European Journal of Operational Research 270 (2018) 931–942

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Characterization of the equivalence of robustification and regularization in linear and matrix regression☆

Dimitris Bertsimas[a,*], Martin S. Copenhaver[b]

[a] Sloan School of Management and Operations Research Center, MIT, United States
[b] Operations Research Center, MIT, United States

ARTICLE INFO

Article history:
Received 9 November 2016
Accepted 20 March 2017
Available online 28 March 2017

Keywords:
Convex programming
Robust optimization
Statistical regression
Penalty methods
Adversarial learning

ABSTRACT

The notion of developing statistical methods in machine learning which are robust to adversarial perturbations in the underlying data has been the subject of increasing interest in recent years. A common feature of this work is that the adversarial *robustification* often corresponds exactly to regularization methods which appear as a loss function plus a penalty. In this paper we deepen and extend the understanding of the connection between robustification and regularization (as achieved by penalization) in regression problems. Specifically,

(a) In the context of linear regression, we characterize precisely under which conditions on the model of uncertainty used and on the loss function penalties robustification and regularization are equivalent.
(b) We extend the characterization of robustification and regularization to matrix regression problems (matrix completion and Principal Component Analysis).

1. Introduction

The development of predictive methods that perform well in the face of uncertainty is at the core of modern machine learning and statistical practice. Indeed, the notion of *regularization*— loosely speaking, a means of controlling the ability of a statistical model to generalize to new settings by trading off with the model's complexity— is at the very heart of such work (Hastie, Tibshirani, & Friedman, 2009). Corresponding regularized statistical methods, such as the Lasso for linear regression (Tibshirani, 1996) and nuclear-norm-based approaches to matrix completion (Candès & Recht, 2012; Recht, Fazel, & Parrilo, 2010), are now ubiquitous and have seen widespread success in practice.

In parallel to the development of such regularization methods, it has been shown in the field of robust optimization that under certain conditions these regularized problems result from the need to immunize the statistical problem against adversarial perturbations in the data (Ben-Tal, Ghaoui, & Nemirovski, 2009; Caramanis, Mannor, & Xu, 2011; Ghaoui & Lebret, 1997; Xu, Caramanis, & Mannor, 2010). Such a *robustification* offers a different perspective on regularization methods by identifying which adversarial perturbations the model is protected against. Conversely, this can help to inform statistical modeling decisions by identifying potential choices of regularizers. Further, this connection between regularization and robustification offers the potential to use sophisticated data-driven methods in robust optimization (Bertsimas, Gupta, & Kallus, 2013; Tulabandhula & Rudin, 2014) to design regularizers in a principled fashion.

With the continuing growth of the adversarial viewpoint in machine learning (e.g. the advent of new deep learning methodologies such as generative adversarial networks (Goodfellow et al., 2014a; Goodfellow, Shlens, & Szegedy, 2014b; Shaham, Yamada, & Negahban, 2015)), it is becoming increasingly important to better understand the connection between robustification and regularization. Our goal in this paper is to shed new light on this relationship by focusing in particular on linear and matrix regression problems. Specifically, our contributions include:

1. In the context of linear regression we demonstrate that in general such a robustification procedure is not equivalent to regularization (via penalization). We characterize precisely under which conditions on the model of uncertainty used and on the loss function penalties one has that robustification is equivalent to regularization.
2. We break new ground by considering problems in the matrix setting, such as matrix completion and Principal Component Analysis (PCA). We show that the nuclear norm, a

☆ Copenhaver is partially supported by the Department of Defense, Office of Naval Research, through the National Defense Science and Engineering Graduate Fellowship.

* Corresponding author.
E-mail addresses: dbertsim@mit.edu (D. Bertsimas), mcopen@mit.edu (M.S. Copenhaver).

http://dx.doi.org/10.1016/j.ejor.2017.03.051
0377-2217/

Table 1
Matrix norms on $\mathbf{\Delta} \in \mathbb{R}^{m\times n}$.

Name	Notation	Definition	Description
p-Frobenius	F_p	$\left(\sum_{ij} \lvert\Delta_{ij}\rvert^p\right)^{1/p}$	Entrywise ℓ_p norm
p-spectral (Schatten)	σ_p	$\lVert\boldsymbol{\mu}(\mathbf{\Delta})\rVert_p$	ℓ_p norm on the singular values
Induced	(h, g)	$\max_{\boldsymbol{\beta}} \frac{g(\mathbf{\Delta}\boldsymbol{\beta})}{h(\boldsymbol{\beta})}$	Induced by norms g, h

popular penalty function used throughout this setting, arises directly through robustification. As with the case of vector regression, we characterize under which conditions on the model of uncertainty there is equivalence of robustification and regularization in the matrix setting.

The structure of the paper is as follows. In Section 2, we review background on norms and consider robustification and regularization in the context of linear regression, focusing both on their equivalence and non-equivalence. In Section 3, we turn our attention to regression with underlying matrix variables, considering in depth both matrix completion and PCA. In Section 4, we include some concluding remarks.

2. A robust perspective of linear regression

2.1. Norms and their duals

In this section, we introduce the necessary background on norms which we will use to address the equivalence of robustification and regularization in the context of linear regression. Given a vector space $V \subseteq \mathbb{R}^n$ we say that $\lVert\cdot\rVert : V \to \mathbb{R}$ is a *norm* if for all $\mathbf{v}, \mathbf{w} \in V$ and $\alpha \in \mathbb{R}$

1. If $\lVert\mathbf{v}\rVert = 0$, then $\mathbf{v} = 0$,
2. $\lVert\alpha\mathbf{v}\rVert = \lvert\alpha\rvert\lVert\mathbf{v}\rVert$ (absolute homogeneity), and
3. $\lVert\mathbf{v}+\mathbf{w}\rVert \leq \lVert\mathbf{v}\rVert + \lVert\mathbf{w}\rVert$ (triangle inequality).

If $\lVert\cdot\rVert$ satisfies conditions 2 and 3, but not 1, we call it a *seminorm*. For a norm $\lVert\cdot\rVert$ on $\mathbb{R}^n$ we define its dual, denoted $\lVert\cdot\rVert_*$, to be

$$\lVert\boldsymbol{\beta}\rVert_* := \max_{\mathbf{x}\in\mathbb{R}^n} \frac{\mathbf{x}'\boldsymbol{\beta}}{\lVert\mathbf{x}\rVert},$$

where $\mathbf{x}'$ denotes the transpose of $\mathbf{x}$ (and therefore $\mathbf{x}'\boldsymbol{\beta}$ is the usual inner product). For example, the ℓ_p norms $\lVert\boldsymbol{\beta}\rVert_p := (\Sigma_i|\beta_i|^p)^{1/p}$ for $p \in [1, \infty)$ and $\lVert\boldsymbol{\beta}\rVert_\infty := \max_i|\beta_i|$ satisfy a well-known duality relation: ℓ_{p^*} is dual to ℓ_p, where $p^* \in [1, \infty]$ with $1/p + 1/p^* = 1$. We call p^* the *conjugate* of p. More generally for matrix norms[1] $\lVert\cdot\rVert$ on $\mathbb{R}^{m\times n}$ the dual is defined analogously:

$$\lVert\mathbf{\Delta}\rVert_* := \max_{\mathbf{A}\in\mathbb{R}^{m\times n}} \frac{\langle\mathbf{A}, \mathbf{\Delta}\rangle}{\lVert\mathbf{A}\rVert},$$

where $\mathbf{\Delta} \in \mathbb{R}^{m\times n}$ and $\langle\cdot,\cdot\rangle$ denotes the trace inner product: $\langle\mathbf{A}, \mathbf{\Delta}\rangle = \mathrm{Tr}(\mathbf{A}'\mathbf{\Delta})$, where $\mathbf{A}'$ denotes the transpose of $\mathbf{A}$. We note that the dual of the dual norm is the original norm (Boyd & Vandenberghe, 2004).

Three widely used choices for matrix norms (see Horn & Johnson, 2013) are Frobenius, spectral, and induced norms. The definitions for these norms are given below for $\mathbf{\Delta} \in \mathbb{R}^{m\times n}$ and summarized in Table 1 for convenient reference.

[1] We treat a matrix norm as any norm on $\mathbb{R}^{m\times n}$ which satisfies the three conditions of a usual vector norm, although some authors reserve the term "matrix norm" for a norm on $\mathbb{R}^{m\times n}$ which also satisfies a submultiplicativity condition (see Horn and Johnson, 2013, pg. 341).

1. The p-Frobenius norm, denoted $\lVert\cdot\rVert_{F_p}$, is the entrywise ℓ_p norm on the entries of $\mathbf{\Delta}$:

$$\lVert\mathbf{\Delta}\rVert_{F_p} := \left(\sum_{ij} \lvert\Delta_{ij}\rvert^p\right)^{1/p}.$$

Analogous to before, F_{p^*} is dual to F_p, where $1/p + 1/p^* = 1$.

2. The p-spectral (Schatten) norm, denoted $\lVert\cdot\rVert_{\sigma_p}$, is the ℓ_p norm on the singular values of the matrix $\mathbf{\Delta}$:

$$\lVert\mathbf{\Delta}\rVert_{\sigma_p} := \lVert\boldsymbol{\mu}(\mathbf{\Delta})\rVert_p,$$

where $\boldsymbol{\mu}(\mathbf{\Delta})$ denotes the vector containing the singular values of $\mathbf{\Delta}$. Again, σ_{p^*} is dual to σ_p.

3. Finally we consider the class of induced norms. If $g : \mathbb{R}^m \to \mathbb{R}$ and $h : \mathbb{R}^n \to \mathbb{R}$ are norms, then we define the induced norm $\lVert\cdot\rVert_{(h,g)}$ as

$$\lVert\mathbf{\Delta}\rVert_{(h,g)} := \max_{\boldsymbol{\beta}\in\mathbb{R}^n} \frac{g(\mathbf{\Delta}\boldsymbol{\beta})}{h(\boldsymbol{\beta})}.$$

An important special case occurs when $g = \ell_p$ and $h = \ell_q$. When such norms are used, (q, p) is used as shorthand to denote (ℓ_q, ℓ_p). Induced norms are sometimes referred to as operator norms. We reserve the term operator norm for the induced norm $(\ell_2, \ell_2) = (2, 2) = \sigma_\infty$, which measures the largest singular value.

2.2. Uncertain regression

We now turn our attention to uncertain linear regression problems and regularization. The starting point for our discussion is the standard problem

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^n} g(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}),$$

where $\mathbf{y} \in \mathbb{R}^m$ and $\mathbf{X} \in \mathbb{R}^{m\times n}$ are data and g is some convex function, typically a norm. For example, $g = \ell_2$ is least squares, while $g = \ell_1$ is known as least absolute deviation (LAD). In favor of models which mitigate the effects of overfitting these are often replaced by the *regularization* problem

$$\min_{\boldsymbol{\beta}} g(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + h(\boldsymbol{\beta}),$$

where $h : \mathbb{R}^n \to \mathbb{R}$ is some penalty function, typically taken to be convex. This approach often aims to address overfitting by penalizing the complexity of the model, measured as $h(\boldsymbol{\beta})$. (For a more formal treatment using Hilbert space theory, (see Bauschke & Combettes, 2011; Bousquet, Boucheron, & Lugosi, 2004). For example, taking $g = \ell_2^2$ and $h = \ell_2^2$, we recover the so-called regularized least squares (RLS), also known as ridge regression (Hastie et al., 2009). The choice of $g = \ell_2^2$ and $h = \ell_1$ leads to Lasso, or least absolute shrinkage and selection operator, introduced in Tibshirani (1996). Lasso is often employed in scenarios where the solution $\boldsymbol{\beta}$ is desired to be sparse, i.e., $\boldsymbol{\beta}$ has very few nonzero entries. Broadly speaking, regularization can take much more general forms; for our purposes, we restrict our attention to regularization that appears in the penalized form above.

In contrast to this approach, one may alternatively wish to reexamine the nominal regression problem $\min_{\boldsymbol{\beta}} g(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ and instead attempt to solve this taking into account adversarial noise in the data matrix $\mathbf{X}$. As in Ghaoui and Lebret (1997), Lewis (2002), Lewis and Pang (2009), Ben-Tal et al. (2009), Xu et al. (2010), this approach may take the form

$$\min_{\boldsymbol{\beta}} \max_{\mathbf{\Delta}\in\mathcal{U}} g(\mathbf{y} - (\mathbf{X} + \mathbf{\Delta})\boldsymbol{\beta}), \tag{1}$$

where the set $\mathcal{U} \subseteq \mathbb{R}^{m\times n}$ characterizes the user's belief about uncertainty on the data matrix $\mathbf{X}$. This set $\mathcal{U}$ is known in the

language of robust optimization (Ben-Tal et al., 2009; Bertsimas, Brown, & Caramanis, 2011) as an uncertainty set and the inner maximization problem $\max_{\mathbf{\Delta}\in\mathcal{U}} g(\mathbf{y}-(\mathbf{X}+\mathbf{\Delta})\boldsymbol{\beta})$ takes into account the worst-case error (measured via g) over $\mathcal{U}$. We call such a procedure *robustification* because it attempts to immunize or robustify the regression problem from structural uncertainty in the data. Such an adversarial or "worst-case" procedure is one of the key tenets of the area of robust optimization (Ben-Tal et al., 2009; Bertsimas et al., 2011).

As noted in the introduction, the adversarial perspective offers several attractive features. Let us first focus on settings when robustification coincides with a regularization problem. In such a case, the robustification identifies the adversarial perturbations the model is protected against, which can in turn provide additional insight into the behavior of different regularizers. Further, technical machinery developed for the construction of data-driven uncertainty sets in robust optimization (Bertsimas et al., 2013; Tulabandhula & Rudin, 2014) enables the potential for a principled framework for the design of regularization schemes, in turn addressing a complex modeling decision encountered in practice.

Moreover, the adversarial approach is of interest in its own right, even if robustification does not correspond directly to a regularization problem. This is evidenced in part by the burgeoning success of generative adversarial networks and other methodologies in deep learning (Goodfellow et al., 2014a; Goodfellow et al., 2014b; Shaham et al., 2015). Further, the worst-case approach often leads to a more straightforward analysis of properties of estimators (Xu et al., 2010) as well as algorithms for finding estimators (Ben-Tal, Hazan, Koren, & Mannor, 2015).

Let us now return to the robustification problem. A natural choice of an uncertainty set which gives rise to interpretability is the set $\mathcal{U}=\{\mathbf{\Delta}\in\mathbb{R}^{m\times n}:\|\mathbf{\Delta}\|\leq\lambda\}$, where $\|\cdot\|$ is some matrix norm and $\lambda>0$. One can then write $\max_{\mathbf{\Delta}\in\mathcal{U}} g(\mathbf{y}-(\mathbf{X}+\mathbf{\Delta})\boldsymbol{\beta})$ as

$$\begin{array}{ll}\max\limits_{\widetilde{\mathbf{X}}} & g(\mathbf{y}-\widetilde{\mathbf{X}}\boldsymbol{\beta})\\ \text{s.t.} & \|\mathbf{X}-\widetilde{\mathbf{X}}\|\leq\lambda,\end{array}$$

or the worst case error taken over all $\widetilde{\mathbf{X}}$ sufficiently close to the data matrix $\mathbf{X}$. In what follows, if $\|\cdot\|$ is a norm or seminorm, then we let $\mathcal{U}_{\|\cdot\|}$ denote the ball of radius λ in $\|\cdot\|$:

$$\mathcal{U}_{\|\cdot\|}=\{\mathbf{\Delta}:\|\mathbf{\Delta}\|\leq\lambda\}.$$

For example, $\mathcal{U}_{F_p}$, $\mathcal{U}_{\sigma_p}$, and $\mathcal{U}_{(h,g)}$ denote uncertainty sets under the norms F_p, σ_p, and (h, g), respectively. We assume $\lambda>0$ fixed for the remainder of the paper.

We briefly mention addressing uncertainty in $\mathbf{y}$. Suppose that we have a set $\mathcal{V}\subseteq\mathbb{R}^m$ which captures some belief about the uncertainty in $\mathbf{y}$. If again we have an uncertainty set $\mathcal{U}\subseteq\mathbb{R}^{m\times n}$, we may attempt to solve a problem of the form

$$\min_{\boldsymbol{\beta}}\max_{\substack{\boldsymbol{\delta}\in\mathcal{V}\\ \mathbf{\Delta}\in\mathcal{U}}} g(\mathbf{y}+\boldsymbol{\delta}-(\mathbf{X}+\mathbf{\Delta})\boldsymbol{\beta}).$$

We can instead work with a new loss function $\bar{g}$ defined as

$$\bar{g}(\mathbf{v}):=\max_{\boldsymbol{\delta}\in\mathcal{V}} g(\mathbf{v}+\boldsymbol{\delta}).$$

If g is convex, then so is $\bar{g}$. In this way, we can work with the problem in the form

$$\min_{\boldsymbol{\beta}}\max_{\mathbf{\Delta}\in\mathcal{U}}\bar{g}(\mathbf{y}-(\mathbf{X}+\mathbf{\Delta})\boldsymbol{\beta}),$$

where there is only uncertainty in $\mathbf{X}$. Throughout the remainder of this paper we will only consider such uncertainty.

Relation to robust statistics

There has been extensive work in the robust statistics community on statistical methods which perform well in noisy, real-world environments. As noted in Ben-Tal et al. (2009), the connection between robust optimization and robust statistics is not clear. We do not put forth any connection here, but briefly describe the development of robust statistics to appropriately contextualize our work. Instead of modeling noise via a distributional perspective, as is often the case in robust statistics, in this paper we choose to model it in a deterministic way using uncertainty sets. For a comprehensive description of the theoretical developments in robust statistics in the last half century, see the texts (Huber & Ronchetti, 2009; Rousseeuw, 1984) and the surveys (Hubert, Rousseeuw, & Aelst, 2008; Morgenthaler, 2007).

A central aspect of work in robust statistics is the development and use of a more general set of loss functions. (This is in contrast to the robust optimization approach, which generally results in the same nominal loss function with a new penalty; see Section 2.3 below.) For example, while least squares (the ℓ_2 loss) is known to perform well under Gaussian noise, it does not perform well under other types of noise, such as contaminated Gaussian noise. (Indeed, the Gaussian distribution was defined so that least squares is the optimal method under Gaussian noise (Rousseeuw, 1984).) In contrast, a method like LAD regression (the ℓ_1 loss) generally performs better than least squares with errors in $\mathbf{y}$, but not necessarily errors in the data matrix $\mathbf{X}$.

A more general class of such methods is M-estimators as proposed in Huber (1973) and since studied extensively (Huber & Ronchetti, 2009; Rousseeuw & Leroy, 1987). However, M-estimators lack desirable *finite sample breakdown* properties; in short, M-estimators perform very poorly in recovering the loadings $\boldsymbol{\beta}^*$ under gross errors in the data $(\mathbf{X}, \mathbf{y})$. To address some of these shortcomings, GM-estimators were introduced (Hampel, 1974; Hill, 1977; Mallows, 1975). Since these, many other estimators have been proposed. One such method is least quantile of squares regression (Rousseeuw, 1984) which has highly desirable robustness properties. There has been significant interest in new robust statistical methods in recent years with the increasing availability of large quantities of high-dimensional data, which often make reliable outlier detection difficult. For commentary on modern approaches to robust statistics, see (Bradic, Fan, & Wang, 2011; Fan, Fan, & Barut, 2014; Hubert et al., 2008) and references therein.

Relation to error-in-variable models

Another class of statistical models which are particularly relevant for the work contained herein are error-in-variable models (Carroll, Ruppert, Stefanski, & Crainiceanu, 2006). One approach to such a problem takes the form

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^n,\ \mathbf{\Delta}\in\mathbb{R}^{m\times n}} g(\mathbf{y}-(\mathbf{X}+\mathbf{\Delta})\boldsymbol{\beta})+P(\mathbf{\Delta}),$$

where P is a penalty function which takes into account the complexity of possible perturbations $\mathbf{\Delta}$ to the data matrix $\mathbf{X}$. A canonical example of such a method is total least squares (Golub & Loan, 1980; Markovsky & Huffel, 2007), which can be written for fixed $\tau>0$ as

$$\min_{\boldsymbol{\beta},\mathbf{\Delta}}\|\mathbf{y}-(\mathbf{X}+\mathbf{\Delta})\boldsymbol{\beta}\|_2+\tau\|\mathbf{\Delta}\|_F.$$

An equivalent way of writing such problems is, instead of penalized form, as constrained optimization problems. In particular, the constrained version generically takes the form

$$\min_{\boldsymbol{\beta}}\min_{\substack{\mathbf{\Delta}:\\ P(\mathbf{\Delta})\leq\eta}} g(\mathbf{y}-(\mathbf{X}+\mathbf{\Delta})\boldsymbol{\beta}), \tag{2}$$

where $\eta>0$ is fixed. Under the representation in (2), the comparison with the robust optimization approach in (1) becomes immediate. While the classical error-in-variables approach takes an optimistic view on uncertainty in the data matrix $\mathbf{X}$, and finds loadings $\boldsymbol{\beta}$ on the new "corrected" data matrix $\mathbf{X}+\mathbf{\Delta}$, the

minimax approach of (1) considers protections against adversarial perturbations in the data which maximally increase the loss.

One of the advantages of the adversarial approach to error-in-variables is that it enables a direct analysis of certain statistical properties, such as asymptotic consistency of estimators (*c.f.* Caramanis et al., 2011; Xu et al., 2010). In contrast, analyzing the consistency of estimators attained by a model such as total least squares is a complex issue (Kukush, Markovsky, & Huffel, 2005).

2.3. Equivalence of robustification and regularization

A natural question is when do the procedures of regularization and robustification coincide. This problem was first studied in Ghaoui and Lebret (1997) in the context of uncertain least squares problems and has been extended to more general settings in Caramanis et al. (2011); Xu et al. (2010) and most comprehensively in Ben-Tal et al. (2009). In this section, we present settings in which robustification is equivalent to regularization. When such an equivalence holds, tools from robust optimization can be used to analyze properties of the regularization problem (*c.f.* Caramanis et al., 2011; Xu et al., 2010).

We begin with a general result on robustification under induced seminorm uncertainty sets.

Theorem 1. *If $g:\mathbb{R}^m\to\mathbb{R}$ is a seminorm which is not identically zero and $h:\mathbb{R}^n\to\mathbb{R}$ is a norm, then for any $\mathbf{z}\in\mathbb{R}^m$ and $\boldsymbol{\beta}\in\mathbb{R}^n$*

$$\max_{\boldsymbol{\Delta}\in\mathcal{U}_{(h,g)}} g(\mathbf{z}+\boldsymbol{\Delta}\boldsymbol{\beta}) = g(\mathbf{z}) + \lambda h(\boldsymbol{\beta}),$$

where $\mathcal{U}_{(h,g)} = \{\boldsymbol{\Delta} : \|\boldsymbol{\Delta}\|_{(h,g)} \le \lambda\}$.

Proof. From the triangle inequality $g(\mathbf{z}+\boldsymbol{\Delta}\boldsymbol{\beta}) \le g(\mathbf{z}) + g(\boldsymbol{\Delta}\boldsymbol{\beta}) \le g(\mathbf{z}) + \lambda h(\boldsymbol{\beta})$ for any $\boldsymbol{\Delta}\in\mathcal{U} := \mathcal{U}_{(h,g)}$. We next show that there exists some $\boldsymbol{\Delta}\in\mathcal{U}$ so that $g(\mathbf{z}+\boldsymbol{\Delta}\boldsymbol{\beta}) = g(\mathbf{z}) + \lambda h(\boldsymbol{\beta})$. Let $\mathbf{v}\in\mathbb{R}^n$ so that $\mathbf{v}\in \operatorname{argmax}_{h^*(\mathbf{v})=1} \mathbf{v}'\boldsymbol{\beta}$, where h^* is the dual norm of h. Note in particular that $\mathbf{v}'\boldsymbol{\beta} = h(\boldsymbol{\beta})$ by the definition of the dual norm h^*. For now suppose that $g(\mathbf{z}) \neq 0$. Define the rank one matrix $\widehat{\boldsymbol{\Delta}} = \frac{\lambda}{g(\mathbf{z})}\mathbf{z}\mathbf{v}'$. Observe that

$$g(\mathbf{z}+\widehat{\boldsymbol{\Delta}}\boldsymbol{\beta}) = g\left(\mathbf{z} + \frac{\lambda h(\boldsymbol{\beta})}{g(\mathbf{z})}\mathbf{z}\right) = \frac{g(\mathbf{z}) + \lambda h(\boldsymbol{\beta})}{g(\mathbf{z})} g(\mathbf{z}) = g(\mathbf{z}) + \lambda h(\boldsymbol{\beta}).$$

We next show that $\widehat{\boldsymbol{\Delta}}\in\mathcal{U}$. Observe that for any $\mathbf{x}\in\mathbb{R}^n$ that

$$g(\widehat{\boldsymbol{\Delta}}\mathbf{x}) = g\left(\frac{\lambda\mathbf{v}'\mathbf{x}}{g(\mathbf{z})}\mathbf{z}\right) = \lambda|\mathbf{v}'\mathbf{x}| \le \lambda h(\mathbf{x}) h^*(\mathbf{v}) = \lambda h(\mathbf{x}),$$

where the final inequality follows by definition of the dual norm. Hence $\widehat{\boldsymbol{\Delta}}\in\mathcal{U}$, as desired.

We now consider the case when $g(\mathbf{z}) = 0$. Let $\mathbf{u}\in\mathbb{R}^m$ so that $g(\mathbf{u}) = 1$ (because g is not identically zero there exists some $\mathbf{u}$ so that $g(\mathbf{u}) > 0$, and so by homogeneity of g we can take $\mathbf{u}$ so that $g(\mathbf{u}) = 1$). Let $\mathbf{v}$ be as before. Now define $\widehat{\boldsymbol{\Delta}} = \lambda\mathbf{u}\mathbf{v}'$. We observe that

$$g(\mathbf{z}+\widehat{\boldsymbol{\Delta}}\boldsymbol{\beta}) = g(\mathbf{z}+\lambda\mathbf{u}\mathbf{v}'\boldsymbol{\beta}) \le g(\mathbf{z}) + \lambda|\mathbf{v}'\boldsymbol{\beta}| g(\mathbf{u}) = \lambda h(\boldsymbol{\beta}).$$

Now, by the reverse triangle inequality,

$$g(\mathbf{z}+\widehat{\boldsymbol{\Delta}}\boldsymbol{\beta}) \ge g(\widehat{\boldsymbol{\Delta}}\boldsymbol{\beta}) - g(\mathbf{z}) = g(\widehat{\boldsymbol{\Delta}}\boldsymbol{\beta}) = \lambda h(\boldsymbol{\beta}),$$

and therefore $g(\mathbf{z}+\widehat{\boldsymbol{\Delta}}\boldsymbol{\beta}) = \lambda h(\boldsymbol{\beta}) = g(\mathbf{z}) + \lambda h(\boldsymbol{\beta})$. The proof that $\widehat{\boldsymbol{\Delta}}\in\mathcal{U}$ is identical to the case when $g(\mathbf{z}) \neq 0$. This completes the proof. □

This result implies as a corollary known results on the connection between robustification and regularization as found in Xu et al. (2010), Ben-Tal et al. (2009), Caramanis et al. (2011) and references therein.

Corollary 1 (Ben-Tal et al., 2009; Caramanis et al., 2011; Xu et al., 2010)**.** *If $p, q\in[1,\infty]$ then*

$$\min_{\boldsymbol{\beta}} \max_{\boldsymbol{\Delta}\in\mathcal{U}_{(q,p)}} \|\mathbf{y} - (\mathbf{X}+\boldsymbol{\Delta})\boldsymbol{\beta}\|_p = \min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_p + \lambda\|\boldsymbol{\beta}\|_q.$$

In particular, for $p = q = 2$ we recover regularized least squares as a robustification; likewise, for $p = 2$ and $q = 1$ we recover the Lasso.[2]

Theorem 2 (Ben-Tal et al., 2009; Caramanis et al., 2011; Xu et al., 2010)**.** *One has the following for any $p, q\in[1,\infty]$:*

$$\min_{\boldsymbol{\beta}} \max_{\boldsymbol{\Delta}\in\mathcal{U}_{F_p}} \|\mathbf{y} - (\mathbf{X}+\boldsymbol{\Delta})\boldsymbol{\beta}\|_p = \min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_p + \lambda\|\boldsymbol{\beta}\|_{p^*},$$

where p^* is the conjugate of p. Similarly,

$$\min_{\boldsymbol{\beta}} \max_{\boldsymbol{\Delta}\in\mathcal{U}_{\sigma_q}} \|\mathbf{y} - (\mathbf{X}+\boldsymbol{\Delta})\boldsymbol{\beta}\|_2 = \min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2 + \lambda\|\boldsymbol{\beta}\|_2.$$

Observe that regularized least squares arises again under all uncertainty sets defined by the spectral norms σ_q when the loss function is $g = \ell_2$. Now we continue with a remark on how Lasso arises through regularization. See Xu et al. (2010) for comprehensive work on the robustness and sparsity implications of Lasso as interpreted through such a robustification considered in this paper.

Remark 1. *As per Corollary 1 it is known that Lasso arises as uncertain ℓ_2 regression with uncertainty set $\mathcal{U} := \mathcal{U}_{(1,2)}$ (Xu et al., 2010). As with Theorem 1, one might argue that the ℓ_1 penalizer arises as an artifact of the model of uncertainty. We remark that one can derive the set $\mathcal{U}$ as an induced uncertainty set defined using the "true" non-convex penalty ℓ_0, where $\|\boldsymbol{\beta}\|_0 := |\{i:\ \beta_i \neq 0\}|$. To be precise, for any $p\in[1,\infty]$ and for $\Gamma = \{\boldsymbol{\beta}\in\mathbb{R}^n : \|\boldsymbol{\beta}\|_p \le 1\}$ we claim that*

$$\mathcal{U}' := \left\{\boldsymbol{\Delta} : \max_{\boldsymbol{\beta}\in\Gamma} \frac{\|\boldsymbol{\Delta}\boldsymbol{\beta}\|_2}{\|\boldsymbol{\beta}\|_0} \le \lambda\right\}$$

satisfies $\mathcal{U} = \mathcal{U}'$. This is summarized, with an additional representation $\mathcal{U}''$ as used in Xu et al. (2010), in the following proposition.

Proposition 1. *If $\mathcal{U} = \mathcal{U}_{(1,2)}$, $\mathcal{U}' = \{\boldsymbol{\Delta} : \|\boldsymbol{\Delta}\boldsymbol{\beta}\|_2 \le \lambda\|\boldsymbol{\beta}\|_0\ \forall\|\boldsymbol{\beta}\|_p \le 1\}$ for an arbitrary $p\in[1,\infty]$, and $\mathcal{U}'' = \{\boldsymbol{\Delta} : \|\boldsymbol{\Delta}_i\|_2 \le \lambda\ \forall i\}$, where $\boldsymbol{\Delta}_i$ is the ith column of $\boldsymbol{\Delta}$, then $\mathcal{U} = \mathcal{U}' = \mathcal{U}''$.*

Proof. We first show that $\mathcal{U} = \mathcal{U}'$. Because $\|\boldsymbol{\beta}\|_1 \le \|\boldsymbol{\beta}\|_0$ for all $\boldsymbol{\beta}\in\mathbb{R}^n$ with $\|\boldsymbol{\beta}\|_p \le 1$, we have that $\mathcal{U}\subseteq\mathcal{U}'$. Now suppose that $\boldsymbol{\Delta}\in\mathcal{U}'$. Then for any $\boldsymbol{\beta}\in\mathbb{R}^n$, we have that

$$\|\boldsymbol{\Delta}\boldsymbol{\beta}\|_2 = \left\|\sum_i \beta_i\boldsymbol{\Delta}\mathbf{e}_i\right\|_2 \le \sum_i |\beta_i|\,\|\boldsymbol{\Delta}\mathbf{e}_i\|_2 \le \sum_i |\beta_i|\lambda = \lambda\|\boldsymbol{\beta}\|_1,$$

where $\{\mathbf{e}_i\}_{i=1}^n$ is the standard orthonormal basis for $\mathbb{R}^n$. Hence, $\boldsymbol{\Delta}\in\mathcal{U}$ and therefore $\mathcal{U}'\subseteq\mathcal{U}$. Combining with the previous direction gives $\mathcal{U} = \mathcal{U}'$.

We now prove that $\mathcal{U} = \mathcal{U}''$. That $\mathcal{U}''\subseteq\mathcal{U}$ is essentially obvious; $\mathcal{U}\subseteq\mathcal{U}''$ follows by considering $\boldsymbol{\beta}\in\{\mathbf{e}_i\}_{i=1}^n$. This completes the proof. □

This proposition implies that ℓ_1 arises from the robustification setting without directly appealing to standard convexity arguments for why ℓ_1 should be used to replace ℓ_0 (which use the fact that ℓ_1 is the so-called convex envelope of ℓ_0 on $[-1,1]^n$, see e.g. Boyd and Vandenberghe (2004).

In light of the above discussion, it is not difficult to show that other Lasso-like methods can also be expressed as an adversarial

[2] Strictly speaking, we recover *equivalent* problems to regularized least squares and Lasso, respectively. We take the usual convention and overlook this technicality (see Ben-Tal et al., 2009 for a discussion). For completeness, we note that one can work directly with the true ℓ_2^2 loss function, although at the cost of requiring more complicated uncertainty sets to recover equivalence results.

robustification, supporting the flexibility and versatility of such an approach. One such example is the elastic net (De Mol, De Vito, & Rosasco, 2009; Mosci, Rosasco, Santoro, Verri, & Villa, 2010; Zou & Hastie, 2005), a hybridized version of ridge regression and the Lasso. An equivalent representation of the elastic net is as follows:

$$\min_{\boldsymbol{\beta}} \|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2+\lambda\|\boldsymbol{\beta}\|_1+\mu\|\boldsymbol{\beta}\|_2.$$

As per Theorem 2, this can be written exactly as

$$\min_{\boldsymbol{\beta}} \max_{\substack{\boldsymbol{\Delta},\boldsymbol{\Delta}':\\ \|\boldsymbol{\Delta}\|_{F_\infty}\le\lambda\\ \|\boldsymbol{\Delta}'\|_{\bar{2}}\le\mu}} \|\mathbf{y}-(\mathbf{X}+\boldsymbol{\Delta}+\boldsymbol{\Delta}')\boldsymbol{\beta}\|_2.$$

Under this interpretation, we see that λ and μ directly control the tradeoff between two different types of perturbations: "feature-wise" perturbations $\boldsymbol{\Delta}$ (controlled via λ and the F_∞ norm) and "global" perturbations $\boldsymbol{\Delta}'$ (controlled via μ and the F_2 norm).

We conclude this section with another example of when robustification is equivalent to regularization for the case of LAD (ℓ_1) and maximum absolute deviation (ℓ_∞) regression under row-wise uncertainty.

Theorem 3 (Xu et al., 2010). *Fix $q\in[1,\infty]$ and let $\mathcal{U}=\{\boldsymbol{\Delta}:\|\boldsymbol{\delta}_i\|_q\le\lambda\ \forall i\}$, where $\boldsymbol{\delta}_i$ is the ith row of $\boldsymbol{\Delta}\in\mathbb{R}^{m\times n}$. Then*

$$\min_{\boldsymbol{\beta}}\max_{\boldsymbol{\Delta}\in\mathcal{U}}\|\mathbf{y}-(\mathbf{X}+\boldsymbol{\Delta})\boldsymbol{\beta}\|_1=\min_{\boldsymbol{\beta}}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_1+m\lambda\|\boldsymbol{\beta}\|_{q^*}$$

and

$$\min_{\boldsymbol{\beta}}\max_{\boldsymbol{\Delta}\in\mathcal{U}}\|\mathbf{y}-(\mathbf{X}+\boldsymbol{\Delta})\boldsymbol{\beta}\|_\infty=\min_{\boldsymbol{\beta}}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_\infty+\lambda\|\boldsymbol{\beta}\|_{q^*}.$$

For completeness, we note that the uncertainty set $\mathcal{U}=\{\boldsymbol{\Delta}:\|\boldsymbol{\delta}_i\|_q\le\lambda\ \forall i\}$ considered in Theorem 3 is actually an induced uncertainty set, namely, $\mathcal{U}=\mathcal{U}_{(q^*,\infty)}$.

2.4. Non-equivalence of robustification and regularization

In contrast to previous work studying robustification for regression, which primarily addresses tractability of solving the new uncertain problem (Ben-Tal et al., 2009) or the implications for Lasso (Xu et al., 2010), we instead focus our attention on characterization of the equivalence between robustification and regularization. We begin with a regularization upper bound on robustification problems.

Proposition 2. *Let $\mathcal{U}\subseteq\mathbb{R}^{m\times n}$ be any non-empty, compact set and $g:\mathbb{R}^m\to\mathbb{R}$ a seminorm. Then there exists some seminorm $\bar{h}:\mathbb{R}^n\to\mathbb{R}$ so that for any $\mathbf{z}\in\mathbb{R}^m$, $\boldsymbol{\beta}\in\mathbb{R}^n$,*

$$\max_{\boldsymbol{\Delta}\in\mathcal{U}} g(\mathbf{z}+\boldsymbol{\Delta}\boldsymbol{\beta})\le g(\mathbf{z})+\bar{h}(\boldsymbol{\beta}),$$

with equality when $\mathbf{z}=\mathbf{0}$.

Proof. Let $\bar{h}:\mathbb{R}^n\to\mathbb{R}$ be defined as

$$\bar{h}(\boldsymbol{\beta}):=\max_{\boldsymbol{\Delta}\in\mathcal{U}} g(\boldsymbol{\Delta}\boldsymbol{\beta}).$$

To show that $\bar{h}$ is a seminorm we must show it satisfies absolute homogeneity and the triangle inequality. For any $\boldsymbol{\beta}\in\mathbb{R}^n$ and $\alpha\in\mathbb{R}$,

$$\begin{aligned}\bar{h}(\alpha\boldsymbol{\beta})&=\max_{\boldsymbol{\Delta}\in\mathcal{U}} g(\boldsymbol{\Delta}(\alpha\boldsymbol{\beta}))=\max_{\boldsymbol{\Delta}\in\mathcal{U}}|\alpha|g(\boldsymbol{\Delta}\boldsymbol{\beta})=|\alpha|\left(\max_{\boldsymbol{\Delta}\in\mathcal{U}} g(\boldsymbol{\Delta}\boldsymbol{\beta})\right)\\&=|\alpha|\bar{h}(\boldsymbol{\beta}),\end{aligned}$$

so absolute homogeneity is satisfied. Similarly, if $\boldsymbol{\beta},\boldsymbol{\gamma}\in\mathbb{R}^n$,

$$\begin{aligned}\bar{h}(\boldsymbol{\beta}+\boldsymbol{\gamma})=\max_{\boldsymbol{\Delta}\in\mathcal{U}} g(\boldsymbol{\Delta}(\boldsymbol{\beta}+\boldsymbol{\gamma}))&\le\max_{\boldsymbol{\Delta}\in\mathcal{U}}\left[g(\boldsymbol{\Delta}\boldsymbol{\beta})+g(\boldsymbol{\Delta}\boldsymbol{\gamma})\right]\\&\le\left(\max_{\boldsymbol{\Delta}\in\mathcal{U}} g(\boldsymbol{\Delta}\boldsymbol{\beta})\right)+\left(\max_{\boldsymbol{\Delta}\in\mathcal{U}} g(\boldsymbol{\Delta}\boldsymbol{\gamma})\right),\end{aligned}$$

and hence the triangle inequality is satisfied. Therefore, $\bar{h}$ is a seminorm which satisfies the desired properties, completing the proof. □

When equality is attained for all pairs $(\mathbf{z},\boldsymbol{\beta})\in\mathbb{R}^m\times\mathbb{R}^n$, we are in the regime of the previous section, and we say that robustification under $\mathcal{U}$ is equivalent to regularization under $\bar{h}$. We now discuss a variety of explicit settings in which regularization only provides upper and lower bounds to the true robustified problem.

Fix $p,q\in[1,\infty]$. Consider the robust ℓ_p regression problem

$$\min_{\boldsymbol{\beta}}\max_{\boldsymbol{\Delta}\in\mathcal{U}_{F_q}}\|\mathbf{y}-(\mathbf{X}+\boldsymbol{\Delta})\boldsymbol{\beta}\|_p,$$

where $\mathcal{U}_{F_q}=\{\boldsymbol{\Delta}\in\mathbb{R}^{m\times n}:\|\boldsymbol{\Delta}\|_{F_q}\le\lambda\}$. In the case when $p=q$ we saw earlier (Theorem 2) that one exactly recovers ℓ_p regression with an ℓ_{p^*} penalty:

$$\min_{\boldsymbol{\beta}}\max_{\boldsymbol{\Delta}\in\mathcal{U}_{F_p}}\|\mathbf{y}-(\mathbf{X}+\boldsymbol{\Delta})\boldsymbol{\beta}\|_p=\min_{\boldsymbol{\beta}}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_p+\lambda\|\boldsymbol{\beta}\|_{p^*}.$$

Let us now consider the case when $p\neq q$. We claim that regularization (with $\bar{h}$) is no longer equivalent to robustification (with $\mathcal{U}_{F_q}$) unless $p\in\{1,\infty\}$. Applying Proposition 2, one has for any $\mathbf{z}\in\mathbb{R}^m$ that

$$\max_{\boldsymbol{\Delta}\in\mathcal{U}_{F_q}}\|\mathbf{z}+\boldsymbol{\Delta}\boldsymbol{\beta}\|_p\le\|\mathbf{z}\|_p+\bar{h}(\boldsymbol{\beta}),$$

where $\bar{h}=\max_{\boldsymbol{\Delta}\in\mathcal{U}_{F_q}}\|\boldsymbol{\Delta}\boldsymbol{\beta}\|_p$ is a norm (when $p=q$, this is precisely the ℓ_{p^*} norm, multiplied by λ). Here we can compute $\bar{h}$. To do this we first define a discrepancy function as follows:

Definition 1. For $a,b\in[1,\infty]$ define the discrepancy function $\delta_m(a,b)$ as

$$\delta_m(a,b):=\max\{\|\mathbf{u}\|_a:\mathbf{u}\in\mathbb{R}^m,\ \|\mathbf{u}\|_b=1\}.$$

This discrepancy function is computable and well-known (see e.g. Horn & Johnson, 2013):

$$\delta_m(a,b)=\begin{cases}m^{1/a-1/b}, & \text{if } a\le b\\ 1, & \text{if } a>b.\end{cases}$$

It satisfies $1\le\delta_m(a,b)\le m$ and $\delta_m(a,b)$ is continuous in a and b. One has that $\delta_m(a,b)=\delta_m(b,a)=1$ if and only if $a=b$ (so long as $m\ge2$). Using this, we now proceed with the theorem. The proof applies basic tools from real analysis and is contained in Appendix A.

Theorem 4.

(a) *For any $\mathbf{z}\in\mathbb{R}^m$ and $\boldsymbol{\beta}\in\mathbb{R}^n$,*

$$\max_{\boldsymbol{\Delta}\in\mathcal{U}_{F_q}}\|\mathbf{z}+\boldsymbol{\Delta}\boldsymbol{\beta}\|_p\le\|\mathbf{z}\|_p+\lambda\delta_m(p,q)\|\boldsymbol{\beta}\|_{q^*}. \tag{3}$$

(b) *When $p\in\{1,\infty\}$, there is equality in (3) for all $(\mathbf{z},\boldsymbol{\beta})$.*

(c) *When $p\in(1,\infty)$ and $p\neq q$, for any $\boldsymbol{\beta}\neq\mathbf{0}$ the set of $\mathbf{z}\in\mathbb{R}^m$ for which the inequality (3) holds at equality is a finite union of one-dimensional subspaces (so long as $m\ge2$). Hence, for any $\boldsymbol{\beta}\neq\mathbf{0}$ the inequality in (3) is strict for almost all $\mathbf{z}$.*

(d) *For $p\in(1,\infty)$, one has for all $\mathbf{z}\in\mathbb{R}^m$ and $\boldsymbol{\beta}\in\mathbb{R}^n$ that*

$$\|\mathbf{z}\|_p+\frac{\lambda}{\delta_m(q,p)}\|\boldsymbol{\beta}\|_{q^*}\le\max_{\boldsymbol{\Delta}\in\mathcal{U}_{F_q}}\|\mathbf{z}+\boldsymbol{\Delta}\boldsymbol{\beta}\|_p. \tag{4}$$

(e) *For $p\in(1,\infty)$, the lower bound in (4) is best possible in the sense that the gap can be arbitrarily small, i.e., for any $\boldsymbol{\beta}\in\mathbb{R}^n$,*

$$\inf_{\mathbf{z}}\left(\max_{\boldsymbol{\Delta}\in\mathcal{U}_{F_q}}\|\mathbf{z}+\boldsymbol{\Delta}\boldsymbol{\beta}\|_p-\|\mathbf{z}\|_p-\frac{\lambda}{\delta_m(q,p)}\|\boldsymbol{\beta}\|_{q^*}\right)=0.$$

Theorem 4 characterizes precisely when robustification under $\mathcal{U}_{F_q}$ is equivalent to regularization for the case of ℓ_p regression. In particular, when $p \neq q$ and $p \in (1, \infty)$, the two are *not* equivalent, and one only has that

$$\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_p + \frac{\lambda}{\delta_m(q,p)}\|\boldsymbol{\beta}\|_{q^*} \leq \min_{\boldsymbol{\beta}} \max_{\boldsymbol{\Delta} \in \mathcal{U}_{F_q}} \|\mathbf{y} - (\mathbf{X} + \boldsymbol{\Delta})\boldsymbol{\beta}\|_p \leq \min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_p + \lambda \delta_m(p,q)\|\boldsymbol{\beta}\|_{q^*}.$$

Further, we have shown that these upper and lower bounds are the *best possible* (Theorem 4, parts (c) and (e)). While ℓ_p regression with uncertainty set $\mathcal{U}_{F_q}$ for $p \neq q$ and $p \in (1, \infty)$ still has both upper and lower bounds which correspond to regularization (with different regularization parameters $\bar{\lambda} \in [\lambda/\delta_m(q,p), \lambda\delta_m(p,q)]$), we emphasize that in this case there is no longer the direct connection between the parameter garnering the magnitude of uncertainty (λ) and the parameter for regularization ($\bar{\lambda}$).

Example 1. As a concrete example, consider the implications of Theorem 4 when $p = 2$ and $q = \infty$. We have that

$$\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2 + \lambda\|\boldsymbol{\beta}\|_1 \leq \min_{\boldsymbol{\beta}} \max_{\boldsymbol{\Delta} \in \mathcal{U}_{F_\infty}} \|\mathbf{y} - (\mathbf{X} + \boldsymbol{\Delta})\boldsymbol{\beta}\|_p \leq \min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2 + \sqrt{m}\lambda\|\boldsymbol{\beta}\|_1.$$

In this case, robustification is *not* equivalent to regularization. In particular, in the regime where there are many data points (i.e. m is large), the gap appearing between the different problems can be quite large.

Let us remark that in general, lower bounds on $\max_{\boldsymbol{\Delta} \in \mathcal{U}} g(\mathbf{z} + \boldsymbol{\Delta}\boldsymbol{\beta})$ will depend on the structure of $\mathcal{U}$ and may not exist (except for the trivial lower bound of $g(\mathbf{z})$) in some scenarios. However, it is easy to show that if $\mathcal{U}$ is compact and zero is in the interior of $\mathcal{U}$, then there exists some $\underline{\lambda} \in (0, 1]$ so that

$$\max_{\boldsymbol{\Delta} \in \mathcal{U}} g(\mathbf{z} + \boldsymbol{\Delta}\boldsymbol{\beta}) \geq g(\mathbf{z}) + \underline{\lambda}\bar{h}(\boldsymbol{\beta}).$$

Before proceeding with other choices of uncertainty sets, it is important to make a further distinction about the general non-equivalence of robustification and regularization as presented in Theorem 4. In particular, it is simple to construct examples (see Appendix B) which imply the following strong existential result:

Theorem 5. *In a setting when robustification and regularization are not equivalent, it is possible for the two problems to have different optimal solutions. In particular,*

$$\boldsymbol{\beta}^* \in \operatorname*{argmin}_{\boldsymbol{\beta}} \max_{\boldsymbol{\Delta} \in \mathcal{U}} g(\mathbf{y} - (\mathbf{X} + \boldsymbol{\Delta})\boldsymbol{\beta})$$

is not necessarily a solution of

$$\min_{\boldsymbol{\beta}} g(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \tilde{\lambda}\bar{h}(\boldsymbol{\beta})$$

for any $\tilde{\lambda} > 0$, and vice versa.

As a result, when robustification and regularization do not coincide, they can induce structurally distinct solutions. In other words, the regularization path (as $\tilde{\lambda} \in (0, \infty)$ varies) and the robustification path (as the radius $\lambda \in (0, \infty)$ of $\mathcal{U}$ varies) can be different.

We now proceed to analyze another setting in which robustification is not equivalent to regularization. The setting, in line with Theorem 2, is ℓ_p regression under spectral uncertainty sets $\mathcal{U}_{\sigma_q}$. As per Theorem 2, one has that

$$\min_{\boldsymbol{\beta}} \max_{\boldsymbol{\Delta} \in \mathcal{U}_{\sigma_q}} \|\mathbf{y} - (\mathbf{X} + \boldsymbol{\Delta})\boldsymbol{\beta}\|_2 = \min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2 + \lambda\|\boldsymbol{\beta}\|_2$$

for any $q \in [1, \infty]$. This result on the "universality" of RLS under a variety of uncertainty sets relies on the fact that the ℓ_2 norm underlies spectral decompositions; namely, one can write any matrix $\mathbf{X}$ as $\sum_i \mu_i \mathbf{u}_i \mathbf{v}_i'$, where $\{\mu_i\}_i$ are the singular values of $\mathbf{X}$, $\{\mathbf{u}_i\}_i$ and $\{\mathbf{v}_i\}_i$ are the left and right singular vectors of $\mathbf{X}$, respectively, and $\|\mathbf{u}_i\|_2 = \|\mathbf{v}_i\|_2 = 1$ for all i.

A natural question is what happens when the loss function ℓ_2, a modeling choice, is replaced by ℓ_p, where $p \in [1, \infty]$. We claim that for $p \notin \{1, 2, \infty\}$, robustification under $\mathcal{U}_{\sigma_q}$ is no longer equivalent to regularization. In light of Theorem 4 this is not difficult to prove. We find that the choice of $q \in [1, \infty]$, as before, is inconsequential. We summarize this in the following proposition:

Proposition 3. *For any $\mathbf{z} \in \mathbb{R}^m$ and $\boldsymbol{\beta} \in \mathbb{R}^n$,*

$$\max_{\boldsymbol{\Delta} \in \mathcal{U}_{\sigma_q}} \|\mathbf{z} + \boldsymbol{\Delta}\boldsymbol{\beta}\|_p \leq \|\mathbf{z}\|_p + \lambda\delta_m(p,2)\|\boldsymbol{\beta}\|_2. \tag{5}$$

In particular, if $p \in \{1, 2, \infty\}$, there is equality in (5) for all $(\mathbf{z}, \boldsymbol{\beta})$. If $p \notin \{1, 2, \infty\}$, then for any $\boldsymbol{\beta} \neq \mathbf{0}$ the inequality in (5) is strict for almost all $\mathbf{z}$ (when $m \geq 2$). Further, for $p \notin \{1, 2, \infty\}$ one has the lower bound

$$\|\mathbf{z}\|_p + \frac{\lambda}{\delta_m(2,p)}\|\boldsymbol{\beta}\|_2 \leq \max_{\boldsymbol{\Delta} \in \mathcal{U}_{\sigma_q}} \|\mathbf{z} + \boldsymbol{\Delta}\boldsymbol{\beta}\|_p,$$

whose gap is arbitrarily small for all $\boldsymbol{\beta}$.

Proof. This result is Theorem 4 in disguise. This follows by noting that

$$\max_{\boldsymbol{\Delta} \in \mathcal{U}_{\sigma_q}} \|\mathbf{z} + \boldsymbol{\Delta}\boldsymbol{\beta}\|_p = \max_{\boldsymbol{\Delta} \in \mathcal{U}_{F_2}} \|\mathbf{z} + \boldsymbol{\Delta}\boldsymbol{\beta}\|_p$$

and directly applying the preceding results. □

We now consider a third setting for ℓ_p regression, this time subject to uncertainty $\mathcal{U}_{(q,r)}$; this is a generalized version of the problems considered in Theorems 1 and 3. From Theorem 1 we know that if $p = r$, then

$$\min_{\boldsymbol{\beta}} \max_{\boldsymbol{\Delta} \in \mathcal{U}_{(q,p)}} \|\mathbf{y} - (\mathbf{X} + \boldsymbol{\Delta})\boldsymbol{\beta}\|_p = \min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_p + \lambda\|\boldsymbol{\beta}\|_q.$$

Similarly, as per Theorem 3, when $r = \infty$ and $p \in \{1, \infty\}$,

$$\min_{\boldsymbol{\beta}} \max_{\boldsymbol{\Delta} \in \mathcal{U}_{(q,\infty)}} \|\mathbf{y} - (\mathbf{X} + \boldsymbol{\Delta})\boldsymbol{\beta}\|_p = \min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_p + \lambda\delta_m(p,\infty)\|\boldsymbol{\beta}\|_q.$$

Given these results, it is natural to inquire what happens for more general choices of induced uncertainty set $\mathcal{U}_{(q,r)}$. As before with Theorem 4, we have a complete characterization of the equivalence of robustification and regularization for ℓ_p regression with uncertainty set $\mathcal{U}_{(q,r)}$:

Proposition 4. *For any $\mathbf{z} \in \mathbb{R}^m$ and $\boldsymbol{\beta} \in \mathbb{R}^n$,*

$$\max_{\boldsymbol{\Delta} \in \mathcal{U}_{(q,r)}} \|\mathbf{z} + \boldsymbol{\Delta}\boldsymbol{\beta}\|_p \leq \|\mathbf{z}\|_p + \lambda\delta_m(p,r)\|\boldsymbol{\beta}\|_q. \tag{6}$$

In particular, if $p \in \{1, r, \infty\}$, there is equality in (5) for all $(\mathbf{z}, \boldsymbol{\beta})$. If $p \in (1, \infty)$ and $p \neq r$, then for any $\boldsymbol{\beta} \neq \mathbf{0}$ the inequality in (6) is strict for almost all $\mathbf{z}$ (when $m \geq 2$). Further, for $p \in (1, \infty)$ with $p \neq r$ one has the lower bound

$$\|\mathbf{z}\|_p + \frac{\lambda}{\delta_m(r,p)}\|\boldsymbol{\beta}\|_q \leq \max_{\boldsymbol{\Delta} \in \mathcal{U}_{(q,r)}} \|\mathbf{z} + \boldsymbol{\Delta}\boldsymbol{\beta}\|_p,$$

whose gap is arbitrarily small for all $\boldsymbol{\beta}$.

Proof. The proof follows the argument given in the proof of Theorem 4. Here we simply note that now one uses the fact that

$$\max_{\boldsymbol{\Delta} \in \mathcal{U}_{(q,r)}} \|\mathbf{z} + \boldsymbol{\Delta}\boldsymbol{\beta}\|_p = \max_{\|\mathbf{u}\|_r \leq \lambda\|\boldsymbol{\beta}\|_q} \|\mathbf{z} + \mathbf{u}\|_p.$$

□

We summarize all of the results on linear regression in Table 2.

Table 2
Summary of equivalencies for robustification with uncertainty set $\mathcal{U}$ and regularization with penalty $\bar{h}$, where $\bar{h}$ is as given in Proposition 2. Here by equivalence we mean that for all $\mathbf{z} \in \mathbb{R}^m$ and $\boldsymbol{\beta} \in \mathbb{R}^n$, $\max_{\boldsymbol{\Delta} \in \mathcal{U}} g(\mathbf{z} + \boldsymbol{\beta}) = g(\mathbf{z}) + \bar{h}(\boldsymbol{\beta})$, where g is the loss function, i.e., the upper bound $\bar{h}$ is also a lower bound. Here δ_m is as in Theorem 4. Throughout $p, q \in [1, \infty]$ and $m \geq 2$. Here $\boldsymbol{\delta}_i$ denotes the ith row of $\boldsymbol{\Delta}$.

Loss function seminorm g	Uncertainty set $\mathcal{U}$ $\mathcal{U}_{(h,g)}$ (h norm)	$\bar{h}(\boldsymbol{\beta})$ $\lambda h(\boldsymbol{\beta})$	Equivalence if and only if always
ℓ_p	$\mathcal{U}_{\sigma_q}$	$\lambda \delta_m(p, 2) \|\boldsymbol{\beta}\|_2$	$p \in \{1, 2, \infty\}$
ℓ_p	$\mathcal{U}_{\ell_q}$	$\lambda \delta_m(p, q) \|\boldsymbol{\beta}\|_{q^*}$	$p \in \{1, q, \infty\}$
ℓ_p	$\mathcal{U}_{(q,r)}$	$\lambda \delta_m(p, r) \|\boldsymbol{\beta}\|_q$	$p \in \{1, r, \infty\}$
ℓ_p	$\{\boldsymbol{\Delta}: \|\boldsymbol{\delta}_i\|_q \leq \lambda \, \forall i\}$	$\lambda m^{1/p} \|\boldsymbol{\beta}\|_{q^*}$	$p \in \{1, \infty\}$

3. On the equivalence of robustification and regularization in matrix estimation problems

A substantial body of problems at the core of modern developments in statistical estimation involves underlying matrix variables. Two prominent examples which we consider here are matrix completion and Principal Component Analysis (PCA). In both cases we show that a common choice of the regularization problem corresponds exactly to a robustification of the nominal problem subject to uncertainty. In doing so we expand the existing knowledge of robustification for vector regression to a novel and substantial domain. We begin by reviewing these two problem classes before introducing a simple model of uncertainty analogous to the vector model of uncertainty.

3.1. Problem classes

In matrix completion problems one is given data $Y_{ij} \in \mathbb{R}$ for $(i, j) \in E \subseteq \{1, \ldots, m\} \times \{1, \ldots, n\}$. One problem of interest is rank-constrained matrix completion

$$\begin{array}{ll} \min\limits_{\mathbf{X}} & \|\mathbf{Y} - \mathbf{X}\|_{\mathrm{P}_{(F_2)}} \\ \text{s.t.} & \operatorname{rank}(\mathbf{X}) \leq k, \end{array} \tag{7}$$

where $\|\cdot\|_{\mathrm{P}_{(F_2)}}$ denotes the projected 2−Frobenius seminorm, namely,

$$\|\mathbf{Z}\|_{\mathrm{P}_{(F_2)}} = \left(\sum_{(i,j) \in E} Z_{ij}^2 \right)^{1/2}.$$

Matrix completion problems appear in a wide variety of areas. One well-known application is in the Netflix challenge (SIGKDD & Netflix, 2007), where one wishes to predict user movie preferences based on a very limited subset of given user ratings. Here rank-constrained models are important in order to obtain parsimonious descriptions of user preferences in terms of a limited number of significant latent factors. The rank-constrained problem (7) is typically converted to a regularized form with rank replaced by the nuclear norm σ_1 (the sum of singular values) to obtain the convex problem

$$\min_{\mathbf{X}} \|\mathbf{Y} - \mathbf{X}\|_{\mathrm{P}_{(F_2)}} + \lambda \|\mathbf{X}\|_{\sigma_1}.$$

In what follows we show that this regularized problem can be written as an uncertain version of a nominal problem $\min_{\mathbf{X}} \|\mathbf{Y} - \mathbf{X}\|_{\mathrm{P}_{(F_2)}}$.

Similarly to matrix completion, PCA typically takes the form

$$\begin{array}{ll} \min\limits_{\mathbf{X}} & \|\mathbf{Y} - \mathbf{X}\| \\ \text{s.t.} & \operatorname{rank}(\mathbf{X}) \leq k, \end{array} \tag{8}$$

where $\|\cdot\|$ is either the usual Frobenius norm $F_2 = \sigma_2$ or the operator norm σ_∞, and $\mathbf{Y} \in \mathbb{R}^{m \times n}$. PCA arises naturally by assuming that $\mathbf{Y}$ is observed as some low-rank matrix $\mathbf{X}$ plus noise: $\mathbf{Y} = \mathbf{X} + \mathbf{E}$. The solution to (8) is well-known to be a truncated singular value decomposition which retains the k largest singular values (Eckart & Young, 1936). PCA is popular for a variety of applications where dimension reduction is desired.

A variant of PCA known as robust PCA (Candès, Li, Ma, & Wright, 2011) operates under the assumption that some entries of $\mathbf{Y}$ may be grossly corrupted. Robust PCA assumes that $\mathbf{Y} = \mathbf{X} + \mathbf{E}$, where $\mathbf{X}$ is low rank and $\mathbf{E}$ is sparse (few nonzero entries). Under this model robust PCA takes the form

$$\min_{\mathbf{X}} \|\mathbf{Y} - \mathbf{X}\|_{F_1} + \lambda \|\mathbf{X}\|_{\sigma_1}. \tag{9}$$

Here again we can interpret $\|\mathbf{X}\|_{\sigma_1}$ as a surrogate penalty for rank. In the spirit of results from compressed sensing on exact ℓ_1 recovery, it is shown in Candès et al. (2011) that (9) can exactly recover the true $\mathbf{X}_0$ and $\mathbf{E}_0$ assuming that the rank of $\mathbf{X}_0$ is small, $\mathbf{E}_0$ is sufficiently sparse, and the eigenvectors of $\mathbf{X}_0$ are well-behaved (see technical conditions contained therein). Below we derive explicit expressions for PCA subject to certain types of uncertainty; in doing so we show that robust PCA does not correspond to an adversarially robust version of $\min_{\mathbf{X}} \|\mathbf{Y} - \mathbf{X}\|_{\sigma_\infty}$ or $\min_{\mathbf{X}} \|\mathbf{Y} - \mathbf{X}\|_{F_2}$ for *any* model of additive linear uncertainty.

Finally let us note that the results we consider here on robust PCA are distinct from considerations in the robust statistics community on robust approaches to PCA. For results and commentary on such methods, see Croux and Ruiz-Gazen (2005), Hubert, Rousseeuw, and den Branden (2005), Salibian-Barrera, Aelst, and Willems (2005), Hubert et al. (2008).

3.2. Models of uncertainty

For these two problem classes we now detail a model of uncertainty. Our underlying problem is of the form $\min_{\mathbf{X}} \|\mathbf{Y} - \mathbf{X}\|$, where $\mathbf{Y}$ is given data (possibly with some unknown entries). As with the vector case, we do not concern ourselves with uncertainty in the observed $\mathbf{Y}$ because modeling uncertainty in $\mathbf{Y}$ simply leads to a different choice of loss function. To be precise, if $\mathcal{V} \subseteq \mathbb{R}^{m \times n}$ and g is convex loss function then

$$\bar{g}(\mathbf{Y} - \mathbf{X}) := \max_{\boldsymbol{\Delta} \in \mathcal{V}} g((\mathbf{Y} + \boldsymbol{\Delta}) - \mathbf{X})$$

is a new convex loss function $\bar{g}$ of $\mathbf{Y} - \mathbf{X}$.

As in the vector case we assume a linear model of uncertainty in the measurement of $\mathbf{X}$:

$$Y_{ij} = X_{ij} + \left(\sum_{\ell k} \Delta_{\ell k}^{(ij)} X_{\ell k} \right) + \epsilon_{ij},$$

where $\boldsymbol{\Delta}^{(ij)} \in \mathbb{R}^{m \times n}$; alternatively, in inner product notation, $Y_{ij} = X_{ij} + \langle \boldsymbol{\Delta}^{(ij)}, \mathbf{X} \rangle + \epsilon_{ij}$. This linear model is in direct analogy with the model for vector regression taken earlier; now $\boldsymbol{\beta}$ is replaced by $\mathbf{X}$, and again we consider linear perturbations of the unknown regression variable.

This linear model of uncertainty captures a variety of possible forms of uncertainty and accounts for possible interactions among different entries of the matrix $\mathbf{X}$. Note that in matrix notation, the nominal problem becomes, subject to linear uncertainty in $\mathbf{X}$,

$$\min_{\mathbf{X}} \max_{\boldsymbol{\Delta} \in \mathcal{U}} \|\mathbf{Y} - \mathbf{X} - \boldsymbol{\Delta}(\mathbf{X})\|,$$

where here $\mathcal{U}$ is some collection of linear maps and $\boldsymbol{\Delta} \in \mathcal{U}$ is defined as $[\boldsymbol{\Delta}(\mathbf{X})]_{ij} = \langle \boldsymbol{\Delta}^{(ij)}, \mathbf{X} \rangle$, where again $\boldsymbol{\Delta}^{(ij)} \in \mathbb{R}^{m \times n}$ (all linear maps can be written in such a form). Note here the direct analogy to the vector case, with the notation $\boldsymbol{\Delta}(\mathbf{X})$ chosen for simplicity. (For clarity, note that $\boldsymbol{\Delta}$ is not itself a matrix, although one could interpret it as a matrix in $\mathbb{R}^{mn \times mn}$, albeit at a notational cost; we avoid this here.)

We now outline some particular choices for uncertainty sets. As with the vector case, one natural set is an induced uncertainty

set. Precisely, if $g, h : \mathbb{R}^{m\times n} \to \mathbb{R}$ are functions, then we define an induced uncertainty set

$$\mathcal{U}_{(h,g)} := \left\{\mathbf{\Delta} : \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n} | \mathbf{\Delta} \text{ linear}, g(\mathbf{\Delta}(\mathbf{X})) \leq \lambda h(\mathbf{X}) \ \forall \mathbf{X} \in \mathbb{R}^{m\times n}\right\}.$$

As before, when g and h are both norms, $\mathcal{U}_{(h,g)}$ is precisely a ball of radius λ in the induced norm

$$\|\mathbf{\Delta}\|_{(h,g)} = \max_{\mathbf{X}} \frac{g(\mathbf{\Delta}(\mathbf{X}))}{h(\mathbf{X})}.$$

There are also many other possible choices of uncertainty sets. These include the spectral uncertainty sets

$$\mathcal{U}_{\sigma_p} = \{\mathbf{\Delta} : \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n} | \mathbf{\Delta} \text{ linear}, \|\mathbf{\Delta}\|_{\sigma_p} \leq \lambda\},$$

where we interpret $\|\mathbf{\Delta}\|_{\sigma_p}$ as the σ_p norm of $\mathbf{\Delta}$ in any, and hence all, of its matrix representations. Other uncertainty sets are those such as $\mathcal{U} = \{\mathbf{\Delta} : \mathbf{\Delta}^{(ij)} \in \mathcal{U}^{(ij)}\}$, where $\mathcal{U}^{(ij)} \subseteq \mathbb{R}^{m\times n}$ are themselves uncertainty sets. These last two models we will not examine in depth here because they are often subsumed by the vector results (note that these two uncertainty sets do not truly involve the matrix structure of $\mathbf{X}$, and can therefore be "vectorized", reducing directly to vector results).

3.3. Basic results on equivalence

We now continue with some underlying theorems for our models of uncertainty. As a first step, we provide a proposition on the spectral uncertainty sets. As noted above, this result is exactly Theorem 2, and therefore we will not consider such uncertainty sets for the remainder of the paper.

Proposition 5. *For any $q \in [1, \infty]$ and any $\mathbf{Y} \in \mathbb{R}^{m\times n}$,*

$$\min_{\mathbf{X}} \max_{\mathbf{\Delta}\in\mathcal{U}_{\sigma_q}} \|\mathbf{Y} - \mathbf{X} - \mathbf{\Delta}(\mathbf{X})\|_{F_2} = \min_{\mathbf{X}} \|\mathbf{Y} - \mathbf{X}\|_{F_2} + \lambda\|\mathbf{X}\|_{F_2}.$$

For what follows, we restrict our attention to induced uncertainty sets. We begin with an analogous result to Theorem 1. The proof is similar and therefore kept concise. Throughout we always assume without loss of generality that if Y_{ij} is not known then $Y_{ij} = 0$ (i.e., we set it to some arbitrary value).

Theorem 6. *If $g : \mathbb{R}^{m\times n} \to \mathbb{R}$ is a seminorm which is not indentically zero and $h : \mathbb{R}^{m\times n} \to \mathbb{R}$ is a norm, then*

$$\min_{\mathbf{X}} \max_{\mathbf{\Delta}\in\mathcal{U}_{(h,g)}} g(\mathbf{Y} - \mathbf{X} - \mathbf{\Delta}(\mathbf{X})) = \min_{\mathbf{X}} g(\mathbf{Y} - \mathbf{X}) + \lambda h(\mathbf{X}).$$

This theorem leads to an immediate corollary:

Corollary 2. *For any norm $\|\cdot\| : \mathbb{R}^{m\times n} \to \mathbb{R}$ and any $p \in [1, \infty]$*

$$\min_{\mathbf{X}} \max_{\mathbf{\Delta}\in\mathcal{U}_{(\sigma_p,\|\cdot\|)}} \|\mathbf{Y} - \mathbf{X} - \mathbf{\Delta}(\mathbf{X})\| = \min_{\mathbf{X}} \|\mathbf{Y} - \mathbf{X}\| + \lambda\|\mathbf{X}\|_{\sigma_p}.$$

In the two sections which follow we study the implications of Theorem 6 for matrix completion and PCA.

3.4. Robust matrix completion

We now proceed to apply Theorem 6 for the case of matrix completion. Note that the projected Frobenius "norm" $\mathrm{P}(F_2)$ is a seminorm. Therefore, we arrive at the following corollary:

Corollary 3. *For any $p \in [1, \infty]$ one has that*

$$\min_{\mathbf{X}} \max_{\mathbf{\Delta}\in\mathcal{U}_{(\sigma_p,\mathrm{P}_{(F_2)})}} \|\mathbf{Y}-\mathbf{X}-\mathbf{\Delta}(\mathbf{X})\|_{\mathrm{P}_{(F_2)}} = \min_{\mathbf{X}} \|\mathbf{Y}-\mathbf{X}\|_{\mathrm{P}_{(F_2)}} + \lambda\|\mathbf{X}\|_{\sigma_p}.$$

In particular, for $p = 1$ one exactly recovers so-called nuclear norm penalized matrix completion:

$$\min_{\mathbf{X}} \|\mathbf{Y} - \mathbf{X}\|_{\mathrm{P}_{(F_2)}} + \lambda\|\mathbf{X}\|_{\sigma_1}.$$

It is not difficult to show by modifying the proof of Theorem 6 that even though $\mathcal{U}_{(\sigma_p,F_2)} \subsetneq \mathcal{U}_{(\sigma_p,\mathrm{P}(F_2))}$, the following holds:

Proposition 6. *For any $p \in [1, \infty]$ one has that*

$$\min_{\mathbf{X}} \max_{\mathbf{\Delta}\in\mathcal{U}_{(\sigma_p,F_2)}} \|\mathbf{Y} - \mathbf{X} - \mathbf{\Delta}(\mathbf{X})\|_{\mathrm{P}(F_2)} = \min_{\mathbf{X}} \|\mathbf{Y} - \mathbf{X}\|_{\mathrm{P}(F_2)} + \lambda\|\mathbf{X}\|_{\sigma_p}.$$

In particular, for $p = 1$ one exactly recovers nuclear norm penalized matrix completion.

Let us briefly comment on the appearance of the nuclear norm in Corollary 3 and Proposition 6. In light of Remark 1, it is not surprising that such a penalty can be derived by working directly with the rank function (nuclear norm is the convex envelope of the rank function on the ball $\{\mathbf{X} : \|\mathbf{X}\|_{\sigma_\infty} \leq 1\}$, which is why the nuclear norm is typically used to replace rank (Fazel, 2002; Recht et al., 2010). We detail this argument as before. For any $p \in [1, \infty]$ and $\Gamma = \{\mathbf{X} \in \mathbb{R}^{m\times n} : \|\mathbf{X}\|_{\sigma_p} \leq 1\}$, one can show that

$$\mathcal{U}_{(\sigma_1,\mathrm{P}(F_2))} = \left\{\mathbf{\Delta} \text{ linear} : \max_{\mathbf{X}\in\Gamma} \frac{\|\mathbf{\Delta}(\mathbf{X})\|_{\mathrm{P}(F_2)}}{\mathrm{rank}(\mathbf{X})} \leq \lambda\right\}. \tag{10}$$

Therefore, similar to the vector case with an underlying ℓ_0 penalty which becomes a Lasso ℓ_1 penalty, rank leads to the nuclear norm from the robustification setting without directly invoking convexity.

3.5. Robust PCA

We now turn our attention to the implications of Theorem 6 for PCA. We begin by noting robust analogues of $\min_{\mathbf{X}} \|\mathbf{Y} - \mathbf{X}\|$ under the F_2 and σ_∞ norms. This is distinct from the considerations in Caramanis et al. (2011) on robustness of PCA with respect to training and testing sets.

Corollary 4. *For any $p \in [1, \infty]$ one has that*

$$\min_{\mathbf{X}} \max_{\mathbf{\Delta}\in\mathcal{U}_{(\sigma_p,F_2)}} \|\mathbf{Y} - \mathbf{X} - \mathbf{\Delta}(\mathbf{X})\|_{F_2} = \min_{\mathbf{X}} \|\mathbf{Y} - \mathbf{X}\|_{F_2} + \lambda\|\mathbf{X}\|_{\sigma_p}$$

and

$$\min_{\mathbf{X}} \max_{\mathbf{\Delta}\in\mathcal{U}_{(\sigma_p,\sigma_\infty)}} \|\mathbf{Y} - \mathbf{X} - \mathbf{\Delta}(\mathbf{X})\|_{\sigma_\infty} = \min_{\mathbf{X}} \|\mathbf{Y} - \mathbf{X}\|_{\sigma_\infty} + \lambda\|\mathbf{X}\|_{\sigma_p}.$$

We continue by considering robust PCA as presented in Candès et al. (2011). Suppose that $\mathcal{U}$ is some collection of linear maps $\mathbf{\Delta} : \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}$ and $\|\cdot\|$ is some norm so that for any $\mathbf{Y}, \mathbf{X} \in \mathbb{R}^{m\times n}$

$$\max_{\mathbf{\Delta}\in\mathcal{U}} \|\mathbf{Y} - \mathbf{X} - \mathbf{\Delta}(\mathbf{X})\| = \|\mathbf{Y} - \mathbf{X}\|_{F_1} + \lambda\|\mathbf{X}\|_{\sigma_1}.$$

It is easy to see that this implies $\|\cdot\| = \|\cdot\|_{F_1}$. These observations, combined with Theorem 6, imply the following:

Proposition 7. *The problem (9) can be written as an uncertain version of $\min_{\mathbf{X}} \|\mathbf{Y} - \mathbf{X}\|$ subject to additive, linear uncertainty in $\mathbf{X}$ if and only if $\|\cdot\|$ is the 1-Frobenius norm F_1. In particular, (9) does not arise as uncertain versions of PCA (using F_2 or σ_∞) under such a model of uncertainty.*

This result is not entirely surprising. This is because robust PCA attempts to solve, based on its model of $\mathbf{Y} = \mathbf{X} + \mathbf{E}$ where $\mathbf{X}$ is low-rank and $\mathbf{E}$ is sparse, a problem of the form

$$\min_{\mathbf{X}} \|\mathbf{Y} - \mathbf{X}\|_{F_0} + \lambda\,\mathrm{rank}(\mathbf{X}),$$

where $\|\mathbf{A}\|_{F_0}$ is the number of nonzero entries of $\mathbf{A}$. In the usual way, F_0 and rank are replaced with surrogates F_1 and σ_1, respectively. Hence, (9) appears as a convex, regularized form of the problem

$$\begin{aligned} \min_{\mathbf{X}} \quad & \|\mathbf{Y} - \mathbf{X}\|_{F_1} \\ \text{s.t.} \quad & \mathrm{rank}(\mathbf{X}) \leq k. \end{aligned}$$

Again, as with matrix completion, it is possible to show that (9) and uncertain forms of PCA with a nuclear norm penalty (as appearing in Corollary 4) can be derived using the true choice of penalizer, rank, instead of imposing an *a priori* assumption of a nuclear norm penalty. We summarize this, without proof, as follows:

Proposition 8. *For any $p \in [1, \infty]$ and any norm $\|\cdot\|$,*

$$\min_{\mathbf{X}\in\Gamma} \max_{\mathbf{\Delta}\in\mathcal{U}_{\Gamma(\text{rank},\|\cdot\|)}} \|\mathbf{Y}-\mathbf{X}-\mathbf{\Delta}(\mathbf{X})\| = \min_{\mathbf{X}\in\Gamma} \|\mathbf{Y}-\mathbf{X}\| + \lambda\|\mathbf{X}\|_{\sigma_1},$$

where $\Gamma = \{\mathbf{X} \in \mathbb{R}^{m\times n} : \|\mathbf{X}\|_{\sigma_p} \leq 1\}$ and

$$\mathcal{U}_{\Gamma(\text{rank},\|\cdot\|)} = \left\{ \mathbf{\Delta} \text{ linear} : \max_{\mathbf{X}\in\Gamma} \frac{\|\mathbf{\Delta}(\mathbf{X})\|}{\text{rank}(\mathbf{X})} \leq \lambda \right\}.$$

3.6. Non-equivalence of robustification and regularization

As with vector regression it is not always the case that robustification is equivalent to regularization in matrix estimation problems. For completeness we provide analogues here of the linear regression results. We begin by stating results which follow over with essentially identical proofs from the vector case; proofs are not included here. Then we characterize precisely when another plausible model of uncertainty leads to equivalence.

We begin with the analogue of Proposition 2.

Proposition 9. *Let $\mathcal{U} \subseteq \{\text{linear maps } \mathbf{\Delta} : \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}\}$ be any non-empty, compact set and $g : \mathbb{R}^{m\times n} \to \mathbb{R}$ a seminorm. Then there exists some seminorm $\overline{h} : \mathbb{R}^{m\times n} \to \mathbb{R}$ so that for any $\mathbf{Z}, \mathbf{X} \in \mathbb{R}^{m\times n}$,*

$$\max_{\mathbf{\Delta}\in\mathcal{U}} g(\mathbf{Z}+\mathbf{\Delta}(\mathbf{X})) \leq g(\mathbf{Z}) + \overline{h}(\mathbf{X}),$$

with equality when $\mathbf{Z} = \mathbf{0}$.

As before with Theorem 4 and Propositions 3 and 4, one can now compute $\overline{h}$ for a variety of problems.

Proposition 10. *For any $\mathbf{Z}, \mathbf{X} \in \mathbb{R}^{m\times n}$,*

$$\|\mathbf{Z}\|_{F_p} + \frac{\lambda}{\delta_{mn}(q,p)}\|\mathbf{X}\|_{F_{q^*}} \leq \max_{\mathbf{\Delta}\in\mathcal{U}_{F_q}} \|\mathbf{Z}+\mathbf{\Delta}(\mathbf{X})\|_{F_p} \tag{11}$$

$$\leq \|\mathbf{Z}\|_{F_p} + \lambda\delta_{mn}(p,q)\|\mathbf{X}\|_{F_{q^*}} \tag{12}$$

where $\|\mathbf{\Delta}\|_{F_q}$ is interpreted as the F_q norm on the matrix representation of $\mathbf{\Delta}$ in the standard basis. In particular, if $p \neq q$ and $p \in (1, \infty)$, then for any $\mathbf{X} \neq \mathbf{0}$ the upper bound in (12) is strict for almost all $\mathbf{Z}$ (so long as $mn \geq 2$). Further, when $p \neq q$ and $p \in (1, \infty)$, the gap in the lower bound in (11) is arbitrarily small for all $\mathbf{X}$.

Proposition 11. *For any $\mathbf{Z}, \mathbf{X} \in \mathbb{R}^{m\times n}$,*

$$\|\mathbf{Z}\|_p + \frac{\lambda}{\delta_{mn}(2,p)}\|\mathbf{X}\|_{F_2} \leq \max_{\mathbf{\Delta}\in\mathcal{U}_{\sigma_q}} \|\mathbf{Z}+\mathbf{\Delta}(\mathbf{X})\|_{F_p} \tag{13}$$

$$\leq \|\mathbf{Z}\|_{F_p} + \lambda\delta_{mn}(p,2)\|\mathbf{X}\|_{F_2}. \tag{14}$$

In particular, if $p \notin \{1, 2, \infty\}$, then for all $\mathbf{X} \neq \mathbf{0}$ the upper bound in (14) is strict for almost all $\mathbf{Z}$ (so long as $mn \geq 2$). Further, if $p \notin \{1, 2, \infty\}$, the gap in the lower bound in (13) is arbitrarily small for all $\mathbf{X}$.

We now turn our attention to non-equivalencies which may arise under different models of uncertainty instead of the general matrix model of linear uncertainty which we have included here, where

$$[\mathbf{\Delta}(\mathbf{X})]_{ij} = \sum_{\ell k} \Delta^{(ij)}_{\ell k} X_{\ell k} = \langle \mathbf{\Delta}^{(ij)}, \mathbf{X} \rangle,$$

with $\mathbf{\Delta}^{(ij)} \in \mathbb{R}^{m\times n}$. Another plausible model of uncertainty is one for which the jth column of $\mathbf{\Delta}(\mathbf{X})$ only depends on $\mathbf{X}_j$, the jth column of $\mathbf{X}$ (or, for example, with columns replaced by rows). We now examine such a model. In this setup, we now have n matrices $\mathbf{\Delta}^{(j)} \in \mathbb{R}^{m\times m}$ and we define the linear map $\mathbf{\Delta}$ so that the jth column of $\mathbf{\Delta}(\mathbf{X}) \in \mathbb{R}^{m\times n}$, denoted $[\mathbf{\Delta}(\mathbf{X})]_j$, is $[\mathbf{\Delta}(\mathbf{X})]_j := \mathbf{\Delta}^{(j)}\mathbf{X}_j$, which is simply matrix vector multiplication. Therefore,

$$\mathbf{\Delta}(\mathbf{X}) = \begin{bmatrix} \mathbf{\Delta}^{(1)}\mathbf{X}_1 & \cdots & \mathbf{\Delta}^{(n)}\mathbf{X}_n \end{bmatrix}. \tag{15}$$

For an example of where such a model of uncertainty may arise, we consider matrix completion in the context of the Netflix problem. If one treats $\mathbf{X}_j$ as user j's true ratings, then such a model addresses uncertainty within a given user's ratings, while not allowing uncertainty to have cross-user effects. This model of uncertainty does not rely on true matrix structure and therefore reduces to earlier results on non-equivalence in vector regression. As an example of such a reduction, we state the following proposition characterizing equivalence. Again, this is a direct modification of Theorem 4 and the proof we do not include here.

Proposition 12. *For the model of uncertainty in (15) with $\mathbf{\Delta}^{(j)} \in \mathcal{U}_{F_{q_j}}$ for $j = 1, \ldots, n$, where $q_j \in [1, \infty]$, one has for the problem $\min_{\mathbf{X}} \max_{\mathbf{\Delta}\in\mathcal{U}} \|\mathbf{Y}-\mathbf{X}-\mathbf{\Delta}(\mathbf{X})\|_{F_p}$ that $\overline{h}$ is defined as*

$$\overline{h}(\mathbf{X}) = \lambda\left(\sum_j \delta_m^p(p,q_j)\|\mathbf{X}_j\|_{q_j^*}^p\right)^{1/p}. \tag{16}$$

Further, under such a model of uncertainty, robustification is equivalent to regularization with $\overline{h}$ if and only if $p \in \{1, \infty\}$ or $p = q_j$ for all $j = 1, \ldots, n$.

While the case of matrix regression offers a large variety of possible models of uncertainty, we see again as with vector regression that this variety inevitably leads to scenarios in which robustification is no longer directly equivalent to regularization. We summarize the conclusions of this section in Table 3.

Table 3
Summary of equivalencies for robustification with uncertainty set $\mathcal{U}$ and regularization with penalty $\overline{h}$, where $\overline{h}$ is as given in Proposition 9. Here by equivalence we mean that for all $\mathbf{Z}, \mathbf{X} \in \mathbb{R}^{m\times n}$, $\max_{\mathbf{\Delta}\in\mathcal{U}} g(\mathbf{Z}+\mathbf{X}) = g(\mathbf{Z}) + \overline{h}(\mathbf{X})$, where g is the loss function, i.e., the upper bound $\overline{h}$ is also a lower bound. Here δ_{mn} is as in Theorem 4. Throughout $p, q \in [1, \infty]$ and $mn \geq 2$.

Loss function seminorm g	Uncertainty set $\mathcal{U}_{(h,g)}$ (h norm)	$\overline{h}(\mathbf{X})$ $\lambda h(\mathbf{X})$	Equivalence if and only if always
F_p	$\mathcal{U}_{\sigma_q}$	$\lambda\delta_{mn}(p,2)\|\mathbf{X}\|_{F_2}$	$p \in \{1, 2, \infty\}$
F_p	$\mathcal{U}_{F_q}$	$\lambda\delta_{mn}(p,q)\|\mathbf{X}\|_{F_{q^*}}$	$p \in \{1, q, \infty\}$
F_p	$\mathcal{U}$ in (15) with $\mathbf{\Delta}^{(j)} \in \mathcal{U}_{F_{q_j}}$	(16)	$(p = q_j\ \forall j)$ or $p \in \{1, \infty\}$

4. Conclusion

In this work we have considered the robustification of a variety of problems from classical and modern statistical regression as subject to data uncertainty. We have taken care to emphasize that there is a fine line between this process of robustification and the usual process of regularization, and that the two are not always directly equivalent. While deepening this understanding we have also extended this connection to new domains, such as in matrix completion and PCA. In doing so, we have shown that the usual regularization approaches to modern statistical regression do not always coincide with an adversarial approach motivated by robust optimization.

Acknowledgments

We thank the reviewer for their comments that helped us improve the paper.

Appendix A.

This appendix contains proofs and additional technical results for the vector regression setting. We prove our results in the vector setting, from which the primary results on matrices follow as a direct corollary.

Proof of Theorem 4.

(a) We begin by proving the upper bound. Here we proceed by showing that $\bar{h}$ above is precisely $\bar{h}(\boldsymbol{\beta}) = \lambda\delta_m(p,q)\|\boldsymbol{\beta}\|_{q^*}$. Now observe that for any $\boldsymbol{\Delta} \in \mathcal{U}_{F_q}$,

$$\begin{aligned}\|\boldsymbol{\Delta}\boldsymbol{\beta}\|_p &\leq \delta_m(p,q)\|\boldsymbol{\Delta}\boldsymbol{\beta}\|_q \leq \delta_m(p,q)\|\boldsymbol{\Delta}\|_{F_q}\|\boldsymbol{\beta}\|_{q^*} \\ &\leq \delta_m(p,q)\lambda\|\boldsymbol{\beta}\|_{q^*}. \end{aligned} \tag{17}$$

The first inequality follows by the definition of the discrepancy function δ_m. The second inequality follows from a well-known matrix inequality: $\|\boldsymbol{\Delta}\boldsymbol{\beta}\|_q \leq \|\boldsymbol{\Delta}\|_{F_q}\|\boldsymbol{\beta}\|_{q^*}$ (this follows from a simple application of Hölder's inequality). Now observe that in the chain of inequalities in (17), if one takes any $\mathbf{u} \in \operatorname{argmax}\delta_m(p,q)$ and any $\mathbf{v} \in \operatorname{argmax}_{\|\mathbf{v}\|_q=1}\mathbf{v}'\boldsymbol{\beta}$, then $\widehat{\boldsymbol{\Delta}} := \lambda\mathbf{u}\mathbf{v}' \in \mathcal{U}_{F_q}$ and $\|\widehat{\boldsymbol{\Delta}}\boldsymbol{\beta}\|_p = \delta_m(p,q)\lambda\|\boldsymbol{\beta}\|_{q^*}$. Hence, $\bar{h}(\boldsymbol{\beta}) = \delta_m(p,q)\lambda\|\boldsymbol{\beta}\|_{q^*}$. This proves the upper bound.

(b) We now prove that for $p \in \{1, \infty\}$ that one has equality for all $(\mathbf{z}, \boldsymbol{\beta}) \in \mathbb{R}^m \times \mathbb{R}^n$. This follows an argument similar to that needed for Theorem 6. First consider the case when $p = 1$. Fix $\mathbf{z} \in \mathbb{R}^m$. Again let $\mathbf{u} \in \operatorname{argmax}\delta_m(1,q)$ and $\mathbf{v} \in \operatorname{argmax}_{\|\mathbf{v}\|_q=1}\mathbf{v}'\boldsymbol{\beta}$. Without loss of generality we may assume that $\operatorname{sign}(z_i) = \operatorname{sign}(u_i)$ for $i = 1, \ldots, m$ (one may change the sign of entries of $\mathbf{u}$ and it is still in $\operatorname{argmax}\delta_m(1,q)$). Then again we have $\widehat{\boldsymbol{\Delta}} := \lambda\mathbf{u}\mathbf{v}' \in \mathcal{U}_{F_q}$ and

$$\begin{aligned}\|\mathbf{z} + \widehat{\boldsymbol{\Delta}}\boldsymbol{\beta}\|_1 &= \|\mathbf{z} + \lambda\mathbf{u}\mathbf{v}'\boldsymbol{\beta}\|_1 = \|\mathbf{z} + \lambda\|\boldsymbol{\beta}\|_{q^*}\mathbf{u}\|_1 \\ &= \|\mathbf{z}\|_1 + \lambda\|\boldsymbol{\beta}\|_{q^*}\|\mathbf{u}\|_1 = \|\mathbf{z}\|_1 + \lambda\|\boldsymbol{\beta}\|_{q^*}\delta_m(1,q).\end{aligned}$$

Hence, one has equality in the upper bound for $p = 1$, as claimed.

We now turn our attention to the case $p = \infty$. Note that $\delta_m(\infty, q) = 1$ because $\|\mathbf{z}\|_\infty \leq \|\mathbf{z}\|_q$ for all $\mathbf{z} \in \mathbb{R}^m$. Fix $\mathbf{z} \in \mathbb{R}^m$, and again let $\mathbf{v} \in \operatorname{argmax}_{\|\mathbf{v}\|_q=1}\mathbf{v}'\boldsymbol{\beta}$. Let $\ell \in \{1, \ldots, m\}$ so that $|z_\ell| = \|\mathbf{z}\|_\infty$. Define $\mathbf{u} = \operatorname{sign}(z_\ell)\mathbf{e}_\ell \in \mathbb{R}^m$, where $\mathbf{e}_\ell$ is the vector whose only nonzero entry is a 1 in the ℓth position. Now observe that $\widehat{\boldsymbol{\Delta}} := \lambda\mathbf{u}\mathbf{v}' \in \mathcal{U}_{F_q}$ and

$$\begin{aligned}\|\mathbf{z} + \widehat{\boldsymbol{\Delta}}\boldsymbol{\beta}\|_\infty &= \|\mathbf{z} + \operatorname{sign}(z_\ell)\lambda\|\boldsymbol{\beta}\|_{q^*}\mathbf{e}_\ell\|_\infty \\ &= \|\mathbf{z}\|_\infty + \lambda\|\boldsymbol{\beta}\|_{q^*}\|\mathbf{e}_\ell\|_\infty = \|\mathbf{z}\|_\infty + \lambda\|\boldsymbol{\beta}\|_{q^*},\end{aligned}$$

which proves equality in (3), as was to be shown.

(c) To proceed, we examine the case where $p \in (1, \infty)$ and consider for which $(\mathbf{z}, \boldsymbol{\beta})$ the inequality in (3) is strict. Fix $\boldsymbol{\beta} \neq \mathbf{0}$. For $p \in (1, \infty)$ and $\mathbf{y}, \mathbf{z} \in \mathbb{R}^m$, one has by Minkowski's inequality that $\|\mathbf{y} + \mathbf{z}\|_p = \|\mathbf{y}\|_p + \|\mathbf{z}\|_p$ if and only if one of $\mathbf{y}$ or $\mathbf{z}$ is a non-negative scalar multiple of the other. To have equality in (3), it must be that there exists some $\boldsymbol{\Delta} \in \operatorname{argmax}_{\boldsymbol{\Delta}\in\mathcal{U}_{F_q}}\|\boldsymbol{\Delta}\boldsymbol{\beta}\|_p$ for which $\|\mathbf{z} + \boldsymbol{\Delta}\boldsymbol{\beta}\|_p = \|\mathbf{z}\|_p + \|\boldsymbol{\Delta}\boldsymbol{\beta}\|_p$. For any $\mathbf{z} \neq \mathbf{0}$ this observation, combined with Minkowski's inequality, implies that

$$\|\boldsymbol{\Delta}\|_{F_q} = \lambda, \quad \boldsymbol{\Delta}\boldsymbol{\beta} = \mu\mathbf{z} \text{ for some } \mu \geq 0, \quad \text{and}$$
$$\|\boldsymbol{\Delta}\boldsymbol{\beta}\|_p = \lambda\delta_m(p,q)\|\boldsymbol{\beta}\|_{q^*}.$$

The first and last equalities imply that $\boldsymbol{\Delta}\boldsymbol{\beta} \in \lambda\|\boldsymbol{\beta}\|_{q^*}\operatorname{argmax}\delta_m(p,q)$. Note that $\operatorname{argmax}\delta_m(p,q)$ is finite whenever $p \neq q$ and $m \geq 2$, a geometric property of ℓ_p balls. Hence, taking any $\mathbf{z}$ which is not a scalar multiple of a point in $\operatorname{argmax}\delta_m(p,q)$ implies by Minkowski's inequality that

$$\max_{\boldsymbol{\Delta}\in\mathcal{U}_{F_q}}\|\mathbf{z} + \boldsymbol{\Delta}\boldsymbol{\beta}\|_p < \|\mathbf{z}\|_p + \lambda\delta_m(p,q)\|\boldsymbol{\beta}\|_{q^*}.$$

Hence, for any $\boldsymbol{\beta} \neq \mathbf{0}$, the inequality in (3) is strict for all $\mathbf{z}$ not in a finite union of one-dimensional subspaces, so long as $p \in (1, \infty)$, $p \neq q$, and $m \geq 2$.

(d) We now prove the lower bound in (4). If $\mathbf{z} = \mathbf{0}$ then there is nothing to show, and therefore we assume $\mathbf{z} \neq \mathbf{0}$. Let $\mathbf{v} \in \mathbb{R}^n$ so that

$$\mathbf{v} \in \operatorname{argmax}_{\|\mathbf{v}\|_q=1}\mathbf{v}'\boldsymbol{\beta}.$$

Hence $\mathbf{v}'\boldsymbol{\beta} = \|\boldsymbol{\beta}\|_{q^*}$ by the definition of the dual norm. Define $\widehat{\boldsymbol{\Delta}} = \frac{\lambda}{\|\mathbf{z}\|_q}\mathbf{z}\mathbf{v}'$. Observe that $\widehat{\boldsymbol{\Delta}} \in \mathcal{U}_{F_q}$. Further, note that $\|\mathbf{z}\|_q \leq \delta_m(q,p)\|\mathbf{z}\|_p$ by definition of δ_m and therefore $1/\delta_m(q,p) \leq \|\mathbf{z}\|_p/\|\mathbf{z}\|_q$. Putting things together,

$$\begin{aligned}\|\mathbf{z}\|_p + \frac{\lambda\|\boldsymbol{\beta}\|_{q^*}}{\delta_m(q,p)} &\leq \|\mathbf{z}\|_p + \frac{\lambda\|\mathbf{z}\|_p\|\boldsymbol{\beta}\|_{q^*}}{\|\mathbf{z}\|_q} \\ &= \|\mathbf{z}\|_p\left(1 + \frac{\lambda\|\boldsymbol{\beta}\|_{q^*}}{\|\mathbf{z}\|_q}\right) = \|\mathbf{z} + \widehat{\boldsymbol{\Delta}}\boldsymbol{\beta}\|_p \\ &\leq \max_{\boldsymbol{\Delta}\in\mathcal{U}_{F_q}}\|\mathbf{z} + \boldsymbol{\Delta}\boldsymbol{\beta}\|_p.\end{aligned}$$

This completes the proof of the lower bound.

(e) To conclude we prove that the gap in (4) can be made arbitrarily small for $p \in (1, \infty)$. We proceed in several steps. We first prove that for any $\mathbf{z} \neq \mathbf{0}$ that

$$\lim_{\alpha\to\infty}\left(\max_{\boldsymbol{\Delta}\in\mathcal{U}_{F_q}}\|\alpha\mathbf{z} + \boldsymbol{\Delta}\boldsymbol{\beta}\|_p - \|\alpha\mathbf{z}\|_p\right) = \frac{\lambda\|\boldsymbol{\beta}\|_{q^*}\|\mathbf{z}^{p-1}\|_{q^*}}{\|\mathbf{z}\|_p^{p-1}}, \tag{18}$$

where we use the shorthand $\mathbf{z}^{p-1}$ to denote the vector in $\mathbb{R}^m$ whose ith entry is $|z_i|^{p-1}$. Observe that

$$\max_{\boldsymbol{\Delta}\in\mathcal{U}_{F_q}}\|\alpha\mathbf{z} + \boldsymbol{\Delta}\boldsymbol{\beta}\|_p = \max_{\|\mathbf{u}\|_q\leq\lambda\|\boldsymbol{\beta}\|_{q^*}}\|\alpha\mathbf{z} + \mathbf{u}\|_p.$$

It is easy to argue that we may assume without any loss of generality that $\mathbf{u} \in \operatorname{argmax}_{\|\mathbf{u}\|_q\leq\lambda\|\boldsymbol{\beta}\|_{q^*}}\|\alpha\mathbf{z} + \mathbf{u}\|_p$ has $\operatorname{sign}(u_i) = \operatorname{sign}(\alpha z_i)$, where

$$\operatorname{sign}(a) = \begin{cases}1, & a \geq 0 \\ -1, & a < 0.\end{cases}$$

Therefore, we restrict our attention to $\mathbf{z} \geq \mathbf{0}$, $\mathbf{z} \neq \mathbf{0}$, and $\mathbf{u} \geq \mathbf{0}$. For any $\mathbf{u}$ such that $\|\mathbf{u}\|_q \leq \lambda\|\boldsymbol{\beta}\|_{q^*}$ and $\mathbf{u} \geq \mathbf{0}$, note that

$$\begin{aligned}\lim_{\alpha\to\infty}\|\alpha\mathbf{z} + \mathbf{u}\|_p - \|\alpha\mathbf{z}\|_p &= \lim_{\alpha\to\infty}\frac{\|\mathbf{z} + \mathbf{u}/\alpha\|_p - \|\mathbf{z}\|_p}{1/\alpha} \\ &= \lim_{\bar{\alpha}\to 0^+}\frac{\|\mathbf{z} + \bar{\alpha}\mathbf{u}\|_p - \|\mathbf{z}\|_p}{\bar{\alpha}} \\ &= \frac{d}{d\bar{\alpha}}\bigg|_{\bar{\alpha}=0}\|\mathbf{z} + \bar{\alpha}\mathbf{u}\|_p = \frac{\mathbf{u}'\mathbf{z}^{p-1}}{\|\mathbf{z}\|_p^{p-1}}.\end{aligned}$$

We can now proceed to finish the claim in (18) (still restricting attention to $\mathbf{z} \geq \mathbf{0}$ without loss of generality). By the above arguments, for any $\mathbf{u} \geq \mathbf{0}$ and any $\epsilon > 0$ there exists some $\hat{\alpha} = \hat{\alpha}(\mathbf{u}) > 0$ sufficiently large so that for all $\alpha > \hat{\alpha}$,

$$\left|\|\alpha\mathbf{z} + \mathbf{u}\|_p - \|\alpha\mathbf{z}\|_p - \frac{\mathbf{u}'\mathbf{z}^{p-1}}{\|\mathbf{z}\|_p^{p-1}}\right| \leq \epsilon.$$

It remains to be shown that for any $\epsilon > 0$ there exists some $\hat{\alpha}$ so that for all $\alpha > \hat{\alpha}$,

$$\left|\left(\max_{\|\mathbf{u}\|_q\leq\lambda\|\boldsymbol{\beta}\|_{q^*}}\|\alpha\mathbf{z} + \mathbf{u}\|_p - \|\alpha\mathbf{z}\|_p\right) - \left(\max_{\|\mathbf{u}\|_q\leq\lambda\|\boldsymbol{\beta}\|_{q^*}}\frac{\mathbf{u}'\mathbf{z}^{p-1}}{\|\mathbf{z}\|_p^{p-1}}\right)\right| \leq \epsilon.$$

We prove this as follows. Let $\epsilon > 0$. Choose points $\{\mathbf{u}_1, \ldots, \mathbf{u}_M\} \subseteq \mathbb{R}^m$ with $\|\mathbf{u}_j\|_q = \lambda\|\boldsymbol{\beta}\|_{q^*}$ $\forall j$ so that for any $\mathbf{u} \in \mathbb{R}^m$ with $\|\mathbf{u}\|_q = \lambda\|\boldsymbol{\beta}\|_{q^*}$, there exists some j so that $\|\mathbf{u} - \mathbf{u}_j\|_p \leq \epsilon/3$ (note that our choice of ℓ_p here is intentional). Now observe that for any α,

$$\begin{aligned}
\max_j \|\alpha\mathbf{z} + \mathbf{u}_j\|_p &\leq \max_{\|\mathbf{u}\|_q \leq \lambda\|\boldsymbol{\beta}\|_{q^*}} \|\alpha\mathbf{z} + \mathbf{u}\|_p \\
&\leq \max_j \left(\max_{\|\mathbf{u}-\mathbf{u}_j\|_p \leq \epsilon/3} \|\alpha\mathbf{z} + \mathbf{u}\|_p \right) \\
&= \max_j \left(\max_{\|\bar{\mathbf{u}}\|_p \leq \epsilon/3} \|\alpha\mathbf{z} + \mathbf{u}_j + \bar{\mathbf{u}}\|_p \right) \\
&\leq \max_j \left(\max_{\|\bar{\mathbf{u}}\|_p \leq \epsilon/3} \|\alpha\mathbf{z} + \mathbf{u}_j\|_p + \|\bar{\mathbf{u}}\|_p \right) \\
&= \epsilon/3 + \max_j \|\alpha\mathbf{z} + \mathbf{u}_j\|_p.
\end{aligned}$$

Similarly, one has for $\bar{\mathbf{z}} = \mathbf{z}^{p-1}/\|\mathbf{z}\|_p^{p-1}$ that $|\max_j \mathbf{u}_j'\bar{\mathbf{z}} - \max_{\|\mathbf{u}\|_q \leq \lambda\|\boldsymbol{\beta}\|_{q^*}} \mathbf{u}'\bar{\mathbf{z}}| \leq \epsilon/3$. (This uses the fact that $\|\bar{\mathbf{z}}\|_{p^*} = 1$.) Now for each j choose $\hat{\alpha}_j$ so that for all $\alpha > \hat{\alpha}_j$,

$$\left| \|\alpha\mathbf{z} + \mathbf{u}_j\|_p - \|\alpha\mathbf{z}\|_p - \mathbf{u}_j'\bar{\mathbf{z}} \right| \leq \epsilon/3.$$

Define $\hat{\alpha} = \max_j \hat{\alpha}_j$. Now observe that by combining the above two observations, one has for any $\alpha > \hat{\alpha}$ that

$$\begin{aligned}
&\left| \left(\max_{\|\mathbf{u}\|_q \leq \lambda\|\boldsymbol{\beta}\|_{q^*}} \|\alpha\mathbf{z} + \mathbf{u}\|_p - \|\alpha\mathbf{z}\|_p \right) - \left(\max_{\|\mathbf{u}\|_q \leq \lambda\|\boldsymbol{\beta}\|_{q^*}} \mathbf{u}'\bar{\mathbf{z}} \right) \right| \\
&\leq 2\epsilon/3 + \left| \left(\max_j \|\alpha\mathbf{z} + \mathbf{u}_j\|_p - \|\alpha\mathbf{z}\|_p \right) - \left(\max_\ell \mathbf{u}_\ell'\bar{\mathbf{z}} \right) \right| \\
&\leq 2\epsilon/3 + \max_j \left| \|\alpha\mathbf{z} + \mathbf{u}_j\|_p - \|\alpha\mathbf{z}\|_p - \mathbf{u}_j'\bar{\mathbf{z}} \right| \\
&\leq 2\epsilon/3 + \epsilon/3 = \epsilon.
\end{aligned}$$

Noting that $\max_{\|\mathbf{u}\|_q \leq \lambda\|\boldsymbol{\beta}\|_{q^*}} \mathbf{u}'\bar{\mathbf{z}} = \lambda\|\boldsymbol{\beta}\|_{q^*}\|\bar{\mathbf{z}}\|_{q^*}$ concludes the proof of (18). We now claim that

$$\min_{\mathbf{z}} \frac{\|\mathbf{z}^{p-1}\|_{q^*}}{\|\mathbf{z}\|_p^{p-1}} = \frac{1}{\delta_m(q,p)}. \tag{19}$$

First note that

$$\min_{\mathbf{z}} \frac{\|\mathbf{z}^{p-1}\|_{q^*}}{\|\mathbf{z}\|_p^{p-1}} = \min_{\tilde{\mathbf{z}}} \frac{\|\tilde{\mathbf{z}}\|_{q^*}}{\|\tilde{\mathbf{z}}\|_{p^*}}. \tag{20}$$

We prove this as follows: given $\mathbf{z}$, let $\tilde{\mathbf{z}} = \mathbf{z}^{p-1}$. Then one can show that $\|\tilde{\mathbf{z}}\|_{p^*}/\|\mathbf{z}\|_p^{p-1} = 1$, and so $\|\tilde{\mathbf{z}}\|_{p^*}/\|\tilde{\mathbf{z}}\|_{q^*} = \|\mathbf{z}\|_p^{p-1}/\|\mathbf{z}^{p-1}\|_{q^*}$. The converse is similar, proving (20). Finally, note that

$$\min_{\tilde{\mathbf{z}}} \frac{\|\tilde{\mathbf{z}}\|_{q^*}}{\|\tilde{\mathbf{z}}\|_{p^*}} = \frac{1}{\delta_m(p^*, q^*)}$$

which follows from an elementary analysis using the definition of δ_m. Combined with the observation that $\delta_m(p^*, q^*) = \delta_m(q, p)$, which follows by a simply duality argument (or by inspecting the formula), we have that (19) is proven. To finish the argument, pick any $\mathbf{z} \in \operatorname{argmin}_{\mathbf{z}} \|\mathbf{z}^{p-1}\|_{q^*}/\|\mathbf{z}\|_p^{p-1}$. Per (19), $\|\mathbf{z}^{p-1}\|_{q^*}/\|\mathbf{z}\|_p^{p-1} = 1/\delta_m(q,p)$. Hence, now applying (18), given any $\epsilon > 0$, there exists some $\alpha > 0$ large enough so that

$$\left| \left(\max_{\mathbf{\Delta} \in \mathcal{U}_{F_q}} \|\alpha\mathbf{z} + \mathbf{\Delta}\boldsymbol{\beta}\|_p \right) - \left(\|\alpha\mathbf{z}\|_p + \frac{\lambda}{\delta_m(q,p)}\|\boldsymbol{\beta}\|_{q^*} \right) \right| \leq \epsilon.$$

Therefore, the gap in the lower bound in (4) can be made arbitrarily small for any $\boldsymbol{\beta} \in \mathbb{R}^n$. This concludes the proof. □

Appendix B.

This appendix includes an example of choice of loss function and uncertainty set under which (a) regularization is not equivalent to robustification in general and (b) there exist problem instances for which the regularization path and robustification path are different. The example we give is in the vector setting for simplicity, although the generalization to matrices is obvious.

In particular, let $m = 2$ and $n = 2$, and consider $\mathcal{U} = \mathcal{U}_{(1,1)}$ and loss function ℓ_2, with $\mathbf{y} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$ and $\mathbf{X} = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}$. In symbols, the problem of interest is

$$\min_{\boldsymbol{\beta}} \max_{\mathbf{\Delta} \in \mathcal{U}_{(1,1)}} \|\mathbf{y} - (\mathbf{X} + \mathbf{\Delta})\boldsymbol{\beta}\|_2. \tag{B.1}$$

For fixed $\boldsymbol{\beta}$, the objective can be rewritten exactly as

$$\begin{aligned}
&\max_{\mathbf{\Delta} \in \mathcal{U}_{(1,1)}} \|\mathbf{y} - (\mathbf{X} + \mathbf{\Delta})\boldsymbol{\beta}\|_2 \\
&= \max_{\substack{\mathbf{u}: \\ \|\mathbf{u}\|_1 \leq \lambda\|\boldsymbol{\beta}\|_1}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} + \mathbf{u}\|_2 \\
&= \max \left\{ \left\| \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \pm \begin{pmatrix} \lambda\|\boldsymbol{\beta}\|_1 \\ 0 \end{pmatrix} \right\|_2, \left\| \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \pm \begin{pmatrix} 0 \\ \lambda\|\boldsymbol{\beta}\|_1 \end{pmatrix} \right\|_2 \right\} \\
&= \max \left\{ \left\| \mathbf{y} - \left(\mathbf{X} + \begin{pmatrix} \pm\lambda & \pm\lambda \\ 0 & 0 \end{pmatrix} \right)\boldsymbol{\beta} \right\|_2, \|\mathbf{y} \right. \\
&\quad \left. - \left(\mathbf{X} + \begin{pmatrix} 0 & 0 \\ \pm\lambda & \pm\lambda \end{pmatrix} \right)\boldsymbol{\beta} \right\|_2 \Bigg\} \\
&= \max_{\mathbf{S} \in \mathcal{S}} \|\mathbf{y} - (\mathbf{X} + \mathbf{S})\boldsymbol{\beta}\|_2,
\end{aligned}$$

where $\mathcal{S}$ is the set of eight matrices $\left\{ \begin{pmatrix} \pm\lambda & \pm\lambda \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ \pm\lambda & \pm\lambda \end{pmatrix} \right\}$. The first step follows by inspecting the definition of $\mathcal{U}_{(1,1)}$; the second step follows from the convexity of $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} + \mathbf{u}\|_2$ (in particular, the maximum of the convex function is attained at an extreme point of $\{\mathbf{u}: \|\mathbf{u}\|_1 \leq \lambda\|\boldsymbol{\beta}\|_1\}$); and the third step follows from the definition of the ℓ_1 norm. Hence, the objective is the maximum of eight modified ℓ_2 losses.

Let us consider $\lambda = 1/2$. We claim that $\boldsymbol{\beta}^* = (1, 1)$ is an optimal solution to (B.1) with objective value $\sqrt{5}$. We will argue that $\boldsymbol{\beta}^*$ is optimal by exhibiting a dual feasible solution with the same objective value. It is easy to see that the dual (lower bounding) problem is

$$\max_{\boldsymbol{\mu} \in \mathbb{R}^{\mathcal{S}}: \sum_{\mathbf{S}} \mu_{\mathbf{S}} = 1, \boldsymbol{\mu} \geq \mathbf{0}} \min_{\boldsymbol{\beta}} \sum_{\mathbf{S}} \mu_{\mathbf{S}} \|\mathbf{y} - (\mathbf{X} + \mathbf{S})\boldsymbol{\beta}\|_2,$$

where there are eight variables $\{\mu_{\mathbf{S}} : \mathbf{S} \in \mathcal{S}\}$, one for each $\mathbf{S} \in \mathcal{S}$. Note that weak duality of the two problems is immediate. Let $\boldsymbol{\mu}^*$ be the dual feasible point with $\mu_{\mathbf{S}} = 0$ except for $\mathbf{S}_1 = \begin{pmatrix} 0 & 0 \\ -1/2 & -1/2 \end{pmatrix}$, where we set $\mu_{\mathbf{S}_1} = 1$. Hence, a lower bound to (B.1) is

$$\min_{\boldsymbol{\beta}} \sum_{\mathbf{S}} \mu_{\mathbf{S}}^* \|\mathbf{y} - (\mathbf{X} + \mathbf{S})\boldsymbol{\beta}\|_2 = \min_{\boldsymbol{\beta}} \|\mathbf{y} - (\mathbf{X} + \mathbf{S}_1)\boldsymbol{\beta}\|_2 = \sqrt{5}.$$

The final step follows by calculus, using that $\mathbf{X} + \mathbf{S}_1 = \begin{pmatrix} 1 & -1 \\ -1/2 & 1/2 \end{pmatrix}$. It follows that $\boldsymbol{\beta}^* = (1, 1)$ (with objective value $\sqrt{5}$) must be optimal to (B.1), as claimed.

We now turn our attention to the central point of interest in this Appendix, namely, that $\boldsymbol{\beta}^* = (1, 1)$ is *not* a solution to the corresponding regularization problem, viz.

$$\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2 + \rho\|\boldsymbol{\beta}\|_1, \tag{B.2}$$

for any $\rho \in (0, \infty)$ (*c.f.* Proposition 4). The solution path of (B.2) ranging over ρ is immediate from the proximal (soft-thresholding) analysis of the Lasso. In particular, it is the set of points $\{(3\alpha, 2\alpha) : \alpha \in [0, 1]\}$. This set does not contain $\boldsymbol{\beta}^* = (1, 1)$, and hence the regularization problem does not solve the robustification problem (B.1) with $\lambda = 1/2$ for any corresponding choice of ρ. (If one does not wish to rely on such an indirect analysis, note that one can solve the equivalent problem to (B.2) of $\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \mu\|\boldsymbol{\beta}\|_1$, ranging over $\mu \in (0, \infty)$. The objective is differentiable at the point $\boldsymbol{\beta}^* = (1, 1)$, and the derivative is $(-2 + \mu, 0 + \mu)$. As this is never $(0, 0)$, $\boldsymbol{\beta}^*$ can never be optimal to this problem, and consequently can never be optimal to (B.2). Despite the more direct analysis, the conclusion is the same.)

To show the converse, we can use the same example. In particular, consider the solution (3/2, 1) to (B.2) (the choice of ρ for which this is optimal is irrelevant for our purposes). We must show that (3/2, 1) is never a solution to (B.1) for any choice of λ. Let us first inspect the objective of (B.1) for $\boldsymbol{\beta}^* = (3/2, 1)$. It can be computed to be $\sqrt{1/4 + (1 + 5\lambda/2)^2}$. We make two observations:

(1) For any $0 \leq \lambda < (\sqrt{19} + 2)/15$, the point (3, 2) has strictly smaller objective (namely, 5λ) than $\boldsymbol{\beta}^*$, and so $\boldsymbol{\beta}^*$ is not optimal to (B.1) whenever $\lambda < (\sqrt{19} + 2)/15 \approx 0.424$.
(2) Similarly, for any $\lambda > (\sqrt{31} - 2)/9$, the point (1, 1) has strictly smaller objective (namely, $\sqrt{4\lambda^2 + 4\lambda + 2}$) than $\boldsymbol{\beta}^*$, and so $\boldsymbol{\beta}^*$ is not optimal to (B.1) whenever $\lambda > (\sqrt{31} - 2)/9 \approx 0.396$.

Because the intervals $[(\sqrt{19} + 2)/15, \infty)$ and $[0, (\sqrt{31} - 2)/9]$ have no overlap, the point $\boldsymbol{\beta}^* = (3/2, 1)$ cannot be a solution to (B.1) for any choice of λ.

Thus, the robustification and regularization solutions for the problems connected via Theorem 4 do not need to coincide. The statement of Theorem 5 follows as desired.

References

Bauschke, H. H., & Combettes, P. L. (2011). *Convex analysis and monotone operator theory in Hilbert spaces*. Springer.

Ben-Tal, A., Ghaoui, L. E., & Nemirovski, A. (2009). *Robust optimization*. Princeton University Press.

Ben-Tal, A., Hazan, E., Koren, T., & Mannor, S. (2015). Oracle-based robust optimization via online learning. *Operations Research, 63*(3), 628–638.

Bertsimas, D., Brown, D. B., & Caramanis, C. (2011). Theory and applications of robust optimization. *SIAM Review, 53*(3), 464–501.

Bertsimas, D., Gupta, V., & Kallus, N. (2017). Data-driven robust optimization. Mathematical Programming.

Bousquet, O., Boucheron, S., & Lugosi, G. (2004). *Advanced lectures on machine learning*. Springer.

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge University Press.

Bradic, J., Fan, J., & Wang, W. (2011). Penalized composite quasi-likelihood for ultra-high dimensional variable selection. *Journal of the Royal Statistical Society, Series B, 73*, 325–349.

Candès, E. J., Li, X., Ma, Y., & Wright, J. (2011). Robust Principal Component Analysis? *Journal of the ACM, 58*(3), 11:1–37.

Candès, E., & Recht, B. (2012). Exact matrix completion via convex optimization. *Communications of the ACM, 55*(6), 111–119.

Caramanis, C., Mannor, S., & Xu, H. (2011). *Optimization for machine learning*. MIT Press.

Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: A modern perspective* (2nd). CRC Press.

Croux, C., & Ruiz-Gazen, A. (2005). High breakdown estimators for principal components: The projection-pursuit approach revisited. *Journal of Multivariate Analysis, 95*, 206–226.

De Mol, C., De Vito, E., & Rosasco, L. (2009). Elastic-net regularization in learning theory. *Journal of Complexity, 25*(2), 201–230.

Eckart, C., & Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika, 1*. 211–8

Fan, J., Fan, Y., & Barut, E. (2014). Adaptive robust variable selection. *The Annals of Statistics, 42*(1), 324–351.

Fazel, M. (2002). *Matrix rank minimization with applications*. (Ph.D. thesis). Stanford University.

Ghaoui, L. E., & Lebret, H. (1997). Robust solutions to least-squares problems with uncertain data. *SIAM Journal of Matrix Analysis and Applications, 18*(4), 1035–1064.

Golub, G. H., & Van Loan, C. F. (1980). An analysis of the total least squares problem. *SIAM Journal of Numerical Analysis, 17*(6), 883–893.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., & Ozair, S. (2014a). Generative adversarial nets. In *Advances in neural information processing systems 27* (pp. 2672–2680).

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014b). Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association, 69*, 383–393.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer.

Hill, R. W. (1977). *Robust regression when there are outlires in the carriers*. (Ph.D. thesis). Harvard University.

Horn, R. A., & Johnson, C. R. (2013). *Matrix analysis* (2nd). Cambridge University Press.

Huber, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *The Annals of Statistics, 1*, 799–821.

Huber, P., & Ronchetti, E. (2009). *Robust statistics* (2nd). Wiley.

Hubert, M., Rousseeuw, P. J., & Aelst, S. V. (2008). High-breakdown robust multivariate methods. *Statistical Science, 23*(1), 92–119.

Hubert, M., Rousseeuw, P., & den Branden, K. V. (2005). ROBPCA: A new approach to robust principal components analysis. *Technometrics, 47*, 64–79.

Kukush, A., Markovsky, I., & Huffel, S. V. (2005). Consistency of the structured total least squares estimator in a multivariate errors-in-variables model. *Journal of Statistical Planning and Inference, 133*, 315–358.

Lewis, A. S. (2002). Robust regularization. *Technical Report*. School of ORIE, Cornell University.

Lewis, A., & Pang, C. (2009). Lipschitz behavior of the robust regularization. *SIAM Journal on Control and Optimization, 48*(5), 3080–3104.

Mallows, C. L. (1975). On some topics in robustness. *Technical Report*. Bell Laboratories.

Markovsky, I., & Huffel, S. V. (2007). Overview of total least-squares methods. *Signal Processing, 87*, 2283–2302.

Morgenthaler, S. (2007). A survey of robust statistics. *Statistical Methods and Applications, 15*, 271–293.

Mosci, S., Rosasco, L., Santoro, M., Verri, A., & Villa, S. (2010). Solving structured sparsity regularization with proximal methods. In *Proceedings of the Joint european conference on machine learning and knowledge discovery in databases* (pp. 418–433). Springer.

Recht, B., Fazel, M., & Parrilo, P. A. (2010). Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review, 52*(3), 471–501.

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association, 79*, 871–880.

Rousseeuw, P., & Leroy, A. (1987). *Robust regression and outlier detection*. Wiley.

Salibian-Barrera, M., Aelst, S. V., & Willems, G. (2005). PCA based on multivariate MM-estimators with fast and robust bootstrap. *Journal of the American Statistical Association, 101*(475), 1198–1211.

Shaham, U., Yamada, Y., & Negahban, S. (2015). Understanding adversarial training: Increasing local stability of neural nets through robust optimization. arXiv preprint arXiv:1511.05432.

SIGKDD, & Netflix (2007). Soft modelling by latent variables: The nonlinear iterative partial least squares (NIPALS) approach. *Proceedings of the KDD Cup and Workshop*.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B, 58*, 267–288.

Tulabandhula, T., & Rudin, C. (2014). Robust optimization using machine learning for uncertainty sets. arXiv preprint arXiv:1407.1097.

Xu, H., Caramanis, C., & Mannor, S. (2010). Robust regression and Lasso. *IEEE Transactions in Information Theory, 56*(7), 3561–3574.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B, 67*(2), 301–320.