Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://amstat.tandfonline.com/loi/uasa20

On Degrees of Freedom of Projection Estimators With Applications to Multivariate Nonparametric Regression

Xi Chen, Qihang Lin & Bodhisattva Sen

To cite this article: Xi Chen, Qihang Lin & Bodhisattva Sen (2020) On Degrees of Freedom of Projection Estimators With Applications to Multivariate Nonparametric Regression, Journal of the American Statistical Association, 115:529, 173-186, DOI: 10.1080/01621459.2018.1537917

To link to this article: https://doi.org/10.1080/01621459.2018.1537917

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2020, VOL. 115, NO. 529, 173–186: Theory and Methods
https://doi.org/10.1080/01621459.2018.1537917

On Degrees of Freedom of Projection Estimators With Applications to Multivariate Nonparametric Regression

Xi Chen[a], Qihang Lin[b], and Bodhisattva Sen[c]

[a]Stern School of Business, New York University, New York, NY; [b]Tippie College of Business, University of Iowa, Iowa City, IA; [c]Department of Statistics, Columbia University, New York, NY

ABSTRACT

In this article, we consider the nonparametric regression problem with multivariate predictors. We provide a characterization of the degrees of freedom and divergence for estimators of the unknown regression function, which are obtained as outputs of linearly constrained quadratic optimization procedures; namely, minimizers of the least-squares criterion with linear constraints and/or quadratic penalties. As special cases of our results, we derive explicit expressions for the degrees of freedom in many nonparametric regression problems, for example, bounded isotonic regression, multivariate (penalized) convex regression, and additive total variation regularization. Our theory also yields, as special cases, known results on the degrees of freedom of many well-studied estimators in the statistics literature, such as ridge regression, Lasso and generalized Lasso. Our results can be readily used to choose the tuning parameter(s) involved in the estimation procedure by minimizing the Stein's unbiased risk estimate. As a by-product of our analysis we derive an interesting connection between bounded isotonic regression and isotonic regression on a general partially ordered set, which is of independent interest. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received February 2017
Revised October 2018

KEYWORDS
Additive model; Bounded isotonic regression; Divergence of an estimator; Generalized Lasso; Multivariate convex regression.

1. Introduction

Consider the problem of nonparametric regression with observations $\{(\mathbf{x}_i, y_i) : i = 1, \ldots, n\}$ satisfying

$$y_i = f(\mathbf{x}_i) + \epsilon_i, \qquad \text{for } i = 1, \ldots, n, \tag{1}$$

where $\epsilon_1, \ldots, \epsilon_n$ are iid $N(0, \sigma^2)$ (unobserved) errors, $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are design points in $\mathbb{R}^d$ ($d \geq 1$) and the regression function f is unknown. In this article, we study the degrees of freedom and divergence of nonparametric estimators of f that are obtained as outputs of linearly constrained quadratic optimization procedures, namely, minimizers of the least-squares criterion with linear constraints and/or quadratic penalties. Letting $\boldsymbol{\theta}^* := (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n))$, these problems are characterized by constraints on $\boldsymbol{\theta}^*$ whereby $\boldsymbol{\theta}^* \in \mathcal{C}$ for some suitable closed convex set $\mathcal{C} \subset \mathbb{R}^n$. We briefly introduce the three main examples we will study in detail in this paper, namely isotonic regression, convex regression, and additive total variation regularization.

Example 1.1 (Isotonic regression). If f is assumed to be nondecreasing and the x_i's are univariate and ordered (i.e., $x_1 < x_2 < \cdots < x_n$), then $\boldsymbol{\theta}^* \in \mathcal{M}$, where

$$\mathcal{M} := \{\boldsymbol{\theta} \in \mathbb{R}^n : \theta_1 \leq \theta_2 \leq \ldots \leq \theta_n\}. \tag{2}$$

Isotonic regression has a long history in statistics; see, for example, Brunk (1955), Ayer et al. (1955), and van Eeden (1958). Isotonic regression can be easily extended to the setup where the predictors take values in any space with a partial order; see Section 5 for the details.

The isotonic least-squares estimator (LSE) $\widehat{\boldsymbol{\theta}}(\mathbf{y})$, which is defined as the Euclidean projection of $\mathbf{y} := (y_1, \ldots, y_n)$ onto $\mathcal{M}$, that is,

$$\widehat{\boldsymbol{\theta}}(\mathbf{y}) := \operatorname*{arg\,min}_{\boldsymbol{\theta} \in \mathcal{M}} \|\mathbf{y} - \boldsymbol{\theta}\|_2^2 \tag{3}$$

(here $\|\cdot\|_2$ denotes the usual Euclidean norm) is a natural estimator in this problem and has many desirable properties (see, e.g., Groeneboom and Jongbloed 2014). However, it suffers from the "spiking" effect (Woodroofe and Sun 1993; Pal 2008), that is, it is inconsistent at the boundary of the covariate domain. For multivariate predictors, this over-fitting of the LSE can be even more pronounced and some recent research has focused on studying the regularized isotonic LSE (see, e.g., Luss, Rosset, and Shahar 2012; Luss and Rosset 2014; Wu, Meyer, and Opsomer 2015). A natural way to regularize the model complexity would be to consider *bounded isotonic* regression: $\boldsymbol{\theta}^*$ is assumed to be nondecreasing and the range of $\boldsymbol{\theta}^*$ is assumed to be bounded by λ, for $\lambda > 0$. In Section 5, we show that for bounded isotonic regression, $\boldsymbol{\theta}^* = (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n))$ belongs to a closed polyhedral set $\mathcal{C}$ (i.e., an intersection of finitely many hyperplanes) that can be expressed in the general form as

$$\mathcal{C} = \{\boldsymbol{\theta} \in \mathbb{R}^n : B\boldsymbol{\theta} \leq \mathbf{c}\} \tag{4}$$

for some suitable matrix $B \in \mathbb{R}^{m \times n}$ and a vector $\mathbf{c} \in \mathbb{R}^{m \times 1}$; here the inequality between vectors is understood in a componentwise sense.

CONTACT Bodhisattva Sen bodhi@stat.columbia.edu Department of Statistics, Columbia University, New York, NY 10027.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/JASA.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

Example 1.2 (Convex regression). In convex regression (see, e.g., Hildreth 1954; Kuosmanen 2008; Seijo and Sen 2011; Lim and Glynn 2012; Xu, Chen, and Lafferty 2016; Han and Wellner 2016) $f : \mathbb{R}^d \to \mathbb{R}$ is known to be a convex function (see Equation (1)) and $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are the design points in $\mathbb{R}^d$, $d \geq 1$. Letting $\boldsymbol{\theta}^* := (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n))$, it can be shown that the convexity of f is equivalent to $\boldsymbol{\theta}^*$ belonging to a convex polyhedral set $\mathcal{C}$. For example, when $d = 1$ and the x_i's are ordered, $\mathcal{C}$ has the following simple characterization:

$$\mathcal{C} = \left\{ \boldsymbol{\theta} \in \mathbb{R}^n : \frac{\theta_2 - \theta_1}{x_2 - x_1} \leq \cdots \leq \frac{\theta_n - \theta_{n-1}}{x_n - x_{n-1}} \right\}. \tag{5}$$

However, for $d \geq 2$, the characterization of the underlying convex set $\mathcal{C}$ is more complex. In this case, there must exist a auxiliary vector $\boldsymbol{\xi} := [\boldsymbol{\xi}_1^\top, \ldots, \boldsymbol{\xi}_n^\top]^\top \in \mathbb{R}^{dn}$ representing the subgradient of $f(\mathbf{x}_j)$, for $j = 1, \ldots, n$, such that $\left\langle \boldsymbol{\xi}_j, \mathbf{x}_i - \mathbf{x}_j \right\rangle \leq \theta_i - \theta_j$, for $i, j = 1, \ldots, n$. Thus, $\mathcal{C}$ can be expressed as the *projection* of the higher-dimensional polyhedron

$$\left\{ (\boldsymbol{\xi}, \boldsymbol{\theta}) \in \mathbb{R}^{dn+n} : \boldsymbol{\xi} = [\boldsymbol{\xi}_1^\top, \ldots, \boldsymbol{\xi}_n^\top]^\top, \right. \\ \left. \left\langle \boldsymbol{\xi}_j, \mathbf{x}_i - \mathbf{x}_j \right\rangle \leq \theta_i - \theta_j, \forall\, i, j = 1, \ldots, n \right\}, \tag{6}$$

onto the space of $\boldsymbol{\theta}$. Although the projection of a polyhedron is still a polyhedron, it is difficult to express $\mathcal{C}$ in the form of (4) explicitly.

As before, a natural estimator of $\boldsymbol{\theta}^*$ in this problem is the LSE defined as in (3) with $\mathcal{M}$ replaced by $\mathcal{C}$. For multivariate designs, the classical convex LSE tends to overfit the data, especially near the boundary of the convex hull of the design points. To avoid this over-fitting, Sen and Meyer (2013) and Lim (2014) proposed a regularization technique using the norm of the subgradients, which leads to penalized convex regression (see Section 4 for the details).

Example 1.3 (Additive total variation regression). Suppose that $d = 1$ and f (as defined in Equation (1)) is a function of bounded variation. In this case, a popular estimator of f is to consider the total variation (TV) regularized regression (Rudin, Osher, and Fatemi 1992; also see Mammen and van de Geer 1997), which can be expressed as

$$\widehat{\boldsymbol{\theta}}(\mathbf{y}) = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^n} \sum_{i=1}^n (y_i - \theta_i)^2 + \lambda \sum_{i=2}^n |\theta_i - \theta_{i-1}|, \tag{7}$$

where $\lambda > 0$ is a tuning parameter. The presence of the ℓ_1-norm in the penalty term in (7) ensures sparsity of the vector $(\widehat{\theta}_2 - \widehat{\theta}_1, \ldots, \widehat{\theta}_n - \widehat{\theta}_{n-1})$; thus $\widehat{\boldsymbol{\theta}}(\mathbf{y})$ is piecewise constant with adaptively chosen break-points. The motivation for using (7) to estimate $\boldsymbol{\theta}^* := (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n))$ comes from the belief that $\boldsymbol{\theta}^*$ lies in the closed convex set $\mathcal{C} = \left\{ \boldsymbol{\theta} \in \mathbb{R}^n : \sum_{i=2}^n |\theta_i - \theta_{i-1}| \leq V \right\}$ for some $V > 0$; indeed (7) expresses the above constraint in the penalized form. TV regularization has many important applications, especially in image processing; also see the closely related method of fused Lasso (Tibshirani et al. 2005).

When we have multidimensional predictors, that is, $d > 1$, to alleviate the curse of dimensionality, it is useful to consider an additive model of the form $f(x_1, \ldots, x_d) := \sum_{j=1}^d f_j(x_j)$, where each $f_j(\cdot)$ is assumed to be of bounded variation. A natural estimator in this scenario, which is an extension of (7), is the additive TV regression (Petersen, Witten, and Simon 2016), where we minimize the sum of squared errors constraining the sum of the variations for each $f_j(\cdot)$. We study this estimator in Section 6.1. In fact, we consider a more general setup where each $f_j(\cdot)$ can have different degrees of "smoothness."

All the above three examples can be succinctly expressed in the Gaussian sequence model:

$$\mathbf{y} = \boldsymbol{\theta}^* + \boldsymbol{\epsilon}, \tag{8}$$

where we observe $\mathbf{y} = (y_1, \ldots, y_n) \in \mathbb{R}^n$, $\boldsymbol{\theta}^* = (\theta_1^*, \ldots, \theta_n^*) \in \mathbb{R}^n$ is the unknown parameter of interest known to belong to a given closed convex set $\mathcal{C} \subseteq \mathbb{R}^n$ (recall that $\boldsymbol{\theta}^*$ corresponds to $(f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n))$), and $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 I_n)$ (I_n is the $n \times n$ identity matrix) is the unobserved error. Let $\widehat{\boldsymbol{\theta}}(\mathbf{y}) := (\widehat{\theta}_1, \ldots, \widehat{\theta}_n)$ be an estimator of $\boldsymbol{\theta}^*$. The "degrees of freedom" of $\widehat{\boldsymbol{\theta}}(\mathbf{y})$ (see Efron 2004) is defined as

$$\mathrm{df}(\widehat{\boldsymbol{\theta}}(\mathbf{y})) := \frac{1}{\sigma^2} \sum_{i=1}^n \mathrm{cov}(\widehat{\theta}_i, y_i). \tag{9}$$

Degrees of freedom (DF) is an important concept in statistical modeling and is often used to quantify the model complexity of a statistical procedure; see, for example, Meyer and Woodroofe (2000), Zou, Hastie, and Tibshirani (2007), Tibshirani and Taylor (2012), and the references therein. Intuitively, the quantity $\mathrm{df}(\widehat{\boldsymbol{\theta}}(\mathbf{y}))$ reflects the effective number of parameters used by $\widehat{\boldsymbol{\theta}}(\mathbf{y})$ in producing the fitted output, for example, in linear regression, if $\widehat{\boldsymbol{\theta}}(\mathbf{y})$ is the LSE of $\mathbf{y}$ onto a subspace of dimension $d < n$, the DF of $\widehat{\boldsymbol{\theta}}(\mathbf{y})$ is simply d. Using Stein's lemma it follows that (see Meyer and Woodroofe 2000; Tibshirani and Taylor 2012)

$$\mathrm{df}(\widehat{\boldsymbol{\theta}}(\mathbf{y})) = \mathbb{E}[D(\mathbf{y})],$$

where

$$D(\mathbf{y}) = \mathrm{div}(\widehat{\boldsymbol{\theta}}(\mathbf{y})) := \sum_{i=1}^n \frac{\partial}{\partial y_i} \widehat{\theta}_i(\mathbf{y}) = \nabla_{\mathbf{y}} \widehat{\boldsymbol{\theta}}(\mathbf{y}) \tag{10}$$

is called the *divergence* of $\widehat{\boldsymbol{\theta}}(\mathbf{y})$. Thus, $D(\mathbf{y})$ is an unbiased estimator of $\mathrm{df}(\widehat{\boldsymbol{\theta}}(\mathbf{y}))$. This has many important implications, for example, Stein's unbiased risk estimate (SURE); see Stein (1981). Aside from plainly estimating the risk of an estimator, one could also use SURE for model selection purposes: if the estimator depends on a tuning parameter, then one could choose this parameter by minimizing SURE. This has been successfully used in many statistical problems; see, for example, Donoho and Johnstone (1995), Xie, Kou, and Brown (2012), Candès, Sing-Long, and Trzasko (2013), and Yi and Zou (2013) for applications in wavelet denoising, heteroscedastic hierarchical models, singular value thresholding, and bandable covariance matrices, respectively. We elaborate on this connection in Section 7.

In this article, we develop a theoretical framework to evaluate the divergence (as defined in Equation (10)) for a broad class of (nonparametric) regression estimators that are minimizers of the least-squares criterion with linear constraints and/or quadratic penalties. Our theory also recovers many existing results (see Section D in the supplementary material), which include the exact expressions for divergence for ridge regression

(see Li 1986) and the active set representation of the divergence for Lasso and generalized Lasso (see Zou, Hastie, and Tibshirani 2007; Tibshirani and Taylor 2012).

In the following, we motivate the general form of the estimators we study in this article. In many regression problems, $\boldsymbol{\theta}^* \in \mathcal{C} \subset \mathbb{R}^n$ where $\mathcal{C}$ is a polyhedron. Moreover, in many of these problem (e.g., convex regression) $\mathcal{C}$ is not easily expressible in the form (4), but can be described as the projection of a higher-dimensional polyhedron of $(\boldsymbol{\xi}, \boldsymbol{\theta})$ onto the space of $\boldsymbol{\theta}$ (see, e.g., Equation (6)). In particular, this higher-dimensional polyhedron can, in general, be represented as

$$\mathcal{Q} := \{(\boldsymbol{\xi}, \boldsymbol{\theta}) \in \mathbb{R}^{p+n} : A\boldsymbol{\xi} + B\boldsymbol{\theta} \leq \mathbf{c}\}, \tag{11}$$

where $\boldsymbol{\xi} \in \mathbb{R}^p$ is the auxiliary variable and $A \in \mathbb{R}^{m\times p}, B \in \mathbb{R}^{m\times n}$ and $\mathbf{c} \in \mathbb{R}^m$ are suitable matrices. The true parameter $\boldsymbol{\theta}^*$ thus belongs to the set $\mathcal{C} := \text{Proj}_{\boldsymbol{\theta}}(\mathcal{Q})$ defined as

$$\text{Proj}_{\boldsymbol{\theta}}(\mathcal{Q}) := \{\boldsymbol{\theta} \in \mathbb{R}^n : \exists\ \boldsymbol{\xi} \in \mathbb{R}^p \text{ such that } (\boldsymbol{\xi}, \boldsymbol{\theta}) \in \mathcal{Q}\}. \tag{12}$$

A natural estimator of $\boldsymbol{\theta}^*$ in this situation is the LSE $\widehat{\boldsymbol{\theta}}(\mathbf{y}) := \arg\min_{\boldsymbol{\theta}\in\text{Proj}_{\boldsymbol{\theta}}(\mathcal{Q})} \frac{1}{2}\|\boldsymbol{\theta}-\mathbf{y}\|_2^2$, which is equivalent to $(\widehat{\boldsymbol{\theta}}(\mathbf{y}), \widehat{\boldsymbol{\xi}}(\mathbf{y})) \in \arg\min_{(\boldsymbol{\theta},\boldsymbol{\xi})\in\mathcal{Q}} \frac{1}{2}\|\boldsymbol{\theta} - \mathbf{y}\|_2^2$. Instead of considering this partially projected LSE, we study a more general formulation by adding *linear* and *quadratic perturbations* in the objective function to accommodate more applications:

$$(\widehat{\boldsymbol{\theta}}(\mathbf{y}), \widehat{\boldsymbol{\xi}}(\mathbf{y})) \in \arg\min_{\boldsymbol{\theta},\boldsymbol{\xi}} \frac{1}{2}\|\boldsymbol{\theta} - \mathbf{y}\|_2^2 + \mathbf{d}^\top\boldsymbol{\xi} + \frac{\lambda}{2}\|\boldsymbol{\xi}\|_2^2 \quad \text{s.t. } A\boldsymbol{\xi} + B\boldsymbol{\theta} \leq \mathbf{c}, \tag{13}$$

where $A = [\mathbf{a}_1, \ldots, \mathbf{a}_m]^\top \in \mathbb{R}^{m\times p}, B = [\mathbf{b}_1, \ldots, \mathbf{b}_m]^\top \in \mathbb{R}^{m\times n}$, $\mathbf{c} \in \mathbb{R}^m$, $\mathbf{d} \in \mathbb{R}^p$ and $\lambda \geq 0$ is a regularization parameter. As we will show below (13) finds many statistical applications beyond the examples described above. Note that the objective function in (13) is strongly convex in $\boldsymbol{\theta}$ and convex in $\boldsymbol{\xi}$; moreover, if $\lambda > 0$, it is strongly convex in both $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$.

Formulation (13) covers a wide range of useful estimators in shape-restricted nonparametric regression, additive total variation regression, and Lasso-related problems. For example, when $\mathbf{d} = \mathbf{0}$, $\lambda = 0$ but A is not a zero matrix, (13) becomes

$$(\widehat{\boldsymbol{\theta}}(\mathbf{y}), \widehat{\boldsymbol{\xi}}(\mathbf{y})) = \arg\min_{(\boldsymbol{\theta},\boldsymbol{\xi})\in\mathcal{Q}} \frac{1}{2}\|\boldsymbol{\theta} - \mathbf{y}\|_2^2, \tag{14}$$

where $\mathcal{Q}$ is defined in (11). This formulation can also be viewed as the projection of $\mathbf{y}$ onto a polyhedron $\text{Proj}_{\boldsymbol{\theta}}(\mathcal{Q})$ defined in (12). This class of problems include the LSE in multivariate convex regression for which DF has not been studied before (see Section 4 for the details). Based on (14), if we further have $\mathbf{d} \neq \mathbf{0}$, then (13) reduces to

$$(\widehat{\boldsymbol{\theta}}(\mathbf{y}), \widehat{\boldsymbol{\xi}}(\mathbf{y})) = \arg\min_{(\boldsymbol{\theta},\boldsymbol{\xi})\in\mathcal{Q}} \frac{1}{2}\|\boldsymbol{\theta} - \mathbf{y}\|_2^2 + \mathbf{d}^\top\boldsymbol{\xi}. \tag{15}$$

This formulation includes many examples in statistics, such as additive TV regression (see Example 1.3) and ℓ_∞-regularized group Lasso (see Section 6). Moreover, when $\mathbf{d} = \mathbf{0}$ and $\lambda > 0$ in (13), the corresponding optimization problem becomes

$$(\widehat{\boldsymbol{\theta}}(\mathbf{y}), \widehat{\boldsymbol{\xi}}(\mathbf{y})) = \arg\min_{(\boldsymbol{\theta},\boldsymbol{\xi})\in\mathcal{Q}} \frac{1}{2}\|\boldsymbol{\theta} - \mathbf{y}\|_2^2 + \frac{\lambda}{2}\|\boldsymbol{\xi}\|_2^2, \tag{16}$$

which includes the example of penalized multivariate convex regression, where the norm of the subgradient $\boldsymbol{\xi}$ is penalized.

In the following we briefly describe some of the main contributions of this article.

1. We characterize the divergence and DF of $\widehat{\boldsymbol{\theta}}(\mathbf{y})$, as defined in (13), by providing easy-to-compute formulas. Our main result, Theorem 3.2, can be used to compute the divergence and DF in any statistical regression problem where the estimator can be expressed in the form (13). A special case of (13)—projection onto a convex polyhedron—has been studied in the literature (Kato 2009; Tibshirani and Taylor 2012) where
$$\widehat{\boldsymbol{\theta}}(\mathbf{y}) = P_{\mathcal{C}}(\mathbf{y}) := \arg\min_{\boldsymbol{\theta}\in\mathcal{C}} \frac{1}{2}\|\boldsymbol{\theta} - \mathbf{y}\|_2^2, \tag{17}$$
and $\mathcal{C} = \{\boldsymbol{\theta} \in \mathbb{R}^n : B\boldsymbol{\theta} \leq \mathbf{c}\}$ is as defined in (4). Our main theorem generalizes these previous results. In particular, when $\mathbf{d} \neq 0$ and $\lambda = 0$ in (13), the problem is challenging as now $\widehat{\boldsymbol{\theta}}(\mathbf{y})$ cannot be written as a projection estimator. When $\lambda > 0$, although (13) can be viewed as a projection problem in a higher dimensional space, the previous results on the projection estimator cannot be directly applied to obtain the divergence of $\widehat{\boldsymbol{\theta}}(\mathbf{y})$ (see Remark 3.1 for details).
2. Using our main result we derive the DF for many estimators, including multivariate convex regression, penalized convex regression, (bounded) isotonic regression, additive TV regression, ℓ_∞-regularized group Lasso, etc. Note that although the divergences and DF for Lasso and generalized Lasso have been characterized in Zou, Hastie, and Tibshirani (2007) and Tibshirani and Taylor (2012) we demonstrate that we recover their results (in the active set representation) as straightforward consequences of Theorem 3.2; see Section D in the supplemental material for the details.
3. For bounded isotonic regression where the design points are allowed to belong to any partially ordered set, we establish the equivalence between the divergence of the isotonic LSE and the number of connected components of the graph induced by the LSE (see Proposition 5.2). This result is not only theoretically interesting but also provides a fast algorithm for computing the divergence in this problem. Moreover, we establish a connection between the LSE for bounded isotonic regression and that for unbounded isotonic regression, a result which is of independent interest. In particular, we show that the bounded isotonic LSE can be easily obtained by appropriately thresholding the unbounded isotonic LSE (see Proposition 5.3). Further, using this property, we show the monotonicity of divergence (and DF) as a function of the model complexity parameter—this shows that DF indeed characterizes model complexity—for bounded isotonic regression.

In the following we compare and contrast our results with some of the recent work on divergence and DF for projection estimators. Kato (2009) characterizes the DF in shrinkage regression where the coefficients belong to a closed convex set. The estimation problem considered by Kato (2009) contains (14) as a special case but his result cannot be directly applied

to (15) when $\mathbf{d} \neq \mathbf{0}$. As a consequence, Kato (2009) can characterize DF for generalized Lasso expressed in a constrained form while we can characterize the DF in the penalized form (as described in Section D of the supplementary material). Hansen and Sokol (2014) consider the closed constraint set $\mathcal{C} = \zeta(\mathcal{B})$ where $\mathcal{B} \subseteq \mathbb{R}^p$ is a closed set and $\zeta : \mathbb{R}^p \to \mathbb{R}^n$ is a (possibly nonlinear) map satisfying some regularity conditions. Their main result (Theorem 3) requires the optimal solution $\widehat{\boldsymbol{\beta}}$ to be in the *interior* of $\mathcal{B}$ (which is almost never the case in the examples of interest to us) and a variant of the Hessian matrix of $\zeta(\widehat{\boldsymbol{\beta}})$ to be full rank (e.g., when $\zeta(\boldsymbol{\beta}) = X\boldsymbol{\beta}$, it requires that $X^\top X$ is full rank). The results in Hansen and Sokol (2014) can only deal with a constraint set that can be explicitly written as a set of inequalities (e.g., the general projected polyhedron $\mathrm{Proj}_\theta(\mathcal{Q})$ in (12) is not allowed) and cannot be applied to regularized estimators (e.g., generalized Lasso as described in Section D of the supplementary material and penalized multivariate convex regression as described in Section 4). Vaiter et al. (2014) studied DF for a class of regularized regression problems that include Lasso and group Lasso as special cases. However, their paper does not consider constrained formulations and thus cannot be applied to shape restricted regression problems. Mikkelsen and Hansen (2018) provide a characterization of DF for a class of estimators, which are locally Lipschitz continuous on each of a finite number of open sets that cover $\mathbb{R}^n$. Rueda (2013) used the results of Meyer and Woodroofe (2000) to study the DF for the specific problem of semiparametric additive (univariate) monotone regression.

In the recent papers Kaufman and Rosset (2014) and Janson, Fithian, and Hastie (2015) the authors argued that in many problems DF might not be an appropriate notion for characterizing model complexity. They provide counter examples of situations where DF is not monotone in the model complexity parameter (or DF is unbounded). However, most of these counter examples either involve nonconvex constraints or non-Gaussian or heteroscedastic noise—in Janson, Fithian, and Hastie (2015) it is argued that such irregular behavior happens "whenever we project onto a nonconvex model." Nevertheless, some of the main applications in our article, namely, bounded isotonic regression and additive total variation regression, correspond to projections onto polyhedral convex sets with iid Gaussian noise so the irregular behavior of DF, observed in some of the counter examples, may not occur here. In fact, in Theorem 5.4 we prove that for bounded isotonic regression, DF is indeed monotone in the model complexity parameter.

The article is organized as follows. In Section 2, we provide some basic results on the divergence of projection estimators. In Section 3, we state our main result. In Sections 4, 5, and 6, we discuss many applications of our main result to different regression problems. In Section 7, we discuss how the characterization of divergence of estimators (computed in the article) can be useful in model selection (choice of tuning parameter) based on SURE, and illustrate this for bounded isotonic regression and penalized multivariate convex regression. We relegate all the technical proofs, graphical illustrations, as well as the derivation of some existing results (such as generalized Lasso) using our main theorem to the supplementary material.

2. An Existing Result on DF

Degrees of Freedom is an important concept in statistical modeling as it provides a quantitative description of the amount of fitting performed by a given procedure. Despite its fundamental role in statistics, its behavior is not completely well-understood, even for widely used estimators.

In this section, we review an important known result on DF and the divergence of the projection estimator $\widehat{\boldsymbol{\theta}}(\mathbf{y})$ when $\mathcal{C}$ is a convex polyhedron as defined in (4); see (17). We will assume that the reader is familiar with basic concepts from convex analysis (see Section A in the supplementary material where we provide a review of some basic concepts: polyhedron, cone, normal cone, affine hull, interior, boundary, relative interior, relative boundary, etc).

The following result, due to Kato (2009, Lemma 3.2) and Tibshirani and Taylor (2012, Lemma 2), shows that the divergence of the projection estimator $\widehat{\boldsymbol{\theta}}(\mathbf{y})$ onto a convex polyhedron as described in (4) can be calculated as the dimension of the affine space that $\widehat{\boldsymbol{\theta}}(\mathbf{y})$ lies on. In fact, Lemma 3.2 in Kato (2009) provides a more general result about the divergence of the projection estimator $\widehat{\boldsymbol{\theta}}(\mathbf{y})$, when $\mathcal{C}$ is a closed convex set with piecewise smooth boundary.

Proposition 2.1. Suppose that the projection estimator $\widehat{\boldsymbol{\theta}}(\mathbf{y})$ is defined in (17) where $\mathcal{C}$ is a convex polyhedron as defined in (4). Then the components of $\widehat{\boldsymbol{\theta}}(\mathbf{y})$ are almost differentiable, and $\nabla\widehat{\theta}_i$ (ith entry of $\nabla\widehat{\boldsymbol{\theta}}(\mathbf{y})$) is an essentially bounded function, for $i = 1, \ldots, n$. Let $J_{\mathbf{y}}$ be the set of indices for all the binding constraints of $\widehat{\boldsymbol{\theta}}(\mathbf{y})$, that is,

$$J_{\mathbf{y}} := \{1 \leq i \leq m : \langle \mathbf{b}_i, \widehat{\boldsymbol{\theta}}(\mathbf{y})\rangle = c_i\}. \tag{18}$$

Then, for a.e. $\mathbf{y} \in \mathbb{R}^n$, there is a neighborhood U of $\mathbf{y}$, such that for every $\mathbf{z} \in U$,

$$\widehat{\boldsymbol{\theta}}(\mathbf{z}) = \operatorname*{arg\,min}_{\boldsymbol{\theta} \in H} \frac{1}{2}\|\boldsymbol{\theta} - \mathbf{z}\|_2^2, \tag{19}$$

where $H = \{\boldsymbol{\theta} : B_{J_{\mathbf{y}}}\boldsymbol{\theta} = \mathbf{c}_{J_{\mathbf{y}}}\}$ is an affine space, $J_{\mathbf{y}}$ is defined in (18) and $B_{J_{\mathbf{y}}}$ is the submatrix of B with rows indexed by $J_{\mathbf{y}}$. As a consequence,

$$D(\mathbf{y}) = n - \mathrm{rank}(B_{J_{\mathbf{y}}}), \qquad \text{for a.e. } \mathbf{y} \in \mathbb{R}^n. \tag{20}$$

Thus, $\mathrm{df}(\widehat{\boldsymbol{\theta}}(\mathbf{y})) = n - \mathbb{E}\left[\mathrm{rank}(B_{J_{\mathbf{y}}})\right]$.

Note that a.e. in (20) stands for "almost everywhere," which means that (20) holds for all $\mathbf{y}$ except on a measure-zero set. Note that, by an almost differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ we mean that f is differentiable everywhere except on a measure-zero set (see Meyer and Woodroofe 2000 for a precise definition); f is essentially bounded if there exists an constant c such that $f^{-1}((c, +\infty))$ is a measure-zero set.

3. Main Result

In this section, we consider the estimator $\widehat{\boldsymbol{\theta}}(\mathbf{y})$ obtained from the optimization problem (13) with the auxiliary variable $\boldsymbol{\xi} \in \mathbb{R}^p$. When $\lambda = 0$ and $\mathbf{d} \neq \mathbf{0}$, the optimization problem (13) may have an unbounded optimal value depending on $\mathbf{d}$. The following result gives the necessary and sufficient condition for (13) to be bounded.

Lemma 3.1. When $\lambda = 0$, the optimization problem in (13) has a bounded optimal value if and only if $-\mathbf{d} = A^\top \mathbf{u}$ for some $\mathbf{u} \geq \mathbf{0}$.

The proof of Lemma 3.1 is based on Farkas's lemma (see, e.g., Rockafellar 1970, Corollary 22.3.1) and is provided in Section B.1 of the supplementary material. Based on the above lemma, for the rest of the article, we will assume that $-\mathbf{d} = A^\top \mathbf{u}$ for some $\mathbf{u} \geq \mathbf{0}$ so that (13) is bounded. When $\mathbf{d} = \mathbf{0}$ such an assumption trivially holds for $\mathbf{u} = \mathbf{0}$. For applications with $\mathbf{d} \neq \mathbf{0}$, for example, additive model, generalized Lasso, and ℓ_∞-regularized group Lasso, we will show that this assumption always holds.

The divergence of $\widehat{\boldsymbol{\theta}}(\mathbf{y})$, as the solution (13), is characterized by the following theorem, which is the main result of the article.

Theorem 3.2. Suppose that $-\mathbf{d} = A^\top \mathbf{u}$ for some $\mathbf{u} \geq \mathbf{0}$ whenever $\lambda = 0$ in (13). For any $\mathbf{y} \in \mathbb{R}^n$, let $(\widehat{\boldsymbol{\theta}}(\mathbf{y}), \widehat{\boldsymbol{\xi}}(\mathbf{y}))$ be any solution for (13) and let

$$J_\mathbf{y} := \{1 \leq i \leq m : \langle \mathbf{a}_i, \widehat{\boldsymbol{\xi}}(\mathbf{y}) \rangle + \langle \mathbf{b}_i, \widehat{\boldsymbol{\theta}}(\mathbf{y}) \rangle = c_i\}, \tag{21}$$

and $A_{J_\mathbf{y}}$ and $B_{J_\mathbf{y}}$ be the submatrices of A and B with rows in the set $J_\mathbf{y}$. Let $I_\mathbf{y} \subseteq J_\mathbf{y}$ be the index set of maximal independent rows of the matrix $[A_{J_\mathbf{y}}, B_{J_\mathbf{y}}]$, that is, the set of vectors $\{[\mathbf{a}_i^\top, \mathbf{b}_i^\top], i \in I_\mathbf{y}\}$ are linearly independent. Then, the following statements hold:

(i) The optimal solution $(\widehat{\boldsymbol{\theta}}(\mathbf{y}), \widehat{\boldsymbol{\xi}}(\mathbf{y}))$ of (13) has unique components $\widehat{\boldsymbol{\theta}}(\mathbf{y})$. The components of $\widehat{\boldsymbol{\theta}}(\mathbf{y})$ are almost differentiable in $\mathbf{y}$ and $\nabla \widehat{\theta}_i(\mathbf{y})$ is an essentially bounded function for each $i = 1, \ldots, n$.

(ii) For a.e. $\mathbf{y}$,

$$D(\mathbf{y}) = \begin{cases} n - \operatorname{trace}\left(B_{I_\mathbf{y}}^\top \left(B_{I_\mathbf{y}} B_{I_\mathbf{y}}^\top + \frac{1}{\lambda} A_{I_\mathbf{y}} A_{I_\mathbf{y}}^\top\right)^{-1} B_{I_\mathbf{y}}\right), \\ \qquad\qquad\qquad\qquad \text{if } \lambda > 0, \\ n - |I_\mathbf{y}| + \operatorname{rank}(A_{I_\mathbf{y}}), \\ \qquad\qquad\qquad\qquad \text{if } \lambda = 0, \end{cases} \tag{22}$$

and $\operatorname{df}(\widehat{\boldsymbol{\theta}}(\mathbf{y})) = \mathbb{E}[D(\mathbf{y})]$ (note that the index set $I_\mathbf{y}$ is random).

First note that any solution $(\widehat{\boldsymbol{\theta}}(\mathbf{y}), \widehat{\boldsymbol{\xi}}(\mathbf{y}))$ of (13) depends on $\mathbf{d}$ and so do $J_\mathbf{y}$ and $I_\mathbf{y}$. Hence, $D(\mathbf{y})$ given in (22) depends on $\mathbf{d}$ implicitly. To simplify notation, we suppress the dependence of $J_\mathbf{y}$, $I_\mathbf{y}$ and $D(\mathbf{y})$ on $\mathbf{d}$. The divergence in (22) holds for any given $\mathbf{d} \in \mathbb{R}^p$ and for every $\mathbf{y} \in \mathbb{R}^n$ expect for a measure-zero set in $\mathbb{R}^n$. The explicit form of this measure zero set is provided in our proof (see Equation (60) in the supplementary material for the case $\lambda = 0$ and (65) when $\lambda > 0$).

We also note that when $\lambda > 0$, $B_{I_\mathbf{y}} B_{I_\mathbf{y}}^\top + \frac{1}{\lambda} A_{I_\mathbf{y}} A_{I_\mathbf{y}}^\top$ is invertible. To see this observe that, from the definition of $I_\mathbf{y}$, the rows of $V := [\frac{1}{\sqrt{\lambda}} A_{I_\mathbf{y}}, B_{I_\mathbf{y}}]$ are linearly independent. Therefore, $B_{I_\mathbf{y}} B_{I_\mathbf{y}}^\top + \frac{1}{\lambda} A_{I_\mathbf{y}} A_{I_\mathbf{y}}^\top = VV^\top$ is invertible. Further, as a simple sanity check of Theorem 3.2, we show in Lemma B.3 (see Section B.4 of the supplementary material that $D(\mathbf{y})$, as defined in (22), is always nonnegative. A few important remarks are in order now.

Remark 3.1. When $\lambda > 0$, we can define $\mathbf{d}_\lambda := \frac{-\mathbf{d}}{\sqrt{\lambda}}$ and can reformulate (13) as a projection problem

$$(\widehat{\boldsymbol{\theta}}(\mathbf{y}, \mathbf{d}_\lambda), \widehat{\boldsymbol{\gamma}}(\mathbf{y}, \mathbf{d}_\lambda)) = \operatorname*{arg\,min}_{\boldsymbol{\theta}, \boldsymbol{\gamma}} \frac{1}{2} \left\| (\boldsymbol{\theta}, \boldsymbol{\gamma}) - (\mathbf{y}, \mathbf{d}_\lambda) \right\|_2^2 \tag{23}$$
$$\text{s.t. } \frac{1}{\sqrt{\lambda}} A\boldsymbol{\gamma} + B\boldsymbol{\theta} \leq \mathbf{c}.$$

It is easy to verify that $\widehat{\boldsymbol{\gamma}} = \sqrt{\lambda}\, \widehat{\boldsymbol{\xi}}$ and that (23) is just an instance of (17) in $\mathbb{R}^{p+n}$ by viewing $(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\gamma}})$, $(\mathbf{y}, \mathbf{d}_\lambda)$ and the feasible domain $\{(\boldsymbol{\theta}, \boldsymbol{\gamma}) \in \mathbb{R}^{p+n} : \frac{1}{\sqrt{\lambda}} A\boldsymbol{\gamma} + B\boldsymbol{\theta} \leq \mathbf{c}\}$ in (23) as $\widehat{\boldsymbol{\theta}}$, $\mathbf{y}$ and $\mathcal{C}$ in (17), respectively. Hence, by applying Proposition 2.1 to (23), we can show that, for a.e. $(\mathbf{y}, \mathbf{d}_\lambda) \in \mathbb{R}^{p+n}$, there is a neighborhood U of $(\mathbf{y}, \mathbf{d}_\lambda)$, such that for every $(\mathbf{z}, \mathbf{b}) \in U$, the solution $(\widehat{\boldsymbol{\theta}}(\mathbf{z}, \mathbf{b}), \widehat{\boldsymbol{\gamma}}(\mathbf{z}, \mathbf{b}))$ defined in (23) is the projection of $(\mathbf{z}, \mathbf{b})$ to the affine space $\{(\boldsymbol{\theta}, \boldsymbol{\gamma}) : \frac{1}{\sqrt{\lambda}} A_{I_\mathbf{y}} \boldsymbol{\gamma} + B_{I_\mathbf{y}} \boldsymbol{\theta} = \mathbf{c}_{I_\mathbf{y}}\}$ with $I_\mathbf{y}$ defined the same as in Theorem 3.2. In other words, for every $(\mathbf{z}, \mathbf{b}) \in U$,

$$\begin{bmatrix} \widehat{\boldsymbol{\theta}}(\mathbf{z}, \mathbf{b}) \\ \widehat{\boldsymbol{\gamma}}(\mathbf{z}, \mathbf{b}) \end{bmatrix} = (I - P) \begin{bmatrix} \mathbf{z} \\ \mathbf{b} \end{bmatrix},$$
$$\text{where } P = \begin{bmatrix} B_{I_\mathbf{y}}^\top \\ \frac{1}{\sqrt{\lambda}} A_{I_\mathbf{y}}^\top \end{bmatrix} \left(B_{I_\mathbf{y}} B_{I_\mathbf{y}}^\top + \frac{1}{\lambda} A_{I_\mathbf{y}} A_{I_\mathbf{y}}^\top\right)^{-1} \left[B_{I_\mathbf{y}}, \frac{1}{\sqrt{\lambda}} A_{I_\mathbf{y}}\right].$$

Therefore, for a.e. $(\mathbf{y}, \mathbf{d}_\lambda) \in \mathbb{R}^{p+n}$, the matrix $I - P$ is the Jacobian matrix of $(\widehat{\boldsymbol{\theta}}(\mathbf{y}, \mathbf{d}_\lambda), \widehat{\boldsymbol{\gamma}}(\mathbf{y}, \mathbf{d}_\lambda))$ and we obtain (22) for $\lambda > 0$ by taking the trace of the $n \times n$ top-left block of $I - P$.

Unfortunately, this argument cannot serve as a proof for Theorem 3.2 when $\lambda > 0$ as the above argument only holds for almost every $(\mathbf{y}, \mathbf{d}_\lambda)$ in $\mathbb{R}^{p+n}$ but *not necessarily* for almost every $\mathbf{y}$ in $\mathbb{R}^n$ for a given $\mathbf{d}_\lambda$. This is because the projection of a zero-measure set in $\mathbb{R}^{p+n}$ (i.e., the set of $(\mathbf{y}, \mathbf{d}_\lambda)$'s) onto the space of $\mathbf{y}$ is not necessarily a zero-measure set in $\mathbb{R}^n$. But our main result in Theorem 3.2 shows that (22) holds for almost every $\mathbf{y} \in \mathbb{R}^n$ and any given $\mathbf{d}_\lambda \in \mathbb{R}^p$. In Section B.5 in the supplementary material, we present a concrete example which shows that the entire set of $(\mathbf{y}, \mathbf{d}_\lambda)$ with a given $\mathbf{d}_\lambda$ falls into the measure-zero part on which the previous results from Kato (2009) and Tibshirani and Taylor (2012) fail.

Remark 3.2. When $\lambda = 0$, using the strong duality of linear programming, we can reformulate (13) and $\widehat{\boldsymbol{\theta}}(\mathbf{y})$ as follows:

$$\widehat{\boldsymbol{\theta}}(\mathbf{y}) \in \operatorname*{arg\,min}_{\boldsymbol{\theta}} \frac{1}{2} \|\boldsymbol{\theta} - \mathbf{y}\|_2^2 + g(\boldsymbol{\theta}), \tag{24}$$

where $g(\boldsymbol{\theta})$ is a piece-wise linear convex function:

$$g(\boldsymbol{\theta}) := \begin{cases} \min\limits_{\boldsymbol{\xi}} \mathbf{d}^\top \boldsymbol{\xi} \text{ s.t. } A\boldsymbol{\xi} \leq \mathbf{c} - B\boldsymbol{\theta} & \text{if } \{\boldsymbol{\xi} | A\boldsymbol{\xi} \leq \mathbf{c} - B\boldsymbol{\theta}\} \neq \emptyset \\ +\infty & \text{if } \{\boldsymbol{\xi} | A\boldsymbol{\xi} \leq \mathbf{c} - B\boldsymbol{\theta}\} = \emptyset. \end{cases} \tag{25}$$
$$= \begin{cases} \max\limits_{\mathbf{u}} (B\boldsymbol{\theta} - \mathbf{c})^\top \mathbf{u} \text{ s.t. } A^\top \mathbf{u} = -\mathbf{d}, \mathbf{u} \geq \mathbf{0} \\ \qquad\qquad\qquad\qquad \text{if } \{\boldsymbol{\xi} | A\boldsymbol{\xi} \leq \mathbf{c} - B\boldsymbol{\theta}\} \neq \emptyset \\ +\infty \qquad\qquad\qquad\qquad \text{if } \{\boldsymbol{\xi} | A\boldsymbol{\xi} \leq \mathbf{c} - B\boldsymbol{\theta}\} = \emptyset. \end{cases}$$

The formulation (24) means that $\widehat{\boldsymbol{\theta}}(\mathbf{y})$ is the proximal mapping of $\mathbf{y}$ with a proximal term g (Definition 1.22 in Rockafellar and Wets 2011). We note that Exercise 13.45 from Rockafellar and Wets (2011) characterizes the generalized Jacobian of a proximal mapping, which can be a potential tool to derive $D(\mathbf{y})$.

However, due to the complicated form of the proximal term g in (25), it is not easy to directly apply their result to derive the explicit expression of the divergence in our Theorem 3.2, and it requires to first introduce many new notions (e.g., second order generalized derivative for nonsmooth functions and graphical derivative) in variational analysis. On the other hand, our proof for the case of $\lambda = 0$ is more elementary and more consistent with the proof when $\lambda > 0$—both of them are based on a general local projection lemma (see Lemma 3.3).

Remark 3.3. The computation of the index set $J_{\mathbf{y}}$ is straightforward. Given a solution $\widehat{\boldsymbol{\xi}}(\mathbf{y})$ and $\widehat{\boldsymbol{\theta}}(\mathbf{y})$ from an optimization solver, we could easily check if $\langle \mathbf{a}_i, \widehat{\boldsymbol{\xi}}(\mathbf{y})\rangle + \langle \mathbf{b}_i, \widehat{\boldsymbol{\theta}}(\mathbf{y})\rangle$ equals c_i, for each $1 \le i \le m$. After obtaining $J_{\mathbf{y}}$, the index set $I_{\mathbf{y}}$ of maximal independent rows can be found by removing all the rows of $[A_{J_{\mathbf{y}}}, B_{J_{\mathbf{y}}}]$ whose removal does not change the rank of the original matrix $[A_{J_{\mathbf{y}}}, B_{J_{\mathbf{y}}}]$. In particular, we start with an index set $K = J_{\mathbf{y}}$. For each row index $k \in K$, if the rank of $[A_{K\setminus\{k\}}, B_{K\setminus\{k\}}]$ is the same as that of $[A_K, B_K]$, we remove k from K. (Note that the rank can be computed easily by singular value decomposition or by directly applying the *rank* function in Matlab or *rankMatrix* function in R.) We repeat this procedure until no additional index in K can be removed without reducing the rank of the matrix. The obtained index set K is $I_{\mathbf{y}}$.

Remark 3.4. When $\lambda = 0$, it is possible that there exist multiple $\widehat{\boldsymbol{\xi}}(\mathbf{y})$'s satisfying (13) and they correspond to different $J_{\mathbf{y}}$'s and $I_{\mathbf{y}}$'s; while when $\lambda > 0$, $\widehat{\boldsymbol{\xi}}(\mathbf{y})$ is unique. Even if $\widehat{\boldsymbol{\xi}}(\mathbf{y})$ and $J_{\mathbf{y}}$ are unique, there can still exist multiple maximal independent sets $I_{\mathbf{y}}$. However, according to our proof, for any given $\widehat{\boldsymbol{\xi}}(\mathbf{y})$, $J_{\mathbf{y}}$ and $I_{\mathbf{y}}$, we show that $D(\mathbf{y})$ equals the quantity on the right hand side of (22). Note that $D(\mathbf{y})$ is well-defined (see its definition in (10)), unique and does not depend on the choice of $\widehat{\boldsymbol{\xi}}(\mathbf{y})$, $J_{\mathbf{y}}$ and $I_{\mathbf{y}}$.

The key tool to proving Theorem 3.2 is to establish the following lemma, which shows that for a.e. $\mathbf{y}$, the solution of (13) is locally an affine projection *with linear and quadratic perturbations*.

Lemma 3.3. Suppose that $-\mathbf{d} = A^\top \mathbf{u}$ for some $\mathbf{u} \ge \mathbf{0}$ whenever $\lambda = 0$ in (13). For any $\mathbf{y} \in \mathbb{R}^n$, let $(\widehat{\boldsymbol{\theta}}(\mathbf{y}), \widehat{\boldsymbol{\xi}}(\mathbf{y}))$ be any solution of (13) and let the index set $J_{\mathbf{y}}$ be as defined in (21). For a.e. $\mathbf{y} \in \mathbb{R}^n$,

$$\widehat{\boldsymbol{\theta}}(\mathbf{z}) = \widetilde{\boldsymbol{\theta}}(\mathbf{z}), \text{ for any } \mathbf{z} \text{ in a neighborhood } U \text{ of } \mathbf{y}, \tag{26}$$

where $\widetilde{\boldsymbol{\theta}}(\mathbf{z})$ is defined as the unique $\boldsymbol{\theta}$-component of the optimal solution of the following optimization problem:

$$(\widetilde{\boldsymbol{\theta}}(\mathbf{z}), \widetilde{\boldsymbol{\xi}}(\mathbf{z})) \in \underset{\boldsymbol{\theta},\boldsymbol{\xi}}{\arg\min} \frac{1}{2}\|\boldsymbol{\theta} - \mathbf{z}\|_2^2 + \mathbf{d}^\top \boldsymbol{\xi} + \frac{\lambda}{2}\|\boldsymbol{\xi}\|_2^2 \quad \text{s.t. } A_{J_{\mathbf{y}}}\boldsymbol{\xi} + B_{J_{\mathbf{y}}}\boldsymbol{\theta} = \mathbf{c}_{J_{\mathbf{y}}}. \tag{27}$$

A rigorous proof of this lemma involves technical arguments from convex analysis, which will be presented in Section B.2 of the supplementary material. The proof of Theorem 3.2, based on Lemma 3.3, will be provided in Section B.3 of the supplementary material.

4. DF of (Penalized) Convex Regression

One important application of Theorem 3.2 is in characterizing DF for the LSE in *multivariate convex regression* (see, e.g., Seijo and Sen 2011). In particular, consider the nonparametric regression problem in (1), where $f : \mathbb{R}^d \to \mathbb{R}$ ($d > 1$) is a convex function and $\mathcal{X} := \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ is the set of design points (with n distinct elements) in $\mathbb{R}^d$. The goal is to estimate $\boldsymbol{\theta}^* = (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n))$. Let $\mathcal{K}_{\text{conv}}$ be the set of all vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n) \in \mathbb{R}^n$ for which there exists a convex function $\psi : \mathbb{R}^d \to \mathbb{R}$ such that $\psi(\mathbf{x}_i) = \theta_i$ for $i = 1, \ldots, n$. It can be shown that $\mathcal{K}_{\text{conv}}$ is a convex cone (see Lemma 2.3 of Seijo and Sen 2011). The multivariate convex LSE is defined as $\widehat{\boldsymbol{\theta}}(\mathbf{y}) := \arg\min_{\boldsymbol{\theta} \in \mathcal{K}_{\text{conv}}} \frac{1}{2}\|\boldsymbol{\theta} - \mathbf{y}\|_2^2$. In fact, Lemma 2.2 from Seijo and Sen (2011) provides the following explicit characterization of $\mathcal{K}_{\text{conv}}$.

Lemma 4.1 (Seijo and Sen 2011). For a vector $\boldsymbol{\theta} \in \mathbb{R}^n$, we have $\boldsymbol{\theta} \in \mathcal{K}_{\text{conv}}$ if and only if there exists a set of n d-dimensional vectors $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n \in \mathbb{R}^d$ such that the following inequalities hold simultaneously:

$$\langle \boldsymbol{\xi}_j, \mathbf{x}_k - \mathbf{x}_j\rangle \le \theta_k - \theta_j, \quad \text{for all } j \ne k \in \{1, \ldots, n\}. \tag{28}$$

Lemma 4.1 is quite intuitive: since f is a multivariate convex function, we have for any pair $\mathbf{x}_k, \mathbf{x}_j \in \mathcal{X}$,

$$f(\mathbf{x}_k) - f(\mathbf{x}_j) \ge \langle g(\mathbf{x}_j), \mathbf{x}_k - \mathbf{x}_j\rangle, \tag{29}$$

where $g(\mathbf{x}_j) \in \partial f(\mathbf{x}_j)$ is a subgradient of the convex function f at $\mathbf{x}_j$. Letting $\boldsymbol{\xi}_j = g(\mathbf{x}_j)$, one can easily see the equivalence between (29) and (28). Using Lemma 4.1, the LSE of multivariate convex regression can be formulated as the following optimization problem (see, e.g., Kuosmanen 2008; Seijo and Sen 2011; Hannah and Dunson 2011; Lim and Glynn 2012):

$$(\widehat{\boldsymbol{\theta}}(\mathbf{y}), \widehat{\boldsymbol{\xi}}(\mathbf{y})) = \underset{\substack{\boldsymbol{\theta} \in \mathbb{R}^n \\ \boldsymbol{\xi} = [\boldsymbol{\xi}_1^\top, \ldots, \boldsymbol{\xi}_n^\top]^\top \in \mathbb{R}^{nd}}}{\arg\min} \frac{1}{2}\|\boldsymbol{\theta} - \mathbf{y}\|_2^2 \quad \text{s.t. } \langle \boldsymbol{\xi}_j, \mathbf{x}_k - \mathbf{x}_j\rangle \le \theta_k - \theta_j, \ \forall\, j \ne k \in \{1, \ldots, n\}, \tag{30}$$

which is a standard linearly constrained quadratic program and can be solved by many off-the-shelf solvers (e.g., SDPT3, Tütüncü, Toh, and Todd 2003). Next, we show that the above optimization problem can be reformulated as a special case of (13) with properly chosen A, B and $\mathbf{c} = \mathbf{0}$, $\mathbf{d} = \mathbf{0}$, and $\lambda = 0$.

Proposition 4.2. The optimization problem for multivariate convex regression in (30) can be formulated as (14) with $p = nd$ and $\boldsymbol{\xi} = [\boldsymbol{\xi}_1^\top, \ldots, \boldsymbol{\xi}_n^\top]^\top \in \mathbb{R}^{nd}$. In this scenario, A in (14) is a $[n(n-1)] \times nd$ matrix and each row of A is indexed by a pair $r = (j, k)$ with $j \ne k \in \{1, \ldots, n\}$ and each column is indexed by a pair $c = (j', s)$ with $j' \in \{1, \ldots, n\}$ and $s \in \{1, \ldots, d\}$. Moreover, we partition A into $[n(n-1)] \times n$ blocks with each block of size $1 \times d$. Let $A_{r,j'}$ be the block of A with row $r = (j, k)$ and column $j' \in \{1, \ldots, n\}$. $A_{r,j'}$ is defined as $A_{r,j'} = \mathbf{x}_k^\top - \mathbf{x}_j^\top$ if $j = j'$ and $A_{r,j'} = \mathbf{0}^\top$ if $j \ne j'$. The corresponding B is a $[n(n-1)] \times n$ matrix and each row of B is indexed by a pair $r = (j, k)$ with $j \ne k \in \{1, \ldots, n\}$ and each column is indexed by $c \in \{1, \ldots, n\}$. Let $B_{r,c}$ be the entry in row $r = (j, k)$ and column c of the matrix B defined as $B_{r,c} = 1$ if $c = j$, $B_{r,c} = -1$ if $c = k$, and $B_{r,c} = 0$ otherwise. The corresponding $\mathbf{c}$ will be an all-zero vector in $\mathbb{R}^{n(n-1)}$.

The proof of Proposition 4.2 is straightforward and thus omitted. Given the matrices A and B defined in Proposition 4.2, one can define the corresponding polyhedron $\mathcal{Q}$ of $(\boldsymbol{\xi}, \boldsymbol{\theta})$ in (11) and it is clear that $\mathcal{K}_{\text{conv}} = \text{Proj}_{\boldsymbol{\theta}}(\mathcal{Q})$, which is a projected convex polyhedron. Given Proposition 4.2, it is straightforward to apply Theorem 3.2 (with $\mathbf{d} = \mathbf{0}$ and $\lambda = 0$) to calculate the DF of the LSE for multivariate convex regression.

Corollary 4.3. For multivariate convex LSE in (30), let the set of tight constraints be $J_{\mathbf{y}} := \{(j,k) : \langle \widehat{\boldsymbol{\xi}}_j, \mathbf{x}_k - \mathbf{x}_j \rangle = \widehat{\theta}_k - \widehat{\theta}_j\}$. Let $I_{\mathbf{y}} \subseteq J_{\mathbf{y}}$ be the index set of maximal independent rows of the matrix $[A_{J_{\mathbf{y}}}, B_{J_{\mathbf{y}}}]$, where A and B are defined in Proposition 4.2. Then for a.e. $\mathbf{y}$, we have $D(\mathbf{y}) = n - |I_{\mathbf{y}}| + \text{rank}(A_{I_{\mathbf{y}}})$ and $\text{df}(\widehat{\boldsymbol{\theta}}(\mathbf{y})) = n - \mathbb{E}[|I_{\mathbf{y}}|] + \mathbb{E}\left[\text{rank}(A_{I_{\mathbf{y}}})\right]$.

The multivariate convex LSE described in (30) tends to overfit the data, especially near the boundary of the convex hull of the design points—the subgradients take large values near the boundary. Thus, we might want to regularize the convex LSE. A natural way to achieve this is to impose bounds on the norm of the subgradients; see, for example, Sen and Meyer (2013) and Lim (2014). In the penalized form this would lead to the following problem:

$$(\widehat{\boldsymbol{\theta}}(\mathbf{y}), \widehat{\boldsymbol{\xi}}(\mathbf{y})) = \underset{\substack{\boldsymbol{\theta} \in \mathbb{R}^n \\ \boldsymbol{\xi} = [\boldsymbol{\xi}_1^\top, \ldots, \boldsymbol{\xi}_n^\top]^\top \in \mathbb{R}^{nd}}}{\arg\min} \frac{1}{2}\|\boldsymbol{\theta} - \mathbf{y}\|_2^2 + \frac{\lambda}{2} \sum_{j=1}^n \|\boldsymbol{\xi}_j\|_2^2 \quad (31)$$
$$\text{s.t. } \langle \boldsymbol{\xi}_j, \mathbf{x}_k - \mathbf{x}_j \rangle \le \theta_k - \theta_j \quad \forall\, j \ne k,$$

which can be formulated as (16) with $p = nd$ and $\boldsymbol{\xi} = [\boldsymbol{\xi}_1^\top, \ldots, \boldsymbol{\xi}_n^\top]^\top \in \mathbb{R}^{nd}$, where A, B and $\mathbf{c}$ are defined in Proposition 4.2. The divergence of the penalized convex regression estimator $\widehat{\boldsymbol{\theta}}(\mathbf{y})$ in (31) can be easily characterized by Theorem 3.2 (with $\mathbf{d} = \mathbf{0}$ and $\lambda > 0$).

Corollary 4.4. For the penalized multivariate convex LSE described in (31), let the set of tight constraints be $J_{\mathbf{y}} := \{(j,k) : \langle \widehat{\boldsymbol{\xi}}_j, \mathbf{x}_k - \mathbf{x}_j \rangle = \widehat{\theta}_k - \widehat{\theta}_j\}$. Let $I_{\mathbf{y}} \subseteq J_{\mathbf{y}}$ be the index set of maximal independent rows of the matrix $[A_{J_{\mathbf{y}}}, B_{J_{\mathbf{y}}}]$, where A and B are defined in Proposition 4.2. Then for a.e. $\mathbf{y}$, we have $D(\mathbf{y}) = n - \text{trace}\left(B_{I_{\mathbf{y}}}^\top \left(B_{I_{\mathbf{y}}} B_{I_{\mathbf{y}}}^\top + \frac{1}{\lambda} A_{I_{\mathbf{y}}} A_{I_{\mathbf{y}}}^\top\right)^{-1} B_{I_{\mathbf{y}}}\right)$ and $\text{df}(\widehat{\boldsymbol{\theta}}(\mathbf{y})) = \mathbb{E}[D(\mathbf{y})]$.

5. DF of (Bounded) Isotonic Regression

Let us consider isotonic regression on a general partially ordered set; see, for example, Robertson, Wright, and Dykstra (1988, Chapter 1). Let $\mathcal{X} := \{x_1, \ldots, x_n\}$ be a set (with n distinct elements) in a metric space with a *partial order*, that is, there exists a binary relation $\lesssim$ over $\mathcal{X}$ that is reflexive ($x \lesssim x$ for all $x \in \mathcal{X}$), transitive ($u, v, w \in \mathcal{X}$, $u \lesssim v$ and $v \lesssim w$ imply $u \lesssim w$), and antisymmetric ($u, v \in \mathcal{X}$, $u \lesssim v$ and $v \lesssim u$ imply $u = v$). Consider (1) where now the real-valued function f is assumed to be *isotonic* with respect to the partial order $\lesssim$, that is, any pair $u, v \in \mathcal{X}$, $u \lesssim v$ implies $f(u) \le f(v)$. This model can be expressed in the sequence form as (8) by letting $\theta_i^* = f(x_i)$ for $i = 1, \ldots, n$. To construct the LSE in this problem, we add *isotonic* constraints on $\boldsymbol{\theta}$, which are of the form $\theta_i \le \theta_j$ if $x_i \lesssim x_j$, for some $i, j \in \{1, \ldots, n\}$. As a special case, let us consider $\mathcal{X} \subset \mathbb{R}$ for the univariate isotonic regression. Assuming without loss of generality that $x_1 \le x_2 \le \cdots \le x_n$, the isotonic constraint set on $\boldsymbol{\theta}$ takes the form of the isotonic cone $\mathcal{M}$ (see Equation (2)) and the LSE is the projection $\widehat{\boldsymbol{\theta}}(\mathbf{y})$ of $\mathbf{y}$ onto $\mathcal{M}$. For the ease of illustration, the isotonic constraints can be represented by an acyclic directed graph $\widetilde{G} = (V, \widetilde{E})$, where $V = \{1, \ldots, n\}$ (corresponding to $\{\theta_i\}_{i=1}^n$) and the set of the directed edges is denoted by

$$\widetilde{E} = \{(i,j) : x_i \lesssim x_j\}. \quad (32)$$

For the univariate isotonic cone $\mathcal{M}$, the edge set $\tilde{E}$ contains $n-1$ edges, where the ith edge runs from node θ_i to θ_{i+1} for $i = 1, \ldots, n-1$, that is, $\tilde{E} = \{(i, i+1) : i = 1, \ldots, n-1\}$.

It is well-known that the projection $\widehat{\boldsymbol{\theta}}(\mathbf{y})$ of $\mathbf{y}$ onto the isotonic constraint set suffers from the *spiking effect*, that is, overfitting near the boundary of the convex hull of the predictor(s) (see Pal 2008; Woodroofe and Sun 1993). However, such monotonic relationships among variables arise naturally in many applications and this has lead to a recent surge of interest in regularized isotonic regression; see, for example, Luss, Rosset, and Shahar (2012), Luss and Rosset (2014), and Wu, Meyer, and Opsomer (2015). Probably the most natural form of regularization involves constraining the range of $\widehat{\boldsymbol{\theta}}(\mathbf{y})$, that is, $\max_i \widehat{\theta}_i - \min_i \widehat{\theta}_i$; this leads to *bounded isotonic regression*. More specifically, when the range of f is known to be bounded (from above) by some $\gamma \ge 0$, we can impose this boundedness restriction of f by adding the boundedness constraints and the corresponding bounded isotonic LSE can be defined as follows.

Definition 5.1. The bounded isotonic LSE (with boundedness parameter γ) is defined as the projection estimator $\widehat{\boldsymbol{\theta}}_\gamma(\mathbf{y}) := \arg\min_{\boldsymbol{\theta} \in \mathcal{C}} \|\mathbf{y} - \boldsymbol{\theta}\|_2^2$, where the constraint set is

$$\mathcal{C} := \Big\{\boldsymbol{\theta} \in \mathbb{R}^n : \theta_i \le \theta_j \,\forall\, (i,j) \in \widetilde{E},\ \theta_i \le \theta_j + \gamma, \quad i \in \max(V), j \in \min(V), i \ne j\Big\}. \quad (33)$$

Here, $\max(V)$ and $\min(V)$ are the maximal and minimal sets of V with respect to this partial order:

$$\max(V) = \{i \in V : \widetilde{n}^+(i) = \emptyset\}$$
$$\text{and} \quad \min(V) = \{i \in V : \widetilde{n}^-(i) = \emptyset\},$$

where for any node i, $\widetilde{n}^+(i) := \{j \in V : (i,j) \in \widetilde{E}\}$ is the set of elements that are "greater than i" with respect to the partial order (i.e., successors of i), and $\widetilde{n}^-(i) := \{j \in V : (j,i) \in \widetilde{E}\}$ is the set of elements that are "smaller than i" (i.e., predecessors of i).

In Definition 5.1, both $\max(V)$ and $\min(V)$ must be nonempty for any nonempty partially ordered set. This is because $\widetilde{G}$ is an acyclic directed graph where there always exist nodes with no successor and nodes with no predecessor. We also note that $\max(V)$ and $\min(V)$ might overlap, for example, when there exist nodes that cannot be compared with any other nodes under the given partial order. For each $i \in \max(V)$ and $j \in \min(V)$ with $i \ne j$, we add a constraint $\theta_i \le \theta_j + \gamma$ to impose the boundedness restriction on the range of f.

Similar to the unbounded case, we can represent the constraints in (33) by a graph $G = (V, E)$ where $V = \{1, \ldots, n\}$ and

$$E := \widetilde{E} \cup \{(i, j) : i \in \max(V), j \in \min(V), i \neq j\}.$$

As a special case, for univariate bounded isotonic regression, the constraint set $\mathcal{C}$ in (33) becomes $\{\boldsymbol{\theta} \in \mathbb{R}^n : \theta_1 \leq \cdots \leq \theta_n,\ \theta_n - \theta_1 \leq \gamma\}$ and the corresponding edge set is $E = \{(i, i+1), i = 1, \ldots, n-1\} \cup \{(n, 1)\}$.

To compute the DF of bounded isotonic LSE $\widehat{\boldsymbol{\theta}}_\gamma(\mathbf{y})$, first notice that the set $\mathcal{C}$ can be easily represented as a convex polyhedron of the form in (4). We note that as compared to unbounded isotonic regression, the $\mathcal{C}$ in (33) is a convex polyhedron rather than a polyhedral cone due to the additional boundedness constraints. Given the fact that bounded isotonic LSE is a projection estimator onto a convex polyhedron, Theorem 3.2 (with $\mathbf{d} = \mathbf{0}$, $\lambda = 0$ and $A = 0$) can be used to compute its DF. Instead of directly applying Theorem 3.2 in its original form, we draw some interesting connections to graph theory, which also leads to a faster computation of the divergence. In particular, let $\omega(G)$ denote the number of connected components of the undirected version of the graph $G = (V, E)$ (removing the directions of edges in G), that is, the number of maximal connected subgraphs of G. The divergence of $\widehat{\boldsymbol{\theta}}_\gamma(\mathbf{y})$ can be characterized using the number of connected components of a subgraph of G as shown in the following proposition (see the proof in Section C.1 in the supplementary material).

Proposition 5.2. The bounded isotonic constraint set $\mathcal{C}$ defined in (33) is a convex polyhedron in the form of (4), where $m = |E|$ and $B \in \mathbb{R}^{|E| \times n}$ is defined as (the rows of B are indexed by the edge set)

$$B_{e,i} = \begin{cases} 1 & \text{if } e = (i, j) \in E \text{ for some } j \neq i \\ -1 & \text{if } e = (j, i) \in E \text{ for some } j \neq i \\ 0 & \text{otherwise} \end{cases} \tag{34}$$

and $\mathbf{c} = (c_e)_{e=1}^{|E|} \in \mathbb{R}^{|E|}$ is defined as

$$c_e = \begin{cases} \gamma & \text{if } e = (i, j) \in E \text{ for } i \in \max(V), j \in \min(V) \\ 0 & \text{otherwise.} \end{cases} \tag{35}$$

Let B_e be the eth row of B and $J_{\mathbf{y}} := \{e \in E : B_e \widehat{\boldsymbol{\theta}}_\gamma(\mathbf{y}) = c_e\}$. Further, let $G_{J_{\mathbf{y}}}$ be the subgraph of G with the edge set $J_{\mathbf{y}}$. The divergence of $\widehat{\boldsymbol{\theta}}_\gamma(\mathbf{y})$ is the number of connected components of $G_{J_{\mathbf{y}}}$ for a.e. $\mathbf{y}$, that is, $D(\mathbf{y}) = \omega(G_{J_{\mathbf{y}}})$, and therefore, $\mathrm{df}(\widehat{\boldsymbol{\theta}}_\gamma(\mathbf{y})) = \mathbb{E}[\omega(G_{J_{\mathbf{y}}})]$.

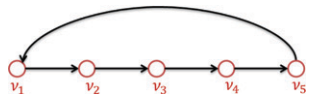
The characterization of divergence in Proposition 5.2 not only has interesting connections to graph theory but also leads to a computationally fast procedure to compute the divergence. In fact, it is easy to compute $\omega(G_{J_{\mathbf{y}}})$ using either breadth-first or depth-first search in linear time in n, which is *computationally much cheaper* than directly calculating the rank of $B_{J_{\mathbf{y}}}$ in Proposition 2.1. To facilitate the understanding of Proposition 5.2, we provide a toy example. Consider the following bounded isotonic constraint set with $n = 5$:

$$\mathcal{C} = \{\boldsymbol{\theta} \in \mathbb{R}^n : \theta_1 \leq \cdots \leq \theta_n, \text{ and } \theta_n - \theta_1 \leq \gamma\}. \tag{36}$$

The set $\mathcal{C}$ can be represented as $\mathcal{C} = \{\boldsymbol{\theta} \in \mathbb{R}^n : B\boldsymbol{\theta} \leq \mathbf{c}\}$ where B is shown in Figure 1(a) and $\mathbf{c}$ only has one nonzero element

$$B = \begin{pmatrix} +1 & -1 & 0 & 0 & 0 \\ 0 & +1 & -1 & 0 & 0 \\ 0 & 0 & +1 & -1 & 0 \\ 0 & 0 & 0 & +1 & -1 \\ -1 & 0 & 0 & 0 & +1 \end{pmatrix}$$

(a) Matrix B

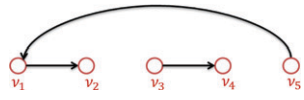


(b) Graph G

Figure 1. The matrix B and the induced graph G.

$$B_{J_{\mathbf{y}}} = \begin{pmatrix} +1 & -1 & 0 & 0 & 0 \\ 0 & 0 & +1 & -1 & 0 \\ -1 & 0 & 0 & 0 & +1 \end{pmatrix}$$

(a) Matrix $B_{J_{\mathbf{y}}}$



(b) Graph $G_{J_{\mathbf{y}}}$

Figure 2. The matrix $B_{J_{\mathbf{y}}}$ and the induced graph $G_{J_{\mathbf{y}}}$.

at the nth position, that is, $c_n = \gamma$. The graph G induced from B, which has only one connected component (i.e., $\omega(G) = 1$), is shown in Figure 1(b).

Now suppose that we have $\widehat{\theta}_{\gamma,1} = \widehat{\theta}_{\gamma,2} < \widehat{\theta}_{\gamma,3} = \widehat{\theta}_{\gamma,4} < \widehat{\theta}_{\gamma,5}$ and $\widehat{\theta}_{\gamma,5} = \widehat{\theta}_{\gamma,1} + \gamma$. Then $J_{\mathbf{y}} = \{1, 3, 5\}$ and the corresponding $B_{J_{\mathbf{y}}}$ and $G_{J_{\mathbf{y}}}$ are presented in Figure 2. From Figure 2, $G_{J_{\mathbf{y}}}$ has 2 connected components $\{\theta_1, \theta_2, \theta_5\}$ and $\{\theta_3, \theta_4\}$ and thus $D(\mathbf{y}) = \omega(G_{J_{\mathbf{y}}}) = 2$. It is of interest to compare this with the univariate unbounded isotonic regression example where the divergence of $\widehat{\boldsymbol{\theta}}_\gamma(\mathbf{y})$ would be 3 (i.e., the number of distinct values of $\widehat{\theta}_i$'s; see Proposition 1 from Meyer and Woodroofe 2000) instead of 2.

Using exactly the same proof technique as that of Proposition 5.2, we can easily derive the following result for the DF of *unbounded isotonic regression* on a partially ordered set. In particular, recall the unbounded isotonic cone $\mathcal{M} = \{\boldsymbol{\theta} \in \mathbb{R}^n : \theta_i \leq \theta_j, \forall (i, j) \in \widetilde{E}\}$ where $\widetilde{E}$ is defined in (32) and the corresponding LSE $\widehat{\boldsymbol{\theta}}(\mathbf{y}) = \arg\min_{\boldsymbol{\theta} \in \mathcal{M}} \|\boldsymbol{\theta} - \mathbf{y}\|_2^2$. The cone $\mathcal{M}$ can be represented as $\mathcal{M} = \{\boldsymbol{\theta} \in \mathbb{R}^n : B\boldsymbol{\theta} \leq \mathbf{0}\}$, where $B \in \mathbb{R}^{|\widetilde{E}| \times n}$ is defined similarly as in (34) (replacing E in (34) by $\widetilde{E}$). Let B_e be the eth row of B, $J_{\mathbf{y}} := \{e \in \widetilde{E} : B_e \widehat{\boldsymbol{\theta}}(\mathbf{y}) = b_e\}$ and $\widetilde{G}_{J_{\mathbf{y}}}$ be the subgraph of $\widetilde{G}$ with the edge set $J_{\mathbf{y}}$. The divergence of $\widehat{\boldsymbol{\theta}}(\mathbf{y})$ for unbounded isotonic regression is $D(\mathbf{y}) = \omega(\widetilde{G}_{J_{\mathbf{y}}})$, and therefore, $\mathrm{df}(\widehat{\boldsymbol{\theta}}(\mathbf{y})) = \mathbb{E}[\omega(\widetilde{G}_{J_{\mathbf{y}}})]$.

In addition to characterizing the DF for general bounded isotonic regression, we also show a useful property of the divergence $D_\gamma(\mathbf{y})$ in Theorem 5.4 (where we make the dependence on the model complexity parameter γ explicit). In particular, we prove that the divergence $D_\gamma(\mathbf{y})$ (and thus the DF) is nondecreasing in γ. To show this we first present an important connection between the solution of bounded isotonic regression and that of unbounded isotonic regression (which can be viewed as a special case of bounded isotonic regression with $\gamma = +\infty$). This result is of independent interest by itself.

We start with some notation. It is well known that the LSE for *unbounded* isotonic regression $\widehat{\boldsymbol{\theta}}$ has a group-constant structure (here $\mathbf{y}$ is suppressed for notational simplicity). That is, there exists a partition $U_1, U_2, \ldots, U_r$ of $V = \{1, \ldots, n\}$ (i.e., U_s's are disjoint and $V = \bigcup_{s=1}^r U_s$) such that $\widehat{\theta}_i = \bar{\theta}_s$ for some

value $\bar{\theta}_s$ for each $i \in U_s$, for $1 \le s \le r$. Moreover, without loss of generality, we assume that $\bar{\theta}_1 < \bar{\theta}_2 < \cdots < \bar{\theta}_r$. Let $\widehat{\boldsymbol{\theta}}_\gamma$ be the LSE for bounded isotonic regression with the boundedness parameter γ. The next proposition shows that $\widehat{\boldsymbol{\theta}}_\gamma$ can be obtained by appropriately thresholding $\widehat{\boldsymbol{\theta}}$.

Proposition 5.3. Let $|U_s| = k_s$ for $s = 1, \ldots, r$ and $H(L, \gamma)$ be a function on $\mathbb{R}^2$ defined as

$$H(L,\gamma) := \sum_{s=1}^{r} k_s \left(L - \bar{\theta}_s\right)_+ + \sum_{s=1}^{r} k_s \left(L + \gamma - \bar{\theta}_s\right)_-, \quad (37)$$

where $(x)_+ = \max\{x, 0\}$ and $(x)_- = \min\{x, 0\}$. For any given γ with $\bar{\theta}_r - \bar{\theta}_1 \ge \gamma \ge 0$, $H(L, \gamma)$ is a continuous and strictly increasing function of L. Moreover, $\lim_{L\to-\infty} H(L, \gamma) = -\infty$ and $\lim_{L\to+\infty} H(L, \gamma) = +\infty$ so that there exists a unique L_γ satisfying $H(L_\gamma, \gamma) = 0$. Then, we have

$$\widehat{\theta}_{\gamma,i} = \max(L_\gamma, \min(L_\gamma + \gamma, \bar{\theta}_s)), \text{ for all } i \in U_s. \quad (38)$$

Moreover, L_γ is nonincreasing in γ.

Proposition 5.3 also provides an efficient way to compute the LSE for bounded isotonic regression. In particular, one can first compute $\widehat{\boldsymbol{\theta}}$ by solving the corresponding unbounded isotonic regression, which can be efficiently computed by using existing off-the-shelf solvers (e.g., SDPT3, Tütüncü, Toh, and Todd 2003). Given $\widehat{\boldsymbol{\theta}}$, one obtains the values of $\bar{\theta}_s$ and k_s for $s = 1, \ldots, r$, which are necessary for constructing the function in (37). If $\gamma > \bar{\theta}_r - \bar{\theta}_1$, the boundedness constraint will be noneffective and $\widehat{\boldsymbol{\theta}}_\gamma = \widehat{\boldsymbol{\theta}}$. On the other hand, if $\bar{\theta}_r - \bar{\theta}_1 \ge \gamma \ge 0$, since $H(L, \gamma)$ is a continuous and strictly increasing function of L, one can use *bisection search* to compute L_γ such that $H(L_\gamma, \gamma) = 0$. Then by (38), we threshold $\widehat{\boldsymbol{\theta}}$ to obtain $\widehat{\boldsymbol{\theta}}_\gamma$: for each U_s, if $\bar{\theta}_s < L_\gamma$, $\widehat{\theta}_{\gamma,i} = L_\gamma$ for all $i \in U_s$; if $\bar{\theta}_s > L_\gamma + \gamma$, $\widehat{\theta}_{\gamma,i} = L_\gamma + \gamma$ for all $i \in U_s$; otherwise $\widehat{\theta}_{\gamma,i}$ is set to $\bar{\theta}_s$ for all $i \in U_s$.

The key to the proof of the above result is to find appropriate values of dual variables such that the primal solutions in (38) and dual solutions together satisfy the KKT condition of $\min_{\boldsymbol{\theta}\in\mathcal{C}} \|\mathbf{y} - \boldsymbol{\theta}\|_2^2$ with $\mathcal{C}$ in (33). We achieve this by designing a *transportation problem*, which is a classical problem in operations research (see, e.g., Chapter 14 in Dantzig 1959). The dual solutions are constructed based on the solution of such a transportation problem. Please refer to the proof in Section C.2 in the supplementary material for details.

Combining Proposition 5.3 and Proposition 5.2, we obtain the following theorem which shows the monotonicity of DF in terms of the boundedness parameter γ in bounded isotonic regression (see Section C.3 in the supplementary material for the proof).

Theorem 5.4. For any given $\mathbf{y} \in \mathbb{R}^n$ the divergence of $\widehat{\boldsymbol{\theta}}_\gamma(\mathbf{y})$ is nondecreasing in γ. This implies that $\mathrm{df}(\widehat{\boldsymbol{\theta}}_\gamma(\mathbf{y}))$ is nondecreasing in γ.

6. Additive TV Regression and Other Applications

In this section, we apply our main result to derive the DF for additive TV regression (see Example 3 in the Introduction) and ℓ_∞-regularized group Lasso. Moreover, our main result (Theorem 3.2) also yields, as special cases, known results on DF of many popular estimators, for example, *Lasso and generalized Lasso*, *linear regression*, and *ridge regression*. Due to space constraints, we illustrate these applications in Section D.3 of the supplementary material; the proofs of the results in this section are also provided in Section D, supplementary material.

6.1. Additive Generalized TV Regression

For each response y_i and input $\mathbf{x}_i = (x_{i1}, \ldots, x_{id})$, where $1 \le i \le n$, the additive model assumes that $\mathbb{E}(y_i|\mathbf{x}_i) = \sum_{j=1}^d f_j(x_{ij})$. Let $\theta_{ji}^* = f_j(x_{ij})$ and $\boldsymbol{\theta}_j^* = (\theta_{j1}, \ldots, \theta_{jn})$, where it is typically assumed that each $\boldsymbol{\theta}_j$ has zero mean (i.e., $\mathbf{1}^\top \boldsymbol{\theta}_j = 0$). Petersen, Witten, and Simon (2016) proposed the following additive TV regularizer. Let $D \in \mathbb{R}^{(n-1)\times n}$ be the discrete first derivative matrix (i.e., the ith row of D only contains two nonzero elements: $D_{i,i} = 1$ and $D_{i,i+1} = -1$) and $P_j \in \mathbb{R}^{n\times n}$ be the permutation matrix that orders the jth feature from least to greatest. The estimation of $\{\boldsymbol{\theta}_j^*\}_{j=1}^d$ in an additive TV regularized regression takes the form:

$$\begin{aligned}\{\widehat{\theta}_0, \{\widehat{\boldsymbol{\theta}}_j\}_{j=1}^d\} = \arg\min_{\{\boldsymbol{\theta}_j\}_{j=1}^d} & \frac{1}{2}\Big\|\mathbf{y} - \sum_{j=1}^d \boldsymbol{\theta}_j - \theta_0 \mathbf{1}\Big\|_2^2 \\ & + \tau \sum_{j=1}^d \|DP_j\boldsymbol{\theta}_j\|_1 \\ \text{s.t.} \quad & \mathbf{1}^\top \boldsymbol{\theta}_j = 0, \quad 1 \le j \le d.\end{aligned}$$

The penalty $\|DP_j\boldsymbol{\theta}_j\|_1$ encourages $\boldsymbol{\theta}_j$ to be piecewise constant with a small number of jumps, depending on the regularization τ. In fact, instead of using the discrete first derivative matrix D, we could impose a higher order smoothness for each component function f_j. More precisely, one can use a higher order discrete difference matrix D_j for each f_j; in the sequel we will consider this more general setup. For example, the second order differencing matrix produces piecewise affine fits, with a few number of kink points. The specific form of higher order discrete difference matrix is given in Equation (41) of Tibshirani (2014). Let us denote D_jP_j by $Q_j \in \mathbb{R}^{n_j\times n}$ for notational simplicity, and we consider the following *additive generalized TV regression*:

$$\begin{aligned}\{\widehat{\theta}_0, \{\widehat{\boldsymbol{\theta}}_j\}_{j=1}^d\} = \arg\min_{\{\boldsymbol{\theta}_j\}_{j=1}^d} & \frac{1}{2}\Big\|\mathbf{y} - \sum_{j=1}^d \boldsymbol{\theta}_j - \theta_0 \mathbf{1}\Big\|_2^2 \\ & + \tau \sum_{j=1}^d \|Q_j\boldsymbol{\theta}_j\|_1 \quad (39) \\ \text{s.t.} \quad & \mathbf{1}^\top \boldsymbol{\theta}_j = 0, \quad 1 \le j \le d.\end{aligned}$$

Let the $\widehat{\boldsymbol{\theta}}(\mathbf{y}) := \sum_{j=1}^d \widehat{\boldsymbol{\theta}}_j(\mathbf{y}) + \widehat{\theta}_0(\mathbf{y})\mathbf{1}$ be the estimated function values at the design points. To characterize its divergence,

we rewrite the optimization problem in (40) as

$$\begin{aligned}&(\widehat{\boldsymbol{\theta}}(\mathbf{y}), \{\widehat{\boldsymbol{\theta}}_j(\mathbf{y})\}_{j=1}^d, \widehat{\theta}_0(\mathbf{y}), \{\widehat{\boldsymbol{\gamma}}_j(\mathbf{y})\}_{j=1}^d)\\ &\quad\in \operatorname*{arg\,min}_{\boldsymbol{\theta},\boldsymbol{\theta}_j,\theta_0,\boldsymbol{\gamma}_j} \frac{1}{2}\|\boldsymbol{\theta}-\mathbf{y}\|_2^2 + \sum_{j=1}^d \tau \mathbf{1}^\top \boldsymbol{\gamma}_j \qquad (40)\\ &\text{s.t. } \boldsymbol{\theta} - \sum_{j=1}^d \boldsymbol{\theta}_j - \theta_0 \mathbf{1} \le \mathbf{0}, \quad -\boldsymbol{\theta} + \sum_{j=1}^d \boldsymbol{\theta}_j + \theta_0 \mathbf{1} \le \mathbf{0}\\ &\quad Q_j \boldsymbol{\theta}_j - \boldsymbol{\gamma}_j \le \mathbf{0}, \quad -Q_j \boldsymbol{\theta}_j - \boldsymbol{\gamma}_j \le \mathbf{0}\\ &\quad \mathbf{1}^\top \boldsymbol{\theta}_j \le 0, \quad -\mathbf{1}^\top \boldsymbol{\theta}_j \le 0, \quad 1 \le j \le d.\end{aligned}$$

With some algebraic manipulations, we show that the optimization in (40) is a special case of (13) with a linear perturbation term $\mathbf{d}^\top \boldsymbol{\xi}$ and $\lambda = 0$ (in particular, in the form of (15)); see the proof in the supplementary material for the details. We then apply Theorem 3.2 to obtain the following result on the DF for $\widehat{\boldsymbol{\theta}}(\mathbf{y})$. In our proof, we also verify that the condition in Theorem 3.2 (i.e., $-\mathbf{d} = A^\top \mathbf{u}$ for some $\mathbf{u} \ge \mathbf{0}$) indeed holds.

Proposition 6.1. For the estimator $\widehat{\boldsymbol{\theta}}(\mathbf{y}) = \sum_{j=1}^d \widehat{\boldsymbol{\theta}}_j(\mathbf{y}) + \widehat{\theta}_0(\mathbf{y})\mathbf{1}$ in (40), the divergence of $\widehat{\boldsymbol{\theta}}(\mathbf{y})$ is,

$$D(\mathbf{y}) = \dim(\operatorname{span}\{\mathbf{1}_{n\times 1}, \ker(K_1), \ldots, \ker(K_d)\}),$$

where, for $j = 1, \ldots, d$, $K_j = \begin{pmatrix} Q_0^j \\ \mathbf{1}_{1\times n} \end{pmatrix}$, Q_0^j is the sub-matrix of Q_j consisting of rows $\mathbf{q}_{ji}$ $(1 \le i \le n_j)$ of Q_j such that $\mathbf{q}_{ji}^\top \widehat{\boldsymbol{\theta}}_j(\mathbf{y}) = 0$ and $\ker(K_j) := \{\mathbf{x} \in \mathbb{R}^n : Q_0^j \mathbf{x} = \mathbf{0} \text{ and } \mathbf{1}_{1\times n}\mathbf{x} = 0\}$ is the kernel of $K_j = \begin{pmatrix} Q_0^j \\ \mathbf{1}_{1\times n} \end{pmatrix}$. Further, $\operatorname{df}(\widehat{\boldsymbol{\theta}}(\mathbf{y})) = \mathbb{E}(D(\mathbf{y}))$.

Remark 6.1. For each j, the matrix K_j can be easily constructed by checking if $\mathbf{q}_{ji}^\top \widehat{\boldsymbol{\theta}}_j(\mathbf{y}) = 0$ for $1 \le i \le n_j$. After obtaining K_j, the basis for the null space $\ker(K_j)$ can be easily computed by transforming K_j into the reduced row echelon form using Gaussian elimination (note that one can use the *null* function in Matlab or the *Null* function in R to compute the basis of $\ker(K_j)$). Then, we construct a matrix using the basis of $\ker(K_j)$ for each j and $\mathbf{1}_{n\times 1}$ as its column so that $D(\mathbf{y})$ can be computed as the rank of this matrix.

6.2. ℓ_∞-regularized Group Lasso

Let $\mathbb{G} = \{\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_l\}$ be a partition of $\{1, 2, \ldots, d\}$. Each element $\mathcal{G} \in \mathbb{G}$ represents a group of variables. The ℓ_∞-regularized group Lasso estimator can be formulated as the following optimization problem (Zhao, Rocha, and Yu 2009; Negahban and Wainwright 2011):

$$\widehat{\boldsymbol{\beta}}(\mathbf{y}) \in \operatorname*{arg\,min}_{\boldsymbol{\beta}\in\mathbb{R}^d} \frac{1}{2}\|\mathbf{y} - X\boldsymbol{\beta}\|_2^2 + \tau \sum_{\mathcal{G}\in\mathbb{G}} \|\boldsymbol{\beta}_{\mathcal{G}}\|_\infty, \qquad (41)$$

where $\boldsymbol{\beta}_{\mathcal{G}}$ is the subvector of $\boldsymbol{\beta}$ consisting of the coordinates indexed by the elements in $\mathcal{G}$. We can easily see that (41) is a special case of the optimization problem (13). In fact, by introducing the variable $\boldsymbol{\gamma} \in \mathbb{R}^l$ and letting $\boldsymbol{\theta} = X\boldsymbol{\beta}$, (41) can be equivalently reformulated as

$$\begin{aligned}(\widehat{\boldsymbol{\theta}}(\mathbf{y}), \widehat{\boldsymbol{\beta}}(\mathbf{y}), \widehat{\boldsymbol{\gamma}}(\mathbf{y})) \in \operatorname*{arg\,min}_{\boldsymbol{\theta},\boldsymbol{\beta},\boldsymbol{\gamma}} &\frac{1}{2}\|\boldsymbol{\theta} - \mathbf{y}\|_2^2 + \tau \mathbf{1}^\top \boldsymbol{\gamma} \qquad (42)\\ \text{s.t. } &X\boldsymbol{\beta} - \boldsymbol{\theta} \le \mathbf{0}, \quad -X\boldsymbol{\beta} + \boldsymbol{\theta} \le \mathbf{0}\\ &\boldsymbol{\beta}_{\mathcal{G}_j} - \gamma_j \mathbf{1}_{|\mathcal{G}_j|} \le \mathbf{0}, \quad -\boldsymbol{\beta}_{\mathcal{G}_j} - \gamma_j \mathbf{1}_{|\mathcal{G}_j|} \le \mathbf{0}.\end{aligned}$$

By setting $\boldsymbol{\xi} = (\boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top)^\top$ and defining E as the $d \times l$ matrix with $E_{ij} = 1$ if $i \in \mathcal{G}_j$ and $E_{ij} = 0$ otherwise, (42) is a special case of (13) with

$$\begin{aligned}&\mathbf{d} = (\mathbf{0}_{1\times d}, \tau \mathbf{1}_{1\times l})^\top, \ \lambda = 0,\\ &A = \begin{pmatrix} X & \mathbf{0}_{n\times l} \\ -X & \mathbf{0}_{n\times l} \\ I_d & -E \\ -I_d & -E \end{pmatrix}, \quad B = \begin{pmatrix} -I_n \\ I_n \\ \mathbf{0}_{d\times n} \\ \mathbf{0}_{d\times n} \end{pmatrix}, \ \mathbf{c} = \mathbf{0}. \qquad (43)\end{aligned}$$

In the next corollary, we characterize the DF of the ℓ_∞-regularized group Lasso estimator using Theorem 3.2.

Corollary 6.2. In the ℓ_∞-regularized group Lasso problem described in (41) and (42), for a.e. $\mathbf{y} \in \mathbb{R}^n$, $\operatorname{df}(\widehat{\boldsymbol{\theta}}(\mathbf{y})) = \operatorname{df}(X\widehat{\boldsymbol{\beta}}(\mathbf{y})) = \mathbb{E}[\operatorname{rank}(X_{J_0^c})]$, where

$$J_0 = \Big\{ i \in \{1, \ldots, d\} : i \in \mathcal{G}_j, \widehat{\beta}_i(\mathbf{y}) = \|\widehat{\boldsymbol{\beta}}_{\mathcal{G}_j}(\mathbf{y})\|_\infty \text{ for some } j \in \{1, 2, \ldots, l\} \Big\},$$

and J_0^c is the complement set of J_0 and $X_{J_0^c}$ consists of the columns of X indexed by J_0^c.

7. Application: SURE and the Choice of Tuning Parameters

Consider the formulation of the problem posited in (8). For notational simplicity, we will use λ to denote the tuning parameter in the regularized/constrained LSE $\widehat{\boldsymbol{\theta}}_\lambda(\mathbf{y})$ (we highlight the dependence of $\widehat{\boldsymbol{\theta}}(\mathbf{y})$ on λ in this section). For example, in bounded isotonic regression the tuning parameter is the choice of the range of $\boldsymbol{\theta}$ (i.e., the parameter γ in (33)); in penalized convex regression (see Equation (31)) the estimator depends on the tuning parameter λ on the norm of the subgradients.

In this section we use SURE to choose the tuning parameter λ. Let

$$L_n(\lambda) = \|\widehat{\boldsymbol{\theta}}_\lambda(\mathbf{y}) - \boldsymbol{\theta}^*\|_2^2 \qquad (44)$$

denote the loss in estimating $\boldsymbol{\theta}^*$ by $\widehat{\boldsymbol{\theta}}_\lambda(\mathbf{y})$. We would ideally like to choose λ by minimizing $L_n(\cdot)$. Let $\lambda^* := \arg\min_{\lambda\ge 0} L_n(\lambda)$. We note that λ^* is a random quantity as $L_n(\lambda)$ is random. Of course, we cannot compute λ^* as we do not know $\boldsymbol{\theta}^*$. However, we can minimize an (unbiased) estimator of L_n, assuming σ is known, as described below. Let

$$U_n(\lambda) := \|\mathbf{y} - \widehat{\boldsymbol{\theta}}_\lambda(\mathbf{y})\|_2^2 + 2\sigma^2 D(\widehat{\boldsymbol{\theta}}_\lambda(\mathbf{y})) - n\sigma^2, \qquad (45)$$

where $D(\widehat{\boldsymbol{\theta}}_\lambda(\mathbf{y}))$ denotes the divergence of $\widehat{\boldsymbol{\theta}}_\lambda(\mathbf{y})$. It is well known that for all $\lambda \ge 0$, $\mathbb{E}[U_n(\lambda)] = \mathbb{E}[L_n(\lambda)]$; see Stein

(1981) (also see Proposition 2 of Meyer and Woodroofe 2000). The quantity U_n in (45) is usually called the SURE. Let

$$\widehat{\lambda} := \arg\min_{\lambda \geq 0} U_n(\lambda) \tag{46}$$

be the minimizer of $U_n(\lambda)$, which can be computed from the data (if σ^2 is assumed known). Note that here we would need to compute the divergence of $\widehat{\boldsymbol{\theta}}_\lambda(\mathbf{y})$, which we can calculate using the results in the previous sections.

We empirically study the behavior of the ratio $L_n(\widehat{\lambda})/L_n(\lambda^*)$ for bounded isotonic regression and penalized convex regression. We also compare the performance of different tuning parameter selection methods—SURE and cross-validation—including the no-tuning parameter approach (e.g., the standard unbounded isotonic regression and un-penalized convex regression) for these two problems.

In Sections 7.1 and 7.2, we provide simulation studies when the true value of the noise variance σ^2 is assumed known for SURE. When σ^2 is known, the SURE method significantly outperforms its competitors. However, we note that the CV method does not require any knowledge of σ^2. In Section 7.3, we estimate σ^2 using an approach proposed in Meyer and Woodroofe (2000). In this case, the performance of SURE and CV are comparable but CV is computationally more expensive than SURE.

7.1. Bounded Isotonic Regression

We generate n iid design points $\mathbf{x}_i \sim \text{Unif}[0,1]^d$, for $i = 1,\ldots,n$. We set the regression function $f : \mathbb{R}^d \to \mathbb{R}$ to be $f(\mathbf{x}) = \|\mathbf{x}\|_2^2$. Recall that $\boldsymbol{\theta}^* = (f(\mathbf{x}_1),\ldots,f(\mathbf{x}_n))$, which is a bounded vector (since $\|\mathbf{x}\|_2^2 \leq d$) and satisfies $\theta_i^* \leq \theta_j^*$ whenever $\mathbf{x}_i \leq \mathbf{x}_j$. We generate the response y_i, for $i = 1,\ldots,n$, according to model (1) with $\sigma^2 = 1$.

Since the true regression function f is a bounded isotonic function, we estimate $\boldsymbol{\theta}^*$ by minimizing $\|\boldsymbol{\theta} - \mathbf{y}\|_2^2$ subject to the following constraints. For each pair (i,j), we put an isotonic constraint $\theta_i \leq \theta_j$ whenever $\mathbf{x}_i \leq \mathbf{x}_j$. We further add one additional *boundedness constraint* $\max\theta_i - \min\theta_i \leq \lambda$, where λ is the tuning parameter (i.e., the parameter γ in (33)). For each given λ, we obtain the LSE $\widehat{\boldsymbol{\theta}}_\lambda(\mathbf{y})$.

We demonstrate the performance of the selected parameter $\widehat{\lambda}$ using SURE. In particular, we compute the ratio $L_n(\widehat{\lambda})/L_n(\lambda^*)$, where $\widehat{\lambda}$ is selected by (46) (we call this the *SURE ratio*). We compare the SURE ratio to the so-called *CV ratio*, where the boundedness parameter is selected by 5-fold cross-validation. We note that when implementing the CV method, for a given training set $\mathcal{T}_{\text{tr}}$, the estimated function value at a point $\mathbf{x}$ is set to $\widehat{f}(\mathbf{x}) := \min_{\mathbf{x}_i \in \mathcal{T}_{\text{tr}}: \mathbf{x}_i \geq \mathbf{x}} \widehat{\theta}_{\lambda,i}$, where $\widehat{\theta}_{\lambda,i}$ the estimated function value at the training data point $\mathbf{x}_i$ obtained from the bounded isotonic LSE. Such a way of extending the estimated function values (on the training set) to new data points ensures that the extended function is monotone and bounded; this extension has also been used by other authors (see, e.g., Chatterjee, Guntuboyina, and Sen 2018). We also compare the performance of the bounded isotonic LSE with the unbounded LSE where we do not include the boundedness constraint $\max\theta_i - \min\theta_i \leq \lambda$ (or equivalently, set $\lambda = +\infty$ and compute $L_n(\infty)/L_n(\lambda^*)$).

We set $d = 2, 5, 7, 10$ and for each fixed d, we vary the sample size $n = 100, 200, 500, 1000, 2000$ and compute the SURE, CV and unbounded ratios over 100 independent replications and plot the results in Figure 3. From Figure 3 one can see that the SURE ratios are, in general, much smaller than the unbounded ratios, illustrating the usefulness of including the boundedness constraint in isotonic regression. When the dimension is very small (e.g., $d = 2$) the CV ratio slightly outperforms the SURE ratio; while for larger d (e.g., $d = 7$ or $d = 10$) the SURE based method significantly outperforms the CV approach. Moreover, for larger sample sizes n, the SURE ratios are close to 1 indicating that the bounded LSE tuned via SURE performs as good as the bounded LSE with oracle tuning.

7.2. Penalized Multivariate Convex Regression

We generate n iid design points $\mathbf{x}_i \sim \text{Unif}[-1,1]^d$, for $i = 1,\ldots,n$. We set the convex regression function $f : \mathbb{R}^d \to \mathbb{R}$ to be $f(\mathbf{x}) = \|\mathbf{x}\|_2^2$, which is symmetric around $\mathbf{0}$. We generate the response y_i, for $i = 1,\ldots,n$, according to model (1) with $\sigma = 0.5$. Let $\boldsymbol{\theta}^* = (f(\mathbf{x}_1),\ldots,f(\mathbf{x}_n))$. We estimate $\boldsymbol{\theta}^*$ by solving the penalized multivariate convex regression problem described in (31) using the SDPT3 package (Tütüncü, Toh, and Todd 2003). We note that since the optimization problem for penalized multivariate convex regression (in (31)) has a lot of constraints and many variables (i.e., $n(n-1)$ constraints and nd variables), we only consider smaller sample sizes (n) in our simulation experiments. Nevertheless, a smaller n is still sufficient to demonstrate the superior performance of the estimator tuned by minimizing SURE. In particular, we consider $d = 4$ and 10, $n = 100$ and 500, and compute the SURE ratio $L_n(\widehat{\lambda})/L_n(\lambda^*)$, where $\widehat{\lambda}$ is defined as in (46). We compare the SURE ratio to the CV ratio, where λ is selected by 5-fold cross-validation. We note that when implementing the CV method, for a given training set $\mathcal{T}_{\text{tr}}$, the estimated function value at any $\mathbf{x}$ is set to

$$\widehat{f}(\mathbf{x}) = \max_{\mathbf{x}_i \in \mathcal{T}_{\text{tr}}} \left(\widehat{\theta}_{\lambda,i} + (\mathbf{x} - \mathbf{x}_i)^\top \widehat{\boldsymbol{\xi}}_{\lambda,i} \right), \tag{47}$$

where $\widehat{\theta}_{\lambda,i}$ and $\widehat{\boldsymbol{\xi}}_{\lambda,i}$ are solutions of the penalized multivariate convex regression problem in (31). The constructed $\widehat{f} : \mathbb{R}^d \to \mathbb{R}$ is clearly a (piecewise affine) convex function; see Section 6.5.5 in Boyd and Vandenberghe (2004). We also include the "un-penalized ratio" $L_n(0)/L_n(\lambda^*)$ as a competitor, that is, the ratio between the loss obtained from the un-penalized multivariate convex regression estimator as defined in (30) and the oracle loss.

We present the results in the form of boxplots in Figure 4, obtained from 100 independent replicates of $\mathbf{y}$ (fixing the design variables). We observe that penalized multivariate convex regression, with the regularization parameter tuned by SURE, has better performance. As we had inferred from Figure 3. Figure 4 also shows that the SURE ratios are much smaller than both the CV ratios and un-penalized ratios and their difference is more pronounced as the dimension d increases. Further, the SURE ratio concentrates near one suggesting that SURE is doing a very good job in selecting the tuning parameter.

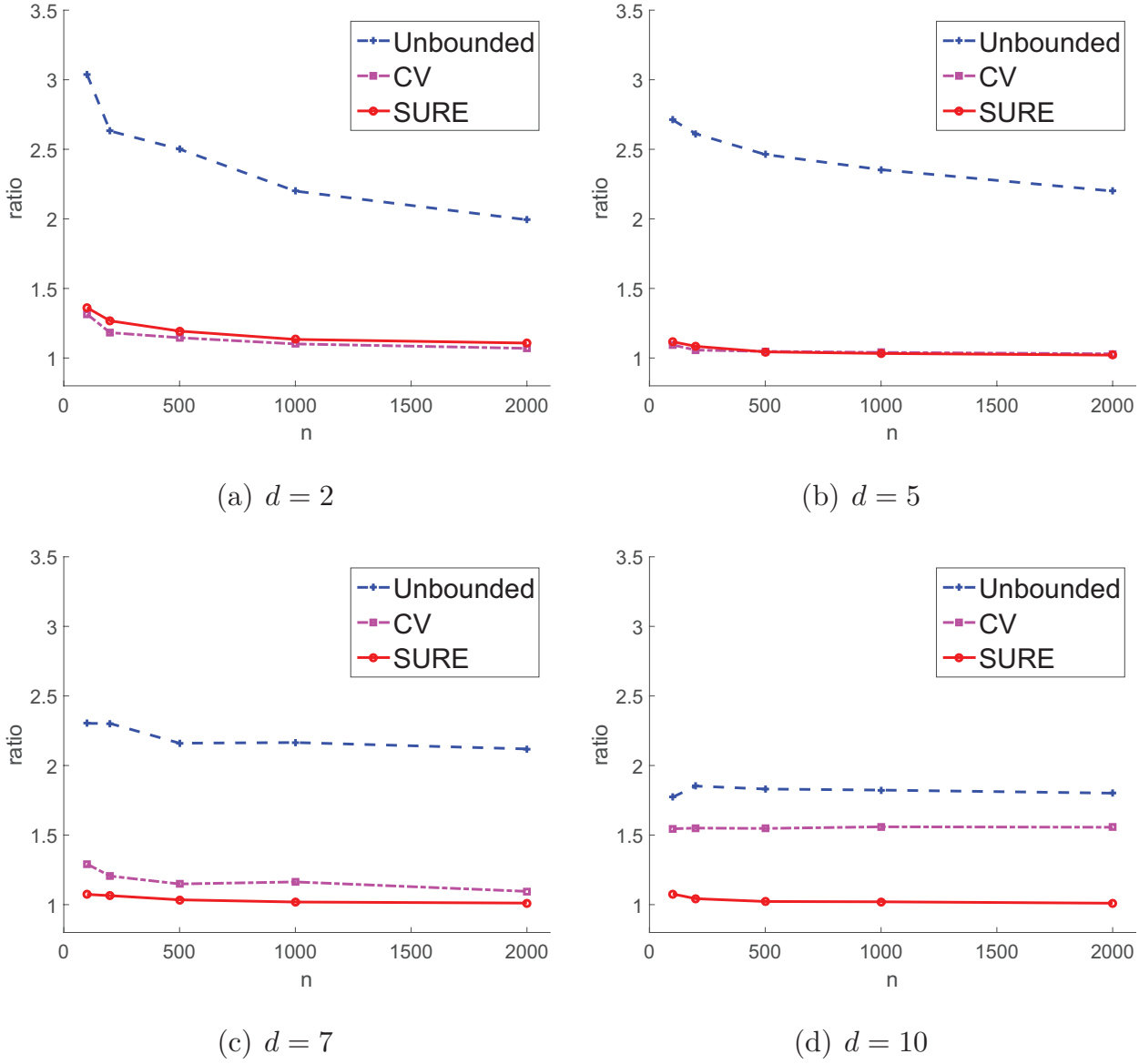


Figure 3. Comparison between the unbounded ratio, the CV ratio and the SURE ratio for isotonic regression.

Table 1. Comparison of the different tuning parameter selection methods for isotonic regression: the unbounded ratio, the CV ratio, the SURE ratio with known σ^2, and the SURE ratio with estimated $\widehat{\sigma^2}$.

n	d	Unbounded	CV	SURE known σ^2	SURE est $\widehat{\sigma^2}$
100	2	3.09 (0.86)	1.28 (0.23)	1.27 (0.22)	1.28 (0.23)
	5	2.66 (0.37)	1.12 (0.11)	1.11 (0.14)	1.47 (0.15)
	10	1.76 (0.25)	1.55 (0.17)	1.09 (0.11)	1.62 (0.17)
1000	2	2.42 (0.50)	1.07 (0.10)	1.10 (0.12)	1.22 (0.15)
	5	2.35 (0.18)	1.04 (0.03)	1.03 (0.05)	1.04 (0.06)
	10	1.80 (0.07)	1.55 (0.05)	1.02 (0.02)	1.48 (0.04)

NOTE: The standard errors are provided in parenthesis.

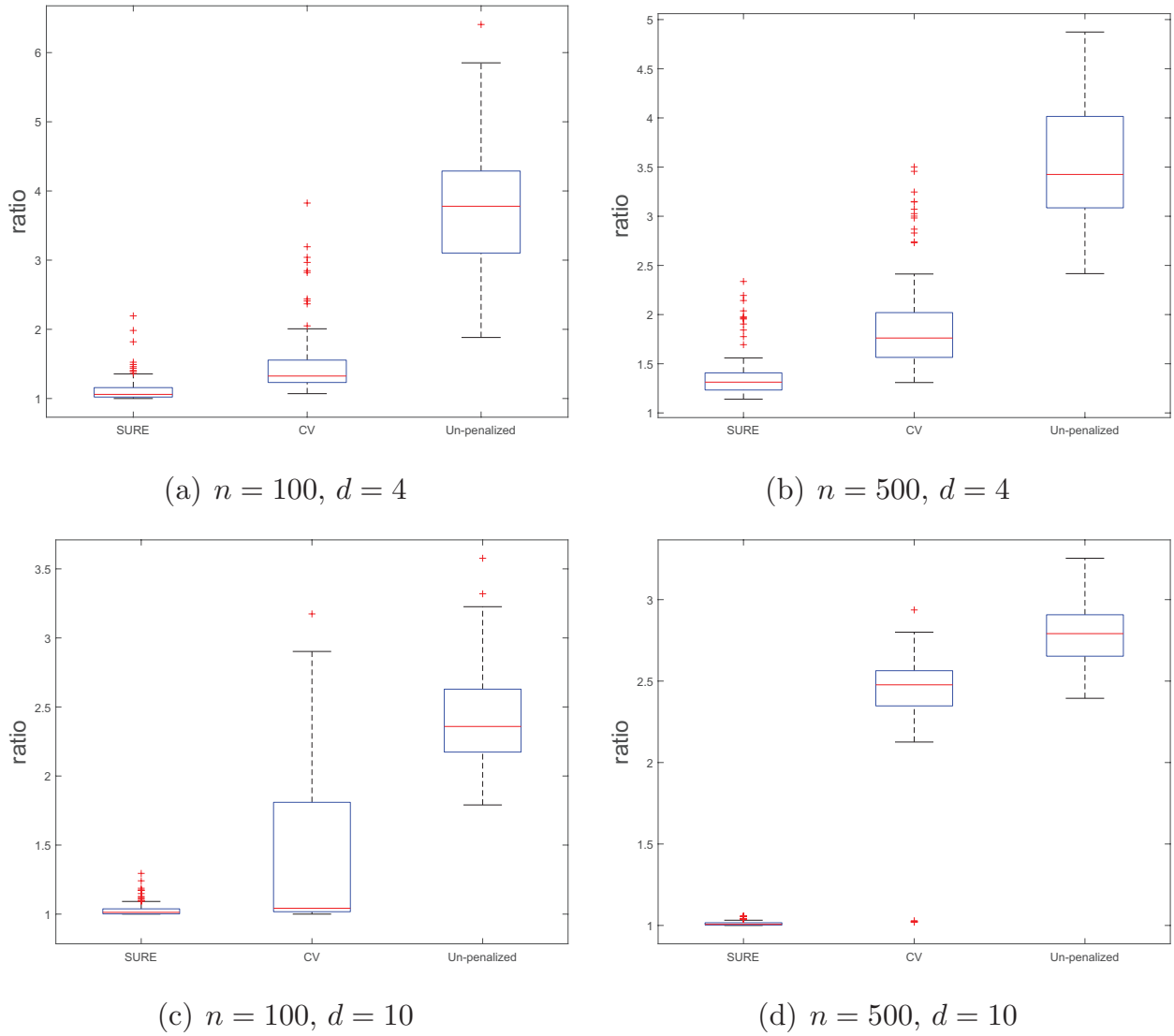
Table 2. Comparison of the different tuning parameter selection methods for convex regression: the un-penalized ratio, the CV ratio, the SURE ratio with known σ^2, and the SURE ratio with estimated $\widehat{\sigma^2}$.

n	d	Un-penalized	CV	SURE known σ^2	SURE est $\widehat{\sigma^2}$
100	2	2.74 (1.12)	1.68 (0.52)	1.35 (0.32)	1.46 (0.39)
	3	3.22 (0.86)	1.42 (0.30)	1.12 (0.22)	1.15 (0.23)
	5	3.62 (0.53)	1.14 (0.25)	1.04 (0.15)	1.30 (0.18)
500	2	2.77 (0.98)	1.20 (0.32)	1.07 (0.11)	1.22 (0.12)
	3	3.47 (0.74)	1.51 (0.29)	1.38 (0.08)	1.49 (0.08)
	5	3.91 (0.50)	1.40 (0.18)	1.05 (0.05)	1.05 (0.06)

NOTE: The standard errors are provided in parenthesis.

7.3. SURE Without the Knowledge of σ^2

In this section, we assume that the noise variance σ^2 in unknown. To estimate σ^2 we adopt a method proposed in Meyer and Woodroofe (2000) and then apply SURE with the estimated σ^2. In particular, we first obtain an initial estimator $\widehat{\boldsymbol{\theta}}$ using unbounded isotonic regression (or un-penalized convex regression) and then estimate σ^2 by $\widehat{\sigma^2} = \frac{\|\widehat{\boldsymbol{\theta}}-\mathbf{y}\|_2^2}{n-2D(\mathbf{y})}$, where $D(\mathbf{y})$ is the divergence of the initial estimator $\widehat{\boldsymbol{\theta}}$. The rationale for this choice comes from Meyer and Woodroofe (2000, Corollary 1) where the authors study (unbiased) estimators for σ^2 in the setup of (8). The averaged ratios $L_n(\widehat{\lambda})/L_n(\lambda^*)$ over 100 independent runs for different tuning parameter selection



(a) $n = 100,\ d = 4$ (b) $n = 500,\ d = 4$

(c) $n = 100,\ d = 10$ (d) $n = 500,\ d = 10$

Figure 4. Boxplots of the SURE ratio, the CV ratio and un-penalized ratio (from left to right) for multivariate convex regression.

methods are provided in Table 1 (for isotonic regression) and Table 2 (for convex regression). For convex regression, the SURE with unknown σ^2 outperforms CV in most cases, whereas for isotonic regression CV performs better in some cases. Moreover, we point out the SURE is computationally more efficient than CV. In particular, 5-fold CV needs to solve five optimization problems for each value of the tuning parameter; thus the SURE method is about five times faster. Moreover, the standard errors of SURE are comparable to those errors of the CV method, and are smaller than the errors for the unbounded and un-penalized cases.

Supplementary Materials

The supplementary material contains all the technical proofs and graphical illustrations.

Acknowledgments

The authors are grateful to the anonymous reviewers and the associate editor for their comments and helpful suggestions.

Funding

Supported by Alibaba innovation research award and Bloomberg data science research grant and Supported by NSF grants DMS-1712822 and AST-1614743.

References

Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W. T., and Silverman, E. (1955), "An Empirical Distribution Function for Sampling With Incomplete Information," *The Annals of Mathematical Statistics*, 26, 641–647. [173]

Boyd, S., and Vandenberghe, L. (2004), *Convex Optimization*, Cambridge: Cambridge University Press. [183]

Brunk, H. D. (1955), "Maximum Likelihood Estimates of Monotone Parameters," *The Annals of Mathematical Statistics*, 26, 607–616. [173]

Candès, E. J., Sing-Long, C. A., and Trzasko, J. D. (2013), "Unbiased Risk Estimates for Singular Value Thresholding and Spectral Estimators," *IEEE Transactions on Signal Processing*, 61, 4643–4657. [174]

Chatterjee, S., Guntuboyina, A., and Sen, B. (2018), "On Matrix Estimation Under Monotonicity Constraints," *Bernoulli*, 24, 1072–1100. [183]

Dantzig, G. B. (1959), *Linear Programming and Extensions*, Princeton, NJ: Princeton University Press. [181]

Donoho, D. L., and Johnstone, I. M. (1995), "Adapting to Unknown Smoothness via Wavelet Shrinkage," *Journal of the American Statistical Association*, 90, 1200–1224. [174]

Efron, B. (2004), "The Estimation of Prediction Error: Covariance Penalties and Cross-Validation," *Journal of the American Statistical Association*, 99, 619–642. [174]

Groeneboom, P., and Jongbloed, G. (2014), *Nonparametric Estimation Under Shape Constraints, Volume 38 of Cambridge Series in Statistical and Probabilistic Mathematics* (Estimators, Algorithms and Asymptotics), New York: Cambridge University Press. [173]

Han, Q., and Wellner, J. A. (2016), "Multivariate Convex Regression: Global Risk Bounds and Adaptation," arxiv preprint arXiv:1601.06844. [174]

Hannah, L. A., and Dunson, D. B. (2011), "Bayesian Nonparametric Multivariate Convex Regression," arXiv preprint arXiv:1109.0322. [178]

Hansen, N. R., and Sokol, A. (2014), "Degrees of Freedom for Nonlinear Least Squares Estimation," arXiv preprint arXiv:1402.2997v3. [176]

Hildreth, C. (1954), "Point Estimates of Ordinates of Concave Functions," *Journal of the American Statistical Association*, 49, 598–619. [174]

Janson, L., Fithian, W., and Hastie, T. J. (2015), "Effective Degrees of Freedom: A Flawed Metaphor," *Biometrika*, 102, 479–485. [176]

Kato, K. (2009), "On the Degrees of Freedom in Shrinkage Estimation," *Journal of Multivariate Analysis*, 100, 1338–1352. [175,176,177]

Kaufman, S., and Rosset, S. (2014), "When Does More Regularization Imply Fewer Degrees of Freedom? Sufficient Conditions and Counterexamples," *Biometrika*, 101, 771–784. [176]

Kuosmanen, T. (2008), "Representation Theorem for Convex Nonparametric Least Squares," *The Econometrics Journal*, 11, 308–325. [174,178]

Li, K. C. (1986), "Asymptotic Optimality of C_L and Generalized Cross-Validation in Ridge Regression With Application to Spline Smoothing," *The Annals of Statistics*, 14, 1101–1112. [175]

Lim, E. (2014), "On Convergence Rates of Convex Regression in Multiple Dimensions," *INFORMS Journal on Computing*, 26, 616–628. [174,179]

Lim, E., and Glynn, P. W. (2012), "Consistency of Multidimensional Convex Regression," *Operations Research*, 60, 196–208. [174,178]

Luss, R., and Rosset, S. (2014), "Generalized Isotonic Regression," *Journal of Computational and Graphical Statistics*, 23(1), 192–210. [173,179]

Luss, R., Rosset, S., and Shahar, M. (2012), "Efficient Regularized Isotonic Regression With Application to Gene-Gene Interaction Search," *Annals of Applied Statistics*, 6, 253–283. [173,179]

Mammen, E., and van de Geer, S. (1997), "Locally Adaptive Regression Splines," *The Annals of Statistics*, 25, 387–413. [174]

Meyer, M., and Woodroofe, M. (2000), "On the Degrees of Freedom in Shape-Restricted Regression," *The Annals of Statistics*, 28, 1083–1104. [174,176,180,183,184]

Mikkelsen, F. R., and Hansen, N. R. (2018), "Degrees of Freedom for Piecewise Lipschitz Estimators," *Annales de l'Institut Henri Poincaré , Probabilités et Statistiques*, *54*, 819–841. [176]

Negahban, S., and Wainwright, M. J. (2011), "Simultaneous Support Recovery in High Dimensions: Benefits and Perils of Block ℓ_1/ℓ_∞-regularization," *IEEE Transactions on Information Theory*, 57, 3841–3863. [182]

Pal, J. K. (2008), "Spiking Problem in Monotone Regression: Penalized Residual Sum of Squares," *Statistics & Probability Letters*, 78, 1548–1556. [173,179]

Petersen, A., Witten, D., and Simon, N. (2016), "Fused Lasso Additive Model," *Journal of Computational and Graphical Statistics*, 25, 1005–1025. [174,181]

Robertson, T., Wright, F. T., and Dykstra, R. L. (1988), *Order Restricted Statistical Inference*. Hoboken, NJ: Wiley. [179]

Rockafellar, R. T. (1970), *Convex Analysis*, Princeton, NJ: Princeton University Press. [177]

Rockafellar, R. T., and Wets, R. J.-B. (2011), *Variational Analysis*, Number 317 in Grundlehren der mathematischen Wissenschaften, Berlin: Springer. [177]

Rudin, L. I., Osher, S., and Fatemi, E. (1992), "Nonlinear Total Variation Based Noise Removal Algorithms," *Physica D: Nonlinear Phenomena*, 60, 259–268. [174]

Rueda, C. (2013), "Degrees of Freedom and Model Selection in Semiparametric Additive Monotone Regression," *Journal of Multivariate Analysis*, 117, 88–99. [176]

Seijo, E., and Sen, B. (2011), "Nonparametric Least Squares Estimation of a Multivariate Convex Regression Function," *The Annals of Statistics*, 39, 1633–1657. [174,178]

Sen, B., and Meyer, M. (2013), "Testing Against a Linear Regression Model Using Ideas From Shape-Restricted Estimation," *Journal of the Royal Statistical Society*, Series B, 79, 423–448. [174,179]

Stein, C. M. (1981), "Estimation of the Mean of a Multivariate Normal Distribution," *The Annals of Statistics*, 9, 1135–1151. [174,183]

Tibshirani, R. J. (2014), "Adaptive Piecewise Polynomial Estimation via Trend Filtering," *The Annals of Statistics*, 42, 285–323. [181]

Tibshirani, R. J., and Taylor, J. (2012), "Degrees of Freedom in Lasso Problems," *The Annals of Statistics*, 40, 1198–1232. [174,175,176,177]

Tibshirani, R. J., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). "Sparsity and Smoothness via the Fused Lasso," *Journal of the Royal Statistical Society*, Series B, 67, 91–108. [174]

Tütüncü, R. H., Toh, K. C., and Todd, M. J. (2003), "Solving Semidefinite-Quadratic-Linear Programs Using SDPT3," *Mathematical Programming*, 95, 189–217. [178,181,183]

Vaiter, S., Deledalle, C.-A., Peyré, G., Fadili, J. M., and Dossal, C. (2014), "The Degrees of Freedom of Partly Smooth Regularizers," arXiv preprint arXiv:1404.5557. [176]

van Eeden, C. (1958), *Testing and Estimating Ordered Parameters of Probability Distributions*. Amsterdam: Mathematical Centre. [173]

Woodroofe, M., and Sun, J. (1993), "A Penalized Maximum Likelihood Estimate of $f(0+)$ When f is Nonincreasing," *Statistica Sinica*, 3, 501–515. [173,179]

Wu, J., Meyer, M. C., and Opsomer, J. D. (2015), "Penalized Isotonic Regression," *Journal of Statistical Planning and Inference*, 161, 12–24. [173,179]

Xie, X. C., Kou, S. C., and Brown, L. D. (2012), "SURE Estimates for a Heteroscedastic Hierarchical Model," *Journal of the American Statistical Association*, 107, 1465–1479. [174]

Xu, M., Chen, M., and Lafferty, J. (2016), "Faithful Variable Screening for High-Dimensional Convex Regression," *The Annals of Statistics*, 44, 2624–2660. [174]

Yi, F., and Zou, H. (2013), "SURE-Tuned Tapering Estimation of Large Covariance Matrices," *Computational Statistics & Data Analysis*, 58, 339–351. [174]

Zhao, P., Rocha, G., and Yu, B. (2009), "Grouped and Hierarchical Model Selection Through Composite Absolute Penalties," *The Annals of Statistics*, 37, 3468–3497. [182]

Zou, H., Hastie, T., and Tibshirani, R. (2007), "On the 'Degrees of Freedom' of the Lasso," *The Annals of Statistics*, 35, 2173–2192. [174,175]