption is easily satisfied, e.g., when there exists a risk-free asset with positive risk-free payoff.

istic function δ_C, i.e.,

$$\sigma_C^*(\mathbb{E}[M\boldsymbol{X}] - \boldsymbol{P}) = \begin{cases} 0 & \mathbb{E}[M\boldsymbol{X}] - \boldsymbol{P} \in C \\ +\infty & \text{else} \end{cases} .$$

Thus, inequality (OA-1) is equivalent to condition $\mathbb{E}[M\boldsymbol{X}] - \boldsymbol{P} \in C$. □

Proof of Corollary 1. By Proposition 1, absence of free lunches in market $(\tilde{Z}_{\{\mathbf{0}\}}, \tilde{\pi}_{\{\mathbf{0}\}}, \tilde{\mathbb{P}})$ is equivalent to the existence of a strictly positive SDF $\tilde{M}$ such that $\tilde{\mathbb{E}}[\tilde{M}\boldsymbol{X}] = \boldsymbol{P}$. Moreover, the absence of free lunches in market $(Z_C, \pi_C, \mathbb{P})$ is equivalent to the existence of a strictly positive SDF M such that $\mathbb{E}[M\boldsymbol{X}] - \boldsymbol{P} \in C$.

□

Proof of Proposition 2. We first prove that $\Pi(C) = -\Delta(C)$. To do so, we rewrite problem $\Pi(C)$ according to the notation given in Bauschke et al. [2017] and obtain its Fenchel-Rockafellar's dual problem. Let $A : L^q \to \mathbb{R}^N$ be the linear function $A(M) := \mathbb{E}[M\boldsymbol{X}]$, and let $g : \mathbb{R}^N \to (-\infty, +\infty]$ denote the function $g(\boldsymbol{\eta}) := \delta_R(\boldsymbol{\eta})$ with $R := \boldsymbol{P} + C = \{\boldsymbol{P} + \boldsymbol{\eta} \ : \ \boldsymbol{\eta} \in C\}$. With the notation $I_f := \mathbb{E}[f(\cdot)]$ for any function $f : \mathbb{R} \to [-\infty, +\infty]$, it follows that $I_{\phi_+}(M) = I_\phi(M) + \delta_{L^q_+}(M)$.[3] Thus we can write:

$$\Pi(C) = \inf_{M \in L^q} \{I_{\phi_+}(M) + g(A(M))\} .$$

As payoffs are in L^p with $1/p + 1/q = 1$, A is continuous, while the properties of ϕ_+ imply that I_{ϕ_+} is a closed convex function. Since C is convex, closed and non-empty, g is convex, closed and proper. In view of these properties, we can obtain the dual problem of Π via [Bauschke et al., 2017, Thm. 15.23]. In order to apply the mentioned theorem, [Bauschke

[3] Remember that ϕ_+ is the restriction of ϕ to $\mathbb{R}_+$, i.e., $\phi_+(x) = \phi(x)$ if $x \geq 0$ and $\phi_+(x) = +\infty$ if $x < 0$.

et al., 2017, Prop. 15.24] shows that we only need to check that:[4]

$$\mathrm{ri}(\mathrm{dom}(g)) \cap A[\mathrm{qri}\ \mathrm{dom}(I_{\phi_+})] \neq \emptyset\ . \tag{OA-2}$$

As shown in [Borwein and Lewis, 1991, Cor. 2.6], our requirement $(0, +\infty) \subset \mathrm{dom}\,\phi$ implies that $A(\mathrm{dom}\, I_\phi \cap L^q_{++}) \subset A[\mathrm{qri}(\mathrm{dom}\, I_{\phi_+})]$. Hence, since $\mathrm{dom}(g) = R$, showing that $A(\mathrm{dom}\, I_\phi \cap L^q_{++}) \cap \mathrm{ri}(R) \neq \emptyset$ is enough to prove (OA-2). Under the assumption that the economy $(\mathcal{Z}_{\tilde{C}}, \pi_{\tilde{C}}, \mathbb{P})$ admits no free lunches, with $\tilde{C} \subset \mathrm{ri}(C)$, by Proposition 1 there exists $\tilde{M} \in \mathrm{dom}\, I_\phi \cap L^q_{++}$, hence $A(\tilde{M}) \in A(\mathrm{dom}\, I_\phi \cap L^q_{++})$. Moreover, $\mathbb{E}[\tilde{M}\boldsymbol{X}] \in \boldsymbol{P} + \tilde{C}$, i.e., $A(\tilde{M}) \in \boldsymbol{P} + \tilde{C} \subset \mathrm{ri}(R)$ holds and condition (OA-2) is satisfied.

Thus, by [Bauschke et al., 2017, Thm. 15.23 and 15.24] we obtain $\Pi(C) = -\Delta(C)$ where

$$\Delta(C) = -\min_{\boldsymbol{\theta} \in \mathbb{R}^N} \{I^*_{\phi_+}(-{}^tA(\boldsymbol{\theta})) + g^*(\boldsymbol{\theta})\}\ , \tag{OA-3}$$

where $I^*_{\phi_+} : L^p \to [0, +\infty]$ is the conjugate function of I_{ϕ_+} and ${}^tA : \mathbb{R}^N \to L^p$ is the adjoint map of A, given by ${}^tA(\boldsymbol{\theta}) = \boldsymbol{X}'\boldsymbol{\theta}$.[5] As ϕ_+ is convex and closed, we can apply [Rockafellar, 1968, Thm. 2] to obtain $I^*_{\phi_+} = I_{\phi^*_+}$. Moreover, for every $\boldsymbol{\theta} \in \mathbb{R}^N$,

$$\begin{aligned} g^*(-\boldsymbol{\theta}) &= \delta^*_R(-\boldsymbol{\theta}) \\ &= \sup_{\boldsymbol{\eta} \in \mathbb{R}^N} \left\{-\boldsymbol{\theta}'\boldsymbol{\eta}\ :\ \boldsymbol{\eta} \in \boldsymbol{P} + C\right\} \\ &= -\boldsymbol{P}'\boldsymbol{\theta} + \sigma_C(-\boldsymbol{\theta})\ . \end{aligned}$$

Thus (OA-3) reads:

$$\Delta(C) = -\min_{\boldsymbol{\theta} \in \mathbb{R}^N} \left\{\mathbb{E}[\phi^*_+(-\boldsymbol{X}'\boldsymbol{\theta})] + \boldsymbol{P}'\boldsymbol{\theta} + \sigma_C(\boldsymbol{\theta})\right\}\ ,$$

[4] See [Bauschke et al., 2017, Def. 6.9] for the definition of relative interior, ri, and quasi relative interior, qri.

[5] The adjoint map of A, ${}^tA$, is characterized by the identity $\mathbb{E}[{}^tA(\boldsymbol{\theta})M] = \boldsymbol{\theta}'\mathbb{E}[M\boldsymbol{X}]$, for each $M \in L^q$ and each portfolio weights $\boldsymbol{\theta} \in \mathbb{R}^N$.

thereby proving strong duality between $\Pi(C)$ and $-\Delta(C)$.

We now prove the relation between the optimal solutions of $\Pi(C)$ and $\Delta(C)$ given in equation (12) of the main text. By [Bauschke et al., 2017, Prop. 19.4], if $\boldsymbol{\theta}_0$ is a dual solution such that $I_{\phi_+^*}$ is Gateaux differentiable in $-\boldsymbol{X}'\boldsymbol{\theta}_0$, then the primal problem either has no solution or it has a unique solution given by $\nabla I_{\phi_+^*}(-\boldsymbol{X}'\boldsymbol{\theta}_0)$. Hereafter we show that indeed $I_{\phi_+^*}$ is Gateaux differentiable and that the derivative is given by $M_0 = (\phi_+^*)'(-\boldsymbol{X}'\boldsymbol{\theta}_0)$.

The assumption that $-\boldsymbol{X}'\boldsymbol{\theta}_0 < \lim_{y\to\infty} \phi(y)/y$ a.s., by [Borwein and Lewis, 1991, Lem. 4.2], implies that $-\boldsymbol{X}'\boldsymbol{\theta}_0$ belongs a.s. to the interior of the domain of ϕ_+^*. This and the strict convexity of ϕ_+ on its domain imply that ϕ_+^* is differentiable a.s. in $-\boldsymbol{X}'\boldsymbol{\theta}_0$, see [Borwein and Lewis, 1991, Thm. 4.6]. Let us denote such derivative by $M_0(\omega) := (\phi_+^*)'(-\boldsymbol{X}'(\omega)\boldsymbol{\theta}_0)$. Hereafter in order to prove that $I_{\phi_+^*}$ is Gateaux differentiable in $-\boldsymbol{X}'\boldsymbol{\theta}_0$, we show that M_0 is the unique element of $\partial I_{\phi_+^*}(-\boldsymbol{X}'\boldsymbol{\theta}_0)$. Now let us consider any element $\bar{M} \in \partial I_{\phi_+^*}(-\boldsymbol{X}'\boldsymbol{\theta}_0)$, then by [Rockafellar, 1970, Thm. 23.5]:

$$I_{\phi_+^*}(-\boldsymbol{X}'\boldsymbol{\theta}_0) + I_{\phi_+}(\bar{M}) = \langle -\boldsymbol{X}'\boldsymbol{\theta}_0, \bar{M} \rangle \ .$$

Explicitly, this gives $\int \left[\phi_+^*(-\boldsymbol{X}'(\omega)\boldsymbol{\theta}_0) + \phi_+(\bar{M}(\omega)) + \boldsymbol{X}'(\omega)\boldsymbol{\theta}_0\bar{M}(\omega)\right] d\mathbb{P}(\omega) = 0$. Here, the integrand, since by Fenchel's inequality is nonnegative, is zero a.s.. Applying again [Rockafellar, 1970, Thm. 23.5] to this integrand, we obtain $\bar{M}(\omega) \in \partial\phi_+^*({}^tA(-\boldsymbol{\theta}_0)(\omega))$ a.s.. However, since ϕ_+^* is differentiable a.s. in $-\boldsymbol{X}'\boldsymbol{\theta}_0$, $\partial\phi_+^*({}^tA(-\boldsymbol{\theta}_0)(\omega))$ has a unique element. Hence, we have that $\bar{M} = M_0$ a.s., which proves that $\partial I_{\phi_+^*}(-\boldsymbol{X}'\boldsymbol{\theta}_0)$ has a unique element and that $I_{\phi_+^*}$ is Gateaux differentiable in $-\boldsymbol{X}'\boldsymbol{\theta}_0$ with derivative given by $(\phi_+^*)'(-\boldsymbol{X}'\boldsymbol{\theta}_0)$. This concludes the proof. □

Proof of Corollary 2. Given constraint set $C = \{\mathbf{0}_{N_S+1}\} \times \{\boldsymbol{\eta} \in \mathbb{R}^{N_D} : \|\boldsymbol{\eta}\|_{2,\boldsymbol{\Sigma}_\zeta^{-1/2}} \leq \tau\}$, the associated penalization function is given by

$$\sigma_C(\boldsymbol{\theta}) = \sup_{\boldsymbol{\eta}\in C} \boldsymbol{\eta}'\boldsymbol{\theta} = \tau \|\boldsymbol{\theta}_D\|_{2,\boldsymbol{\Sigma}_\zeta^{1/2}}$$

using the definition of dual norm and the self-conjugation of the l_2-norm. Clearly, $\text{dom}(\sigma_C) = \mathbb{R}^N$, hence by [Clark, 1993, Thm. 6], the no free lunches condition in market $(\mathcal{Z}_C, \pi_C, \mathbb{P})$ is equivalent to the standard no arbitrage condition. Moreover, by Proposition 1, absence of free lunches in this market is equivalent to the existence of strictly positive S–SDF, M_{APT}, such that $\mathbb{E}[M_{APT}] = 1$, $\mathbb{E}[M_{APT}\boldsymbol{F}^e] = \mathbf{0}$ and $\|\mathbb{E}[M_{APT}\boldsymbol{R}^e]\|_{2,\boldsymbol{\Sigma}_{\boldsymbol{\zeta}}^{-1/2}} \leq \tau$. □

Proof of Corollary 3. The relative interior of $C = \{\mathbf{0}_{N_S+1}\} \times \{\boldsymbol{\eta} \in \mathbb{R}^{N_D} : \|\boldsymbol{\eta}\|_{\Sigma_{\boldsymbol{\zeta}}^{-1/2}} \leq \tau\}$ contains $\{\mathbf{0}_{N_S+1}\} \times \{\boldsymbol{\eta} \in \mathbb{R}^{N_D} : \|\boldsymbol{\eta}\|_{2,\Sigma_{\boldsymbol{\zeta}}^{-1/2}} < \tau\}$, which in turn contains $\tilde{C}$. Since $\tilde{C}$ is convex, closed and contains $\mathbf{0}$, the result follows from Proposition 2, where in dual portfolio problem $\Delta(C)$ the penalization function is given by

$$\sigma_C(\theta) = \sup_{\boldsymbol{\eta} \in C} \boldsymbol{\eta}'\boldsymbol{\theta} = \tau \|\boldsymbol{\theta}_D\|_{*,\boldsymbol{\Sigma}_{\boldsymbol{\zeta}}^{1/2}} \ .$$

□

Proposition OA-1 (Equivalence with dispersion constrained minimum pricing error problems). *Let* $C_\tau := \{\boldsymbol{\eta} \in \mathbb{R}^N \ : \ h(\boldsymbol{\eta}) \leq \tau\}$ *where* $\tau \in \mathbb{R}$ *and* h *is a convex, proper and closed function. Suppose that* $\Pi(C_\tau)$ *and*

$$\mathcal{P}(\nu) := \inf_{M \in \mathcal{M}_+} \{h(\mathbb{E}[M\boldsymbol{X} - \boldsymbol{P}]) \ : \ \mathbb{E}[\phi(M)] \leq \nu\} \tag{OA-4}$$

with $\nu \geq 0$ *are finite and attained. (i) If* M_0 *solves* $\Pi(C_\tau)$*, then there exists* $\nu_0 \geq 0$ *such that* M_0 *solves* $\mathcal{P}(\nu_0)$*. (ii) If* M_0 *is the unique solution of* $\mathcal{P}(\nu)$*, then there exists* $\tau_0 \in \mathbb{R}$ *such that* M_0 *solves* $\Pi(C_{\tau_0})$*.*

Proof. (i) Let $\nu_0 := \mathbb{E}[\phi(M_0)]$. By strict convexity of ϕ and [Borwein and Lewis, 1991, Prop. 2.11], M_0 is the unique solution of $\Pi(C_\tau)$. Therefore, M_0 is the unique element of $\{M \in \mathcal{M}_+ \ : \ \mathbb{E}[\phi(M)] \leq \nu_0, \ h(\mathbb{E}[M\boldsymbol{X} - \boldsymbol{P}]) \leq \tau\}$. Thus, M_0 solves $\mathcal{P}(\nu_0)$. (ii) Let $\tau_0 := h(\mathbb{E}[M_0\boldsymbol{X} - \boldsymbol{P}])$. If M_0 is the unique solution of $\mathcal{P}(\nu)$, it is the unique element of $\{M \in \mathcal{M}_+ \ : \ \mathbb{E}[\phi(M)] \leq \nu, \ h(\mathbb{E}[M\boldsymbol{X} - \boldsymbol{P}]) \leq \tau_0\}$. Thus, M_0 solves $\Pi(C_{\tau_0})$. □

Lemma OA-1. *Consider norms h_1 and h_∞ in equations (27) and (28) of the main text. Then, penalization $\sigma_{C_i}(\boldsymbol{\theta}) = \tau h_{i*,\boldsymbol{\Sigma}_{\boldsymbol{\zeta}}^{1/2}}(\boldsymbol{\theta})$ $(i = 1, \infty)$ from Corollary 3 in the main text is given in closed-from as follows.*

(i) If $\lambda = 1$: $h_{1*}(\boldsymbol{\theta}) = h_{\infty*}(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_2$.

(ii) If $\lambda = 0$: $h_{1*}(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_\infty$ and $h_{\infty*}(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_1 / \sqrt{N}$.

(iii) If $\lambda \in (0, 1)$:

$$h_{1*}(\boldsymbol{\theta}) = \min_{\boldsymbol{z} \in \mathbb{R}^N} \left\{ \max \left\{ \frac{\|\boldsymbol{z}\|_\infty}{1-\lambda}, \frac{\|\boldsymbol{\theta} - \boldsymbol{z}\|_2}{\lambda} \right\} \right\} ,$$

$$h_{\infty*}(\boldsymbol{\theta}) = \min_{\boldsymbol{z} \in \mathbb{R}^N} \left\{ \max \left\{ \frac{\|\boldsymbol{z}\|_1}{\sqrt{N}(1-\lambda)}, \frac{\|\boldsymbol{\theta} - \boldsymbol{z}\|_2}{\lambda} \right\} \right\} .$$

Proof of Lemma OA-1. Since the cases $\lambda = 0, 1$ are obvious, let $\lambda \in (0, 1)$. Denoting by $||\cdot||_A$, $||\cdot||_B$ two norms, following identity holds for any $\boldsymbol{\eta} \in \mathbb{R}^N$:

$$||\boldsymbol{\eta}|| := \lambda ||\boldsymbol{\eta}||_A + (1-\lambda)||\boldsymbol{\eta}||_B = \left|\left| \begin{pmatrix} \lambda ||\boldsymbol{\eta}||_A \\ (1-\lambda)||\boldsymbol{\eta}||_B \end{pmatrix} \right|\right|_1 .$$

From [Combettes et al., 2019, Thm. 2.5], we obtain for any $\boldsymbol{\theta} \in \mathbb{R}^N$:

$$\begin{aligned} ||\boldsymbol{\theta}||_* &= \inf_{(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \in \mathbb{R}^{2N}} \left\{ \left|\left| \begin{pmatrix} \frac{1}{\lambda}||\boldsymbol{\theta}_1||_{A*} \\ \frac{1}{1-\lambda}||\boldsymbol{\theta}_2||_{B*} \end{pmatrix} \right|\right|_{1*} : \boldsymbol{\theta}_1 + \boldsymbol{\theta}_2 = \boldsymbol{\theta} \right\} \\ &= \min_{\boldsymbol{z} \in \mathbb{R}^N} \left\{ \max \left\{ \frac{1}{\lambda}||\boldsymbol{z}||_{A*}, \frac{1}{1-\lambda}||\boldsymbol{\theta} - \boldsymbol{z}||_{B*} \right\} \right\} , \end{aligned}$$

since $||\cdot||_{1*} = ||\cdot||_\infty$. The statement of the lemma finally follows by further recalling that $||\cdot||_{2*} = ||\cdot||_2$ and $||\cdot||_{\infty*} = ||\cdot||_1$. □

B ONLINE APPENDIX – Theory of S–SDFs and Applications

B.1 Cressie-Read Dispersions

We collect for easy reference relevant explicit examples of $\phi-$dispersions from the Cressie-Read family, together with their corresponding functions ϕ_+ and ϕ_+^*; see also Kitamura et al. [2004] and Newey and Smith [2004], among others.

Examples 1. 1. Kullback-Leibler dispersion:

$$\phi_+(x) = \begin{cases} x\ln x - x + 1 & x > 0 \\ 1 & x = 0 \\ +\infty & x < 0 \end{cases}, \quad \text{with} \quad \phi_+^*(y) = \exp(y) - 1 \ .$$

2. Negative entropy:

$$\phi_+(x) = \begin{cases} -\ln x + x - 1 & x > 0 \\ +\infty & x \leq 0 \end{cases}, \quad \text{with} \quad \phi_+^*(y) = \begin{cases} -\ln(1-y), & y < 1 \\ +\infty & y \geq 1 \end{cases} .$$

3. Power dispersion:

   (a) for $\gamma > 1$ and $\beta = \gamma/(\gamma - 1)$,

$$\phi_+(x) = \begin{cases} x^\gamma/\gamma & x \geq 0 \\ +\infty & x < 0 \end{cases}, \quad \text{with} \quad \phi_+^*(y) = \max\{y, 0\}^\beta/\beta \ ,$$

(b) for $0 < \gamma < 1$ and $\beta = \gamma/(\gamma - 1)$,

$$\phi_+(x) = \begin{cases} \frac{\gamma x - x^\gamma}{1-\gamma} & x \geq 0 \\ +\infty & x < 0 \end{cases}, \quad \text{with} \quad \phi_+^*(y) = \begin{cases} (1 + y/\beta)^\beta - 1 & y < -\beta \\ +\infty & y \geq -\beta \end{cases},$$

(c) for $\gamma < 0$ and $\beta = \gamma/(\gamma - 1)$,

$$\phi_+(x) = \begin{cases} \frac{1 - x^\gamma + \gamma(x-1)}{\gamma} & x > 0 \\ +\infty & x \leq 0 \end{cases}, \quad \text{with} \quad \phi_+^*(y) = \begin{cases} -\frac{(1-y)^\beta + 1}{\beta} & y < 1 \\ +\infty & y \geq 1 \end{cases}.$$

In the above settings, function ϕ_+ is always closed and strictly convex, while function ϕ_+^* is always differentiable in the interior of its domain. Moreover, function ϕ_+^* is strictly convex in all cases, but case (3a).

The next lemma summarizes the analytical expressions for minimum dispersion S–SDFs corresponding to the particular choices of $\phi-$dispersions from Examples 1.

Lemma OA-2. *Under the conditions of Proposition 2 in the main text, consider the $\Phi-$dispersions induced by functions ϕ_+ in Examples 1. It then follows:*

(1) $M_0 = \exp(-\boldsymbol{X}'\boldsymbol{\theta}_0)$.

(2) $M_0 = 1/(1 + \boldsymbol{X}'\boldsymbol{\theta}_0)$, *if* $-\boldsymbol{X}'\boldsymbol{\theta}_0 < 1$.

(3a) $M_0 = \max\{-\boldsymbol{X}'\boldsymbol{\theta}_0, 0\}^{\beta-1}$.

(3b) $M_0 = (1 - \boldsymbol{X}'\boldsymbol{\theta}_0/\beta)^{\beta-1}$, *if* $-\boldsymbol{X}'\boldsymbol{\theta}_0 < -\beta$.

(3c) $M_0 = (1 + \boldsymbol{X}'\boldsymbol{\theta}_0)^{\beta-1}$, *if* $-\boldsymbol{X}'\boldsymbol{\theta}_0 < 1$.

Proof of Lemma *OA-2.* For every $\phi-$dispersion considered under Examples 1, ϕ_+ is strictly convex on its domain. Moreover, for $\phi-$dispersions (1) and (3a) the interior of the domain of ϕ_+^*, $\text{int}(\text{dom}\,\phi_+^*)$ is the entire real line, i.e., there is no restriction on the

values of the optimal portfolio payoff. With regard to the other cases, the optimal portfolio payoff is restricted to be almost surely in the interior of the domain of ϕ_+^*, which is $(-\infty, 1)$ in cases (2) and (3c) and $(-\infty, -\beta)$ in case (3b), respectively. □

B.2 S–SDFs induced by short selling constraints and bid-ask spreads

Minimum variance SDFs incorporating bid-ask spreads and short-sale constraints in arbitrage-free markets have been studied in Luttmer [1996]. In general, such market frictions can be modelled with a portfolio constraint $\boldsymbol{\theta} \in K$, using a finitely generated convex cone $K \subset \mathbb{R}^N$. Luttmer [1996] shows that SDFs in such markets imply pricing errors constrained to belong to the polar cone of K, which we denote by K^0.[6] For instance, short-sale constraints are defined by non-negative portfolio positions, i.e., $K = \mathbb{R}^N_+$. Since the polar cone of $\mathbb{R}^N_+$ is $\mathbb{R}^N_-$, it follows that $\mathbb{E}[M\boldsymbol{X}] \leq \boldsymbol{P}$. In our framework, these market frictions are easily modelled via the characteristic function of K:

$$\sigma_{K^0}(\boldsymbol{\theta}) = \delta_K(\boldsymbol{\theta}) := \begin{cases} 0 & \boldsymbol{\theta} \in K \\ +\infty & \boldsymbol{\theta} \notin K \end{cases}.$$

Literally, investors incur in zero transaction costs if portfolio positions belong to K and arbitrarily large transaction costs otherwise. The S–SDF valuations obtained in Luttmer [1996] are constrained to belong to set $\boldsymbol{P} + K^0$.

B.3 Theories of Good-Deal-Bounds and robust identification of investor beliefs

Consider first the Good-Deal Bounds theory in Cochrane and Saa-Requejo [2000], which specifies a maximal price $\overline{P}_D$ and a minimal price $\underline{P}_D$ for a newly issued security with payoff X_D, given an exact pricing condition for a set of traded securities and under an upper bound on the maximum attainable Sharpe Ratio. This bound is introduced to rule

[6] The polar cone of set K is the set $K^0 := \{\boldsymbol{y} \in \mathbb{R}^N : \boldsymbol{x}'\boldsymbol{y} \leq 0, \text{ for all } \boldsymbol{x} \in K\}$.

out unreasonably good deals in the extended market including the new security. Let M_S be the minimum dispersion SDF that minimizes the SDF dispersion for a given convex dispersion function ϕ and which prices exactly the set of sure securities.[7] Therefore, M_S solves the optimization problem:

$$\inf\{\mathbb{E}[\phi(M)] \ : \ M \in \mathcal{M}_+\} \ .$$

Let $\nu_S := \mathbb{E}[\phi(M_S)]$ and denote by $P_D := \mathbb{E}[M_S X_D]$ the reference price for the newly issued payoff under SDF M_S. The minimal and maximal price for the newly issued security are obtained for an upper dispersion threshold $\nu \geq \nu_S$ as follows:

$$\underline{P}_D := P_D + \inf_{M \in \mathcal{M}_+} \{\mathbb{E}[MX_D] - P_D \ : \ \mathbb{E}[\phi(M)] \leq \nu\} \ , \tag{OA-5}$$

$$\overline{P}_D := P_D - \inf_{M \in \mathcal{M}_+} \{-(\mathbb{E}[MX_D] - P_D) \ : \ \mathbb{E}[\phi(M)] \leq \nu\} \ . \tag{OA-6}$$

By construction, these prices satisfy the inequalities $\underline{P}_D \leq P_D \leq \overline{P}_D$. Moreover, they are supported by corresponding S–SDFs $\underline{M}_0$ and $\overline{M}_0$, which solve each one of the minimizations on the RHS of identities (OA-5)-(OA-6):

$$\underline{M}_0 = \arg \min_{M \in \mathcal{M}_+} \{\underline{h}(\mathbb{E}[MX_D] - P_D) \ : \ \mathbb{E}[\phi(M)] \leq \nu\} \ , \tag{OA-7}$$

$$\overline{M}_0 := \arg \min_{M \in \mathcal{M}_+} \left\{\overline{h}(\mathbb{E}[MX_D] - P_D)) \ : \ \mathbb{E}[\phi(M)] \leq \nu\right\} \ , \tag{OA-8}$$

with corresponding pricing metrics given by $\underline{h}(\boldsymbol{\eta}) = \boldsymbol{\eta}$ and $\overline{h} = -\boldsymbol{\eta}$. It thus follows from Proposition OA-1 of Appendix A that $\underline{P}_D$ and $\overline{P}_D$ can be equivalently obtained by computing the price of the new security's payoff X_D under these two minimum dispersion S–SDFs.

Consider now Chen et al. [2020] theory for a robust identification of investor beliefs.

[7] In Cochrane and Saa-Requejo [2000] Good-Deal Bounds theory the notion of dispersion minimized is the SDF variance.

This theory relies on the determination of a S–SDF $\underline{M}_0$ solving problem (OA-7) under robust notions of S–SDF dispersion (cases 1 and 3a of Section B.1) and while imposing the normalization constraint $\mathbb{E}[M] = 1$, which can be naturally incorporated in constrained set $\mathcal{M}_+$ as a sure pricing constraint. Such normalization gives rise to the interpretation of $\underline{M}_0$ as a density reproducing a recovered robust investor belief with respect to the reference probability used to describe risky asset payoffs. Such interpretation if further motivated by the fact that in this theory "sure payoffs" $\boldsymbol{X}_S$ are not asset payoffs, but instead asset payoffs that have been already stochastically discounted by a parametric SDF $M^\star$ implied by economic theory. Therefore, in this setting $\mathbb{E}[\underline{M}_0 X_D]$ has the interpretation of a conservative expected payoff under the recovered investor belief, instead of a lowest arbitrage-free price for payoff X_D.

B.4 Existence of S–SDFs for markets with convex costs

We have shown in Section 2.2 of the main text that S–SDFs are characterized by the absence of free lunches in an economy with sublinear costs. In this section, we show that S–SDFs arise more generally also in an economy with convex costs, under a no-arbitrage condition stronger than the absence of free lunches, i.e., market viability, as defined in, e.g., Harrison and Kreps [1979].[8] To this end, let portfolio positions involve costs measured by closed convex function $\sigma : \mathbb{R}^N \to [0, \infty]$. As in the main text, we define the set of traded payoffs as the set of portfolio payoffs involving finite costs:

$$\mathcal{Z} := \{Z = \boldsymbol{X}'\boldsymbol{\theta} \ : \ \sigma(\boldsymbol{\theta}) < +\infty\} \ .$$

Accordingly, we define a pricing functional π on $\mathcal{Z}$ by:

$$\pi(Z) := \inf_{\boldsymbol{\theta} \in \mathbb{R}^N} \{\boldsymbol{P}'\boldsymbol{\theta} + \sigma(\boldsymbol{\theta}) \ : \ Z = \boldsymbol{X}'\boldsymbol{\theta}\} \ .$$

[8] A market $(\mathcal{Z}, \pi, \mathbb{P})$ is viable if there exists an agent (represented by preference $\gtrsim$) and $Z^* \in \mathcal{Z}$ such that $\pi(Z^*) \leq 0$ and $Z^* \gtrsim Z$ for all $Z \in \mathcal{Z}$ with $\pi(Z) \leq 0$. The preference relation $\gtrsim$ is assumed to be convex, continuous and strictly increasing.

Since cost function σ is closed and convex, π is a convex pricing functional and $\mathcal{Z}$ is a closed and convex set of traded payoffs. With these definitions, the next proposition provides the closed-form supporting market in which the existence of a S–SDF is characterized by market viability.

Proposition OA-2. *Market $(\mathcal{Z}, \pi, \mathbb{P})$ is viable if and only if there exists an almost surely strictly positive SDF $M \in L^q(\mathbb{P})$, such that*

$$\mathbb{E}[M\boldsymbol{X}] - \boldsymbol{P} \in C_\sigma \ ,$$

where

$$C_\sigma := \{\boldsymbol{\eta} \in \mathbb{R}^N \ : \ \boldsymbol{\eta}'\boldsymbol{\theta} \leq \sigma(\boldsymbol{\theta}) \ \ \textit{for all } \boldsymbol{\theta} \in \mathbb{R}^N\} \ .$$

Proof of Proposition OA-2. Convexity of σ implies convexity of pricing functional π and the set of traded payoffs $\mathcal{Z}$. By [Chen, 2001, Thm. 1 and Cor. 1], market viability is equivalent to the existence of an a.s. strictly positive SDF, M, such that $\mathbb{E}[MZ] \leq \pi(Z)$ for any $Z \in \mathcal{Z}$. Now, by definition of pricing functional π and traded payoff space $\mathcal{Z}$, it equivalently follows for any $\boldsymbol{\theta} \in \mathbb{R}^N$ that $\boldsymbol{\theta}'\boldsymbol{P} + \sigma(\boldsymbol{\theta}) \geq \mathbb{E}[M\boldsymbol{\theta}'\boldsymbol{X}]$, i.e.,

$$\boldsymbol{\theta}' \left(\mathbb{E}[M\boldsymbol{X}] - \boldsymbol{P}\right) \leq \sigma(\boldsymbol{\theta}) \ . \tag{OA-9}$$

Equivalently, by definition of C_σ, we have $\mathbb{E}[M\boldsymbol{X}] - \boldsymbol{P} \in C_\sigma$.

□

Given a closed convex set $C = \{\boldsymbol{\eta} \in \mathbb{R}^N \ : \ h(\boldsymbol{\eta}) \leq \tau\}$ defining a S–SDF with closed convex function h and threshold $\tau \geq 0$, the convex arbitrage-free economy supporting such constraint set C is based on a cost function given by:

$$\sigma = h^* + \tau \ .$$

While in Proposition 1 of the main text the necessary and sufficient condition for the existence of a strictly positive S–SDF in an economy with sublinear cost function σ_C is the absence of free-lunches, when h^* is not sublinear the no-arbitrage condition equivalent to the existence of a strictly positive S–SDF is market viability.

C ONLINE APPENDIX – Empirics of APT S–SDFs

C.1 S–SDF tradeoff between pricing accuracy and time series explanatory power

Figure OA-4 in Section D of this Online Appendix documents the effect of the choice of the notion of S–SDF dispersion for the S–SDF tradeoff between time series and cross-sectional explanatory powers. The two left panels of Figure OA-4 show that the resulting tradeoff of S–SDFs minimizing notions of dispersion such as Kullback-Leibler or Hellinger divergence is slightly Pareto dominated by minimum variance S–SDFs on non perturbed data.[9] Moreover, minimum variance S–SDFs produce a clearly more consistent tradeoff on perturbed data, both for the low- and intermediate dimensional datasets.

C.2 Basic out-of-sample framework and evidence

For every in-sample window of 30 years of monthly observations with last observation in month y, the in-sample S–SDF estimation requires the specification of a corresponding threshold, denoted by $\hat{\tau}_y$.[10] Similar to the in-sample analysis, our theoretical findings offer a natural way to determine the range of relevant empirical thresholds. We obtain the maximal admissible threshold $\hat{\tau}_y^{max}$ as the empirical version of equation (26) in the main

[9] Kullback-Leibler and Hellinger divergence are members of the family of Cressie-Read notions of stochastic dispersion detailed in Section B.1 of this Online Appendix (see case 1 and case 3b for $\gamma = 1/2$, respectively).

[10] Given our data sample, this gives us an out-of-sample period of monthly observations from July 1963 to June 2018. The first in-sample window of 30 years consists of monthly observations from July 1933 to June 1963.

text. For instance, under a set of traded risk factors including the market excess return alone, this corresponds to the empirical threshold implied by a minimum variance SDF that is only required to price exactly the market excess return in the given estimation window. The minimal admissible threshold $\hat{\tau}_y^{min}$ is obtained as in Section 5.3 in the main text, as the smallest threshold value for which in the given estimation window no empirical duality failure arises. This yields for any estimation window a range $[\hat{\tau}_y^{min}, \hat{\tau}_y^{max}]$ of admissible thresholds $\hat{\tau}_y$ and a range $[\hat{\kappa}_y^{min}, 1]$ of admissible relative thresholds $\hat{\kappa}_y$, where $\hat{\kappa}_y^{min} := \hat{\tau}_y^{min}/\hat{\tau}_y^{max}$. Given a sequence of admissible relative thresholds $\{\hat{\kappa}_y\}$, we apply the estimation methodology in Section 5.2 of the main text, to obtain a sequence $\{\hat{\boldsymbol{\theta}}_y\}$ of estimated S–SDF dual portfolio weights, which is updated at a semi-annual frequency. For each of the six months following month y, we denote by $\boldsymbol{X}_{y+m}$ the vector of excess returns in month $y+m$ and estimate with Corollary 3 the corresponding sequence of out-of-sample monthly minimum variance S–SDFs, i.e., $\hat{M}_{y+m} := \max\{-\hat{\boldsymbol{\theta}}_y' \boldsymbol{X}_{y+m}, 0\}$. In this way, we obtain a monthly time series $\{\hat{M}_{y+m}\}$ of data driven APT S–SDFs, for which we can evaluate the out-of-sample time series and cross-sectional explanatory power for the sequence of out-of-sample return vectors $\{\boldsymbol{X}_{y+m}\}$.

Regarding the choice of the sequence of relative thresholds $\{\hat{\kappa}_y\}$, we adopt in this section a simple approach based on constant relative thresholds $\hat{\kappa}_y = \overline{\kappa}$. This approach produces a useful description of the attainable out-of-sample tradeoffs between time series and cross-sectional S–SDF explanatory power, which may arise from data-driven threshold selections holding the relative threshold roughly constant across estimation windows, but it does not produce a criterion for a non forward-looking choice of an optimal threshold. The first two rows of Figure OA-5 report the resulting curves of out-of-sample average time series R^2 (31) and cross-sectional GLS R^2 (32), in dependence of the value of constant relative threshold $\overline{\kappa} \in [\max_y\{\hat{\kappa}_y^{min}\}, 1]$, for minimum variance S–SDFs incorporating the market excess return as the single traded risk factor. All out-of-sample GLS R^2 curves are inversely U–shaped. In the low (intermediate) dimensional data set, the maximal GLR R^2 is as high

as about 70% (45%) for the APT S–SDF with l_2-pricing constraints, when the relative threshold $\overline{\kappa}$ is about 40% (65%). These are large increases in pricing accuracy, compared, e.g., to the mildly negative out-of-sample GLR R^2s of an empirical CAPM S–SDF exactly pricing only the market return in-sample ($\overline{\kappa} = 1$).

Importantly, we observe a tradeoff between time series and cross-sectional S–SDF out-of-sample explanatory power, which is summarized in the two bottom panels of in Figure OA-5, for relative thresholds between the threshold value at which the maximum of the GLS R^2 curve is attained and the maximum admissible value $\overline{\kappa} = 1$. Similar to the in-sample findings of Section 5.4.2 in the main text, APT S–SDFs with pricing metric different from the l_2-norm imply a Pareto dominated out-of-sample tradeoff between cross-sectional and time series explanatory power, further confirming that S–SDF sparsity is quite costly also in an out-of-sample perspective.

D ONLINE APPENDIX – Additional Figures

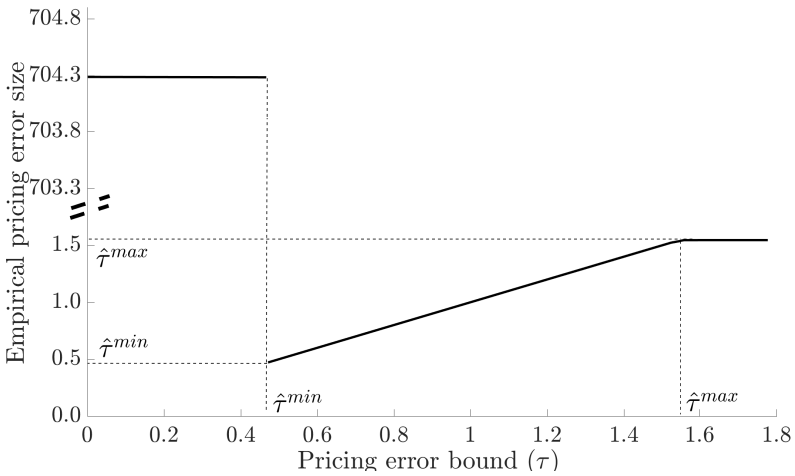
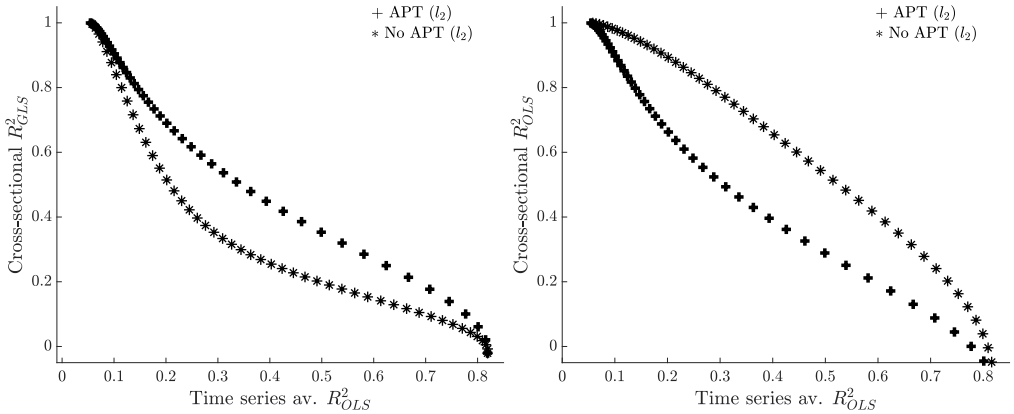
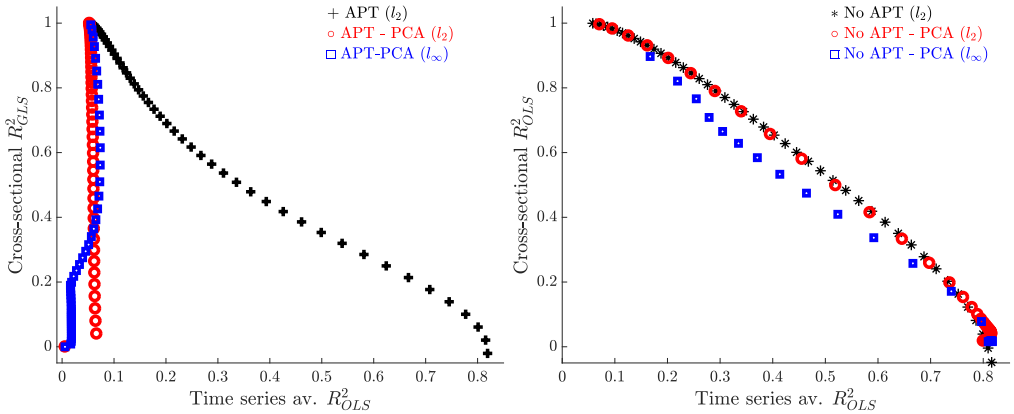


Figure OA-1: **Empirical duality failure.** We compute the minimum variance APT S–SDF M_0 with l_2-pricing function for varying thresholds $\tau \geq 0$. On the y-axis we report the estimate of the function value $h(\mathbb{E}[M_0\boldsymbol{X} - \boldsymbol{P}])$ for each τ. The point of discontinuity in the plot identifies the smallest threshold τ, for which a solution of the empirical primal S–SDF problem exists. The largest threshold $\hat{\tau}^{max}$ in the plot is computed as the sample version of the maximal threshold τ^{max} in equation (26) of the main text. Sorted portfolio returns are used to construct the S–SDFs without assuming any observable traded factor. All calculations are based on the intermediate dimensional dataset from June 1990 to June 2018.

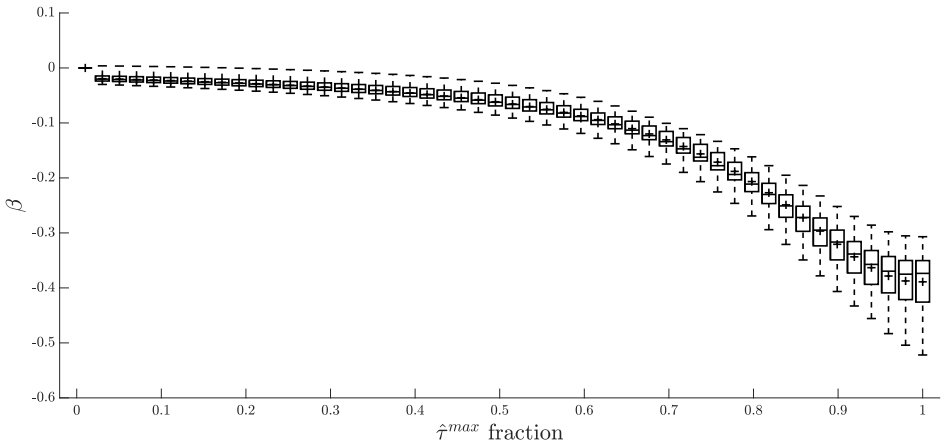


(1) **Panel (A)**

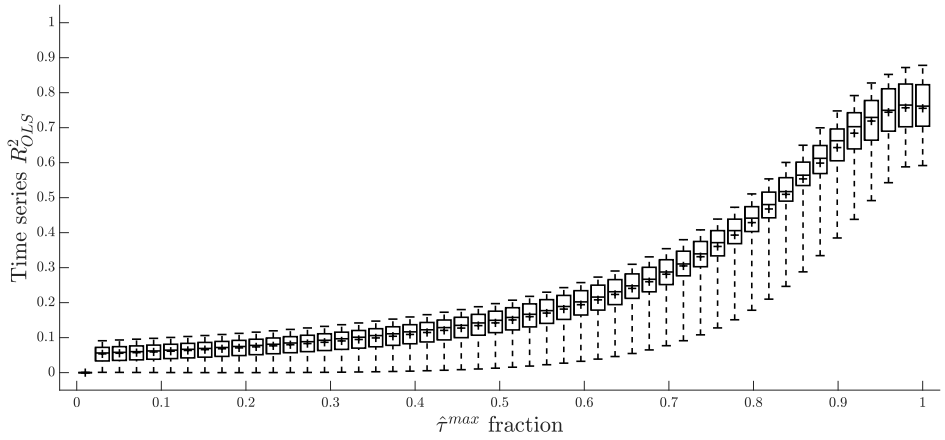


(2) **Panel (B)**

Figure OA-2: **Tradeoff between time series and cross-sectional S–SDF explanatory power of APT S–SDFs, non APT S–SDFs and principal components based S–SDFs.** For various minimum variance S–SDFs, the figure reports the tradeoff between average time series R^2 metric (32) and cross-sectional GLS R^2 metric (31) (denoted by R^2_{GLS}, left panels) in the main text, or cross-sectional R^2 metric (34) (denoted by R^2_{OLS}, right panels) in the main text. The following S–SDFs are compared in Panel (A). First, APT S–SDFs based on the l_2-metric. Second, non APT S–SDFs based on the l_2-metric, but with a diagonal weighting matrix $\mathbf{\Sigma_\eta} = \boldsymbol{I}_{N_D \times N_D}$. Both these S–SDFs take the first principal component of excess returns as the single traded factor. In each figure of Panel (B), results for two additional principal components based S–SDFs are reported, which impose the corresponding pricing constraints on the principal components of sorted portfolio excess returns, instead of the original excess returns. The first and second of these S-SDFs impose in the left (right) figure an APT l_2- and $l_\infty-$ pricing constraint (a non APT l_2- and $l_\infty-$ pricing constraint based on a diagonal weighting matrix $\mathbf{\Sigma_\eta} = \boldsymbol{I}_{N_D \times N_D}$), respectively. All results are based on the low dimensional dataset running from 1963 to 2018.

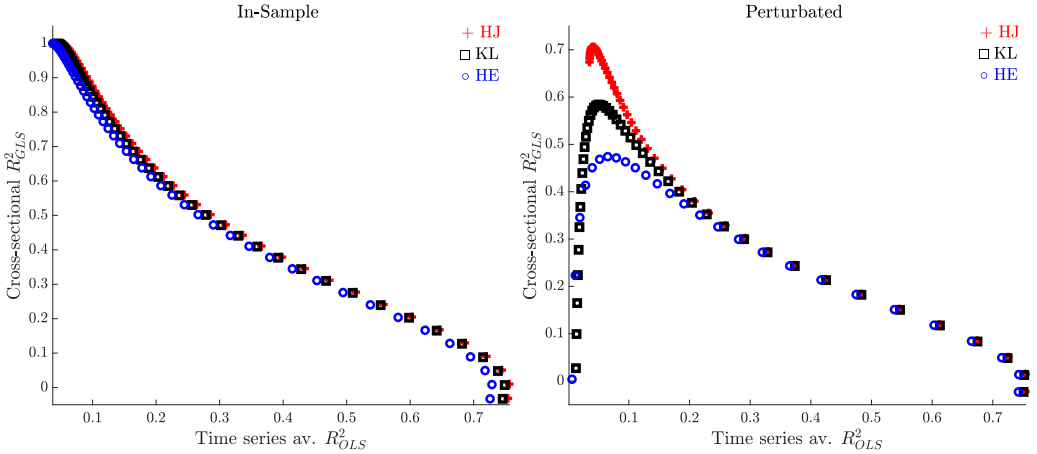


(1) **Panel (A)**

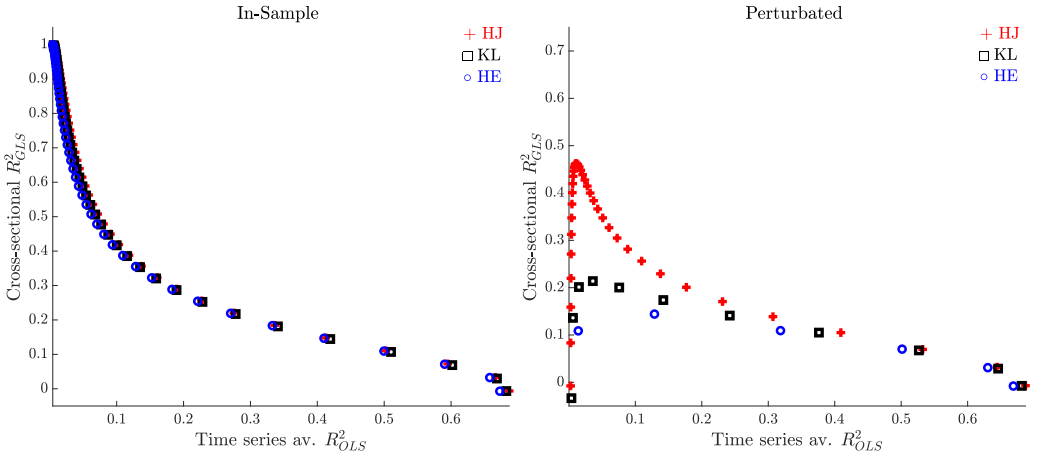


(2) **Panel (B)**

Figure OA-3: **Cross-sectional distributions of APT S–SDF betas and time series R^2 metrics.** The figure reports in **Panel (A)** (**Panel (B)**) the cross-sectional distribution of S–SDF betas (time series R^2 metrics (32) in the main text) as a function of threshold τ, for minimum variance S–SDFs satisfying APT pricing constraint (16) of the main text under an l_2-metric, and while taking the market excess return as the single traded factor. In both panels, plus signs (horizontal lines) in the boxes indicate cross-sectional averages (medians). All computations are based on the low dimensional dataset from 1963 to 2018.

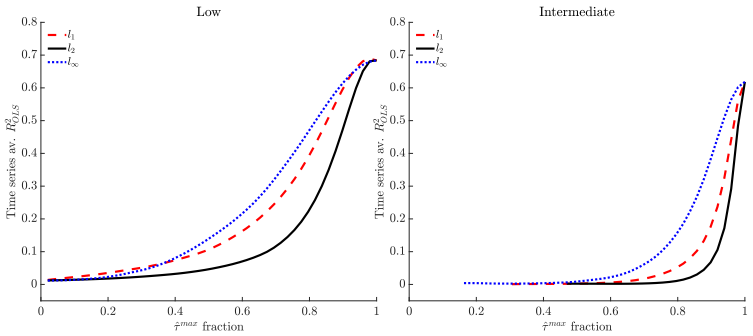


(1) **Panel (A)**

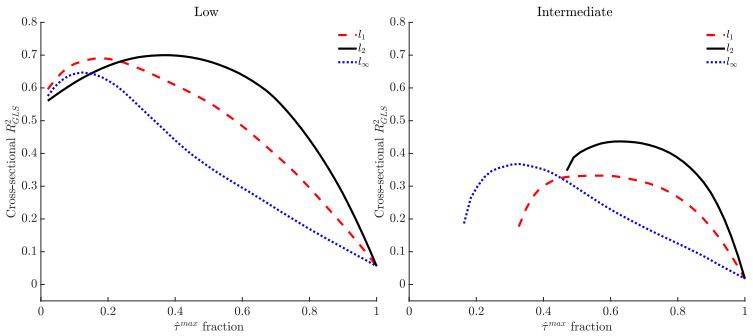


(2) **Panel (B)**

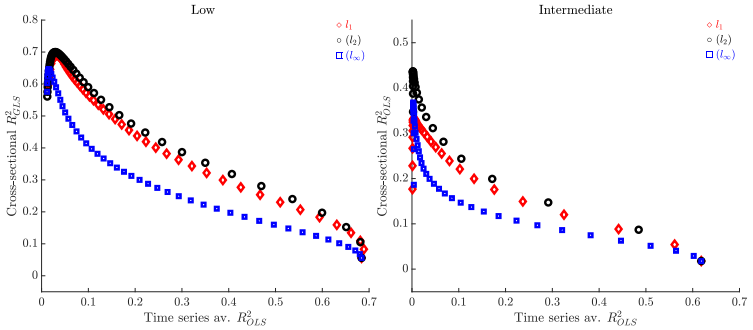
Figure OA-4: **Tradeoff between time series and cross-sectional APT S–SDF explanatory power, under different notions of S–SDF dispersion and data perturbations.** The figure reports the tradeoff between average time series R^2 metric (32) and cross-sectional GLS R^2 metric (31) in the main text for minimum dispersion APT S–SDFs based on the l_2-metric and on different notions of S–SDF dispersion from the Cressie and Read [1984] family in Section B.1 of this Online Appendix: Variance (label HJ, case 3(a)), Kullback-Leibler divergence (label KL, case 1) and Hellinger divergence (label HE, case 3(b)). **Panel (A)** reports results for the low dimensional dataset and **Panel (B)** for the intermediate dimensional dataset. Left panels report in-sample results, where the estimation and evaluation periods overlap on the 40 (50) year window 1965-2005 (1955-2005) for the low (intermediate) dimensional dataset. For the low (intermediate) dimensional dataset, right panels report results for a 40 (50) year estimation window 1965-2005 (1955-2015) and a translated 40 (50) year evaluation window 1975-2015 (1965-2015) of 10 years.



(1) **Panel (A)**



(2) **Panel (B)**



(3) **Panel (C)**

Figure OA-5: **Out-of-sample tradeoff between time-series and cross-sectional APT S–SDF explanatory power.** The figure reports the out-of-sample average time-series R^2 metric (32) (top panels) and the out-of-sample cross-sectional GLS R^2 metric (31) (middle panels) from the main text, for the sequence of minimum variance APT S–SDFs from Section C.2. These S–SDFs are estimated on rolling windows of 30 years under a constant relative threshold $\overline{\kappa} \in [\max_y\{\hat{\kappa}_y^{min}\}, 1]$ and based on a l_1-, l_2-, and (scaled) $l_\infty-$metric, respectively. In each estimation window, S–SDFs correctly price the risk-free asset and the market return. Bottom panels directly report the tradeoff between out-of-sample average time series R^2 metric (32) and out-of-sample cross-sectional GLS R^2 metric (31). Results are based on the low dimensional (left panels) and the intermediate dimensional (right panels) datasets.

References

Heinz H Bauschke and Patrick L Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*, volume 408. Springer, 2011.

Heinz H Bauschke, Patrick L Combettes, et al. *Convex analysis and monotone operator theory in Hilbert spaces*. Springer, 2017.

Jonathan M Borwein. A lagrange multiplier theorem and a sandwich theorem for convex relations. *Mathematica Scandinavica*, 48:189–204, 1981.

Jonathan M Borwein and Adrian S Lewis. Duality relationships for entropy-like minimization problems. *SIAM Journal on Control and Optimization*, 29(2):325–338, 1991.

Lawrence D Brown and R Purves. Measurable selections of extrema. *The annals of statistics*, 1(5):902–912, 1973.

Xiaohong Chen, Lars P Hansen, and Peter G Hansen. Robust identification of investor beliefs. Technical report, 2020.

Zhiwu Chen. Viable costs and equilibrium prices in frictional securities markets. *Annals of Economics and Finance*, 2(2):297–323, 2001.

Stephen A Clark. The valuation problem in arbitrage price theory. *Journal of Mathematical Economics*, 22(5):463–478, 1993.

John H Cochrane and Jesus Saa-Requejo. Beyond arbitrage: Good-deal asset price bounds in incomplete markets. *Journal of Political Economy*, 108(1):79–119, 2000.

Patrick L Combettes, Andrew M McDonald, Charles A Micchelli, and Massimiliano Pontil. Learning with optimal interpolation norms. *Numerical Algorithms*, 81(2):695–717, 2019.

Noel Cressie and Timothy RC Read. Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 46(3):440–464, 1984.

James Davidson. *Stochastic limit theory: An introduction for econometricians*. OUP Oxford, 1994.

J Michael Harrison and David M Kreps. Martingales and arbitrage in multiperiod securities markets. *Journal of Economic theory*, 20(3):381–408, 1979.

Yuichi Kitamura, Gautam Tripathi, and Hyungtaik Ahn. Empirical likelihood-based inference in conditional moment restriction models. *Econometrica*, 72(6):1667–1714, 2004.

Erzo GJ Luttmer. Asset pricing in economies with frictions. *Econometrica*, 64(6):1439–1467, 1996.

Whitney K Newey and Daniel McFadden. Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4:2111–2245, 1994.

Whitney K Newey and Richard J Smith. Higher order properties of gmm and generalized empirical likelihood estimators. *Econometrica*, 72(1):219–255, 2004.

Ralph T Rockafellar. Integrals which are convex functionals. *Pacific Journal of Mathematics*, 24(3):525–539, 1968.

Ralph T Rockafellar. *Convex analysis*. Princeton University Press, 1970.