http://pubsonline.informs.org/journal/ijoc/

INFORMS JOURNAL ON COMPUTING
Articles in Advance, pp. 1–14
ISSN 1091-9856 (print), ISSN 1526-5528 (online)

Robust Maximum Likelihood Estimation

Dimitris Bertsimas,[a] Omid Nohadani[b]

[a] Operations Research Center and Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139;
[b] Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, Illinois 60208

Contact: dbertsim@mit.edu, http://orcid.org/0000-0002-1985-1003 (DB); nohadani@northwestern.edu, http://orcid.org/0000-0001-6332-3403 (ON)

Received: January 10, 2016
Revised: February 12, 2017; December 19, 2017
Accepted: April 30, 2018
Published Online in Articles in Advance: April 26, 2019

https://doi.org/10.1287/ijoc.2018.0834

Copyright: © 2019 INFORMS

Abstract. In many applications, statistical estimators serve to derive conclusions from data, for example, in finance, medical decision making, and clinical trials. However, the conclusions are typically dependent on uncertainties in the data. We use robust optimization principles to provide robust maximum likelihood estimators that are protected against data errors. Both types of input data errors are considered: (a) the adversarial type, modeled using the notion of uncertainty sets, and (b) the probabilistic type, modeled by distributions. We provide efficient local and global search algorithms to compute the robust estimators and discuss them in detail for the case of multivariate normally distributed data. The estimator performance is demonstrated on two applications. First, using computer simulations, we demonstrate that the proposed estimators are robust against both types of data uncertainty and provide more accurate estimates compared with classical estimators, which degrade significantly, when errors are encountered. We establish a range of uncertainty sizes for which robust estimators are superior. Second, we analyze deviations in cancer radiation therapy planning. Uncertainties among plans are caused by patients' individual anatomies and the trial-and-error nature of the process. When analyzing a large set of past clinical treatment data, robust estimators lead to more reliable decisions when applied to a large set of past treatment plans.

History: Accepted by Karen Aardal, Area Editor for Design and Analysis of Algorithms.
Funding: O. Nohadani received support from the National Science Foundation [Grant CMMI-1463489].
Supplemental Material: The online supplement is available at https://doi.org/10.1287/ijoc.2018.0834.

Keywords: optimization • robust optimization • robust statistics • maximum likelihood estimator • radiation therapy

1. Introduction

Maximum likelihood is widely used to successfully construct statistical estimators for parameters of a probability distribution. This method is prevalent in many applications, ranging from econometrics and machine learning to many areas of science and engineering, where the nature of the data motivates a functional form of the underlying distribution. The parameters of this distribution remain to be estimated (Pfanzagl 1994). For uncertain distributions, the estimators can be located in a family of them by leveraging the minimax asymptotic variance (Huber et al. 1964, Huber 1996). This is also possible in the common case of symmetric contamination (Jaeckel 1971). Maximum likelihood estimator (MLE) methods typically assume data to be complete, precise, and free of errors. In reality, however, data are often insufficient. Moreover, any input data can be subject to errors and perturbations. These can stem from (a) measurement errors, (b) input errors, (c) implementation errors, (d) numerical errors, or (e) model errors. These sources of uncertainty affect the quality of the estimators and can degrade outcomes significantly, so much so that we might lose the advantages of maximum likelihood estimators completely. Therefore, it is instrumental to the success of the application to construct estimators that are intrinsically robust against possible sources of uncertainty.

In this work, we seek to estimate the parameter θ for the probability density function $f(\theta, \mathbf{x})$ of an ensemble of n data points $\mathbf{x}_i \in \mathbb{R}^m$, which can be contaminated by errors $\Delta\mathbf{x}_i \in \mathbb{R}^m$ in some uncertainty set $\mathcal{U}$. The robust MLE maximizes the worst-case likelihood via

$$\max_{\theta} \min_{\Delta\mathbf{X} \in \mathcal{U}} \prod_{i=1}^{n} f(\theta; \mathbf{x}_i - \Delta\mathbf{x}_i), \tag{1}$$

following the robust optimization (RO) paradigm.

Robust optimization has increasingly been used as an effective way to immunize solutions against data uncertainty. In principle, if errors are not taken into account, an otherwise optimal solution may turn out to be suboptimal, or even in some cases infeasible. RO, however, considers errors to reside within an uncertainty set and aims to calculate solutions that are robust to such uncertainty. There is a sizable body of literature on various aspects of RO, and we refer to Ben-Tal et al. (2009) and Bertsimas et al. (2011). In the context of

simulation-based problems, that is, problems not given by a closed form solution, a local search algorithm was proposed that provides robust solutions to unconstrained (Bertsimas et al. 2010b) and constrained (Bertsimas et al. 2010a) problems without exploiting the structure of the problem or the uncertainty set.

The effects of measurement errors on statistical estimators have been addressed extensively in the literature, for example, by Buonaccorsi (2010) and the references within. Measurement error models assume a distribution of the observed values given the true values of a certain quantity (Fuller 2009). This is the reverse of the Berkson error model, which assumes a distribution on the true values given the observed values (Berkson 1950). The rich literature for error correction provides a plethora of techniques for correcting additive errors in linear regression (Cheng et al. 1999).

El Ghaoui and Lebret (1997) showed that robust least squares problems for erroneous but bounded data can be formulated as second-order cone or semidefinite optimization problems, and thus become efficiently solvable. In the context of maximum likelihood, Calafiore and El Ghaoui (2001) elaborated on estimators in linear models in the presence of Gaussian noise whose parameters are uncertain. The proposed estimators maximize a lower bound on the worst-case likelihood using semidefinite optimization.

In the context of robust statistics, Huber (1980) introduced estimators that are insensitive to perturbations. The robustness of estimators is measured in different ways: For instance, the breakdown point is defined as the minimum amount of contamination that causes the estimator to become unreliable. Another measure is the influence curve that describes the impact of outliers to an estimator (Hampel 1974).

In this paper, we introduce a robust MLE method to produce estimators that are robust to data uncertainty. Our approach differs from Huber's (1980) in multiple facets, as summarized in Table 1. Because the likelihood is not proportional to the error in the estimation, our proposed method considers the worst case directly in the likelihood. Correspondingly, we believe that our proposed approach is directly relevant to real-world data, where errors are observed on the data and a priori information on distributions is not available.

Table 1. Comparison of Huber's (1980) Approach to the Proposed RO Approach

Comparison	Huber's approach	Our approach
Estimators	Functionals on the space of distributions	Functions on the observed data
Worst case	In the value of the estimator	In the value of the likelihood
Observables	Observed distributions reside in a neighborhood around the true distribution	True data reside in a neighborhood around the observed data

In principle, errors may be of an adversarial nature, where no probabilistic information about their source is known, or of a distributional nature, where the source is known to be probabilistic. Correspondingly, we discuss two kinds of robust maximum likelihood estimators:

1. *Adversarially robust*: The worst-case scenario is calculated among possible errors that reside in some uncertainty set. To compute these estimators, we propose two methods: a first-order gradient descent algorithm, which is highly efficient and warrants local optimal robust estimators, and a global search method based on robust simulated annealing, which provides global optimal robust estimators at higher computationally expense.

2. *Distributionally robust*: The worst-case scenario is evaluated among errors that are independent and follow some distribution residing in some set of distributions. Such errors resemble persistent errors. Using distributional robust optimization techniques, we show that their estimators are a particular case of adversarially robust estimators.

To demonstrate the performance of the methods, we apply our methods to two types of data sets. First, we conduct numerical experiments on simulated data to ensure a controlled setting and to be able to determine the deviation from true data. We show that for small-sized errors, both the local and the global RO methods yield comparable estimates. For larger errors, however, we observe that the robust simulated annealing method outperforms the local search method. Moreover, we show that the proposed estimators are also immune against the source of uncertainty; that is, even if the errors follow a different distribution than anticipated, the estimators remain robustly optimal. Furthermore, the proposed robust estimators turn out to be significantly more accurate compared with classical maximum likelihood estimators, which degrade sizably, when errors are encountered. Finally, we establish the range within which the robust estimators are most effective. This range can inform practitioners about the appropriate size of the uncertainty set. The error size–dependent observation can be generalized to a broader range of RO approaches.

In the second application, past patient data for cancer radiation therapy serve to probe the method. In clinical practice, the quality of treatment plans is typically examined based on the spatial dose distribution, summarized in five specific observable dose points and compared with internally recommended criteria. However, the recording and evaluation of these criteria are subject to human uncertainty. To evaluate the overall performance of an individual clinician, a team, or an entire institution, statistical estimators are employed. The resulting decisions remain highly sensitive to the

uncertainty in data. We analyze 491 treatment plans for various tumor sites that have already undergone radiation therapy and compute robust estimators to support physicians in arriving at more dependable conclusions. We show that robust estimators have a significantly reduced and stable spread over different samples, when compared with nominal estimators, offering more reliable and sample-independent decision making.

The structure of this paper is as follows: In Section 2, we define the robust estimators and introduce the RO problem for robust estimators. In Section 2.1, we discuss the robust maximum likelihood estimators, along with a corresponding local as well as a global search algorithm. Section 2.1 also details the method to compute the corresponding robust estimates when the data follow a specific distribution, namely, the multivariate normal distribution. In Section 3, we report on the performance of the proposed robust estimators on simulated data. In Section 4, we compute robust estimators for a large set of clinical cancer radiation therapy data. In Section 5, we conclude our findings.

2. Robust Maximum Likelihood Estimators

To introduce the robust estimators, we first differentiate between observed and true data. Consider the following setting: we can only observe samples $\mathbf{x}_i^{\text{obs}}$, $i = 1, 2, \ldots, n$, which may include errors. This is expressed via

$$\mathbf{x}_i^{\text{obs}} = \mathbf{x}_i^{\text{true}} + \Delta\mathbf{x}_i, \; i = 1, 2, \ldots, n,$$

where $\mathbf{x}_i^{\text{true}}$ is the error-free (but not observable) data, and $\Delta\mathbf{x}_i$ is the error in the ith sample. The error-free data $\mathbf{x}_i^{\text{true}}$ are assumed to be distributed according to a distribution $\mathcal{W}(\theta)$, with probability density function $f(\theta; \mathbf{x})$, where θ is the parameter we wish to estimate.

Maximum likelihood estimator seeks to find θ that maximizes the probability density function

$$\prod_{i=1}^{n} f(\theta; \mathbf{x}_i^{\text{true}}) \equiv \prod_{i=1}^{n} f(\theta; \mathbf{x}_i^{\text{obs}} - \Delta\mathbf{x}_i), \tag{2}$$

or equivalently maximizes the log-likelihood density function

$$\psi(\theta; \mathbf{X}^{\text{obs}} - \Delta\mathbf{X}) \equiv \log\left(\prod_{i=1}^{n} f(\theta; \mathbf{x}_i^{\text{obs}} - \Delta\mathbf{X}_i)\right), \tag{3}$$

where $\mathbf{X}^{\text{obs}}$ denotes the ensemble of the observed data $\mathbf{x}_i^{\text{obs}}$, and $\Delta\mathbf{X}$ the ensemble of $\Delta\mathbf{x}_i$ as

$$\mathbf{X}^{\text{obs}} = [\mathbf{x}_1^{\text{obs}}, \mathbf{x}_2^{\text{obs}}, \ldots, \mathbf{x}_n^{\text{obs}}]^\top, \text{ and}$$
$$\Delta\mathbf{X} = [\Delta\mathbf{x}_1, \Delta\mathbf{x}_2, \ldots, \Delta\mathbf{x}_n]^\top.$$

In what will follow, the errors $\Delta\mathbf{x}_i$, $i = 1, 2, \ldots, n$, are modeled in two different ways:

(a) no further knowledge about the nature of errors is available, and we consider them to reside within an uncertainty set, which leads to adversarially robust estimators (AREs), introduced in Section 2.1;

(b) errors can be considered as random variables with known support, which leads to designing distributionally robust estimators (DREs), introduced in Section 2.3.

In both cases, we assume the magnitude of $\Delta\mathbf{X}$ to be sizably smaller than that of $\mathbf{X}^{\text{true}}$, making the estimators identifiable. When they are of comparable sizes, the distinction of their parameters fades. However, a discussion on identifiability for general cases is beyond the scope of this work. With respect to likelihood methods, our approach can be regarded as *semiparametric* because we specify only the distribution of $\mathbf{X}^{\text{true}}$ parametrically, and not that of $\Delta\mathbf{X}$.

2.1. Adversarial Robust Maximum Likelihood Estimators

In most real-world applications, the knowledge about the error distribution is not accessible and at times not even existent. For example, medical records are often manually transferred from a diagnostic device onto paper and later into an electronic form, and during each step, copying errors and uncertainties from unit conversion may occur. The nature of these errors cannot be associated to a known distribution. Therefore, following the RO paradigm, we model such errors $\Delta\mathbf{X}$ as belonging to an uncertainty set, which is assumed to be a convex set and denoted by $\mathcal{U}$. The set $\mathcal{U}$ is typically determined by the underlying application; for example, the accuracy of a measurement device determines the size of measurement errors in data. When the Euclidean norm of the errors is bounded by a parameter $\Gamma > 0$, the corresponding uncertainty set can be expressed as

$$\mathcal{U} = \left\{ \Delta\mathbf{X} = [\Delta\mathbf{x}_1, \Delta\mathbf{x}_2, \ldots, \Delta\mathbf{x}_n]^\top \,\middle|\, \|\Delta\mathbf{x}_i\|_2 \le \Gamma, \; i = 1, 2, \ldots, n \right\}. \tag{4}$$

Although this set serves to clarify the exposition, our gradient descent approach does not leverage this specific structure, as will be discussed. We seek to find θ that maximizes the log-likelihood in Equation (3) against all errors (in particular the worst-case one) in $\mathcal{U}$. Therefore, the ARE is the solution to

$$\max_{\theta} \min_{\Delta\mathbf{X} \in \mathcal{U}} \sum_{i=1}^{n} \log\left(f(\theta; \mathbf{x}_i^{\text{obs}} - \Delta\mathbf{x}_i)\right). \tag{5}$$

Note that if there are no errors, $\mathcal{U} = \{0\}$, and problem (5) corresponds to classical maximum likelihood estimation. As the size of $\mathcal{U}$ increases, the ARE may become more conservative, that is, we attempt to immunize against larger errors, potentially at the expense of lower likelihood.

2.2. Computation of AREs

To compute adversarial maximum likelihood estimates, we first discuss the gradient-based robust optimization method for nonconvex cost functions before extending it to compute local optimal estimators. We then introduce a global search method based on robust simulated annealing that warrants global robust optimality.

2.2.1. Robust Nonconvex Optimization. In general, for a continuously differentiable and possibly nonconvex cost function $f(\mathbf{x})$, where $\mathbf{x} \in \mathbb{R}^n$ is the decision or data vector, the *robust optimization problem* is given through

$$\min_{\mathbf{x}} g(\mathbf{x}) \equiv \min_{\mathbf{x}} \max_{\Delta\mathbf{x} \in \mathcal{U}} f(\mathbf{x} + \Delta\mathbf{x}). \tag{6}$$

Here, the errors $\Delta\mathbf{x}$ directly affect the decision variables and are bound in an uncertainty set $\mathcal{U}$. The robust problem can be solved by updating $\mathbf{x}$ along *descent directions* that reduce g by excluding worst errors. In other words, $\mathbf{d}$ is a descent direction for the robust optimization problem (6) at $\mathbf{x}$, if the directional derivative in direction $\mathbf{d}$ satisfies the following condition:

$$g'(\mathbf{x}) < 0. \tag{7}$$

Note that for problem (6), it may not be possible to find $\Delta\mathbf{x}^* = \arg\max_{\Delta\mathbf{x} \in \mathcal{U}} f(\mathbf{x} + \Delta\mathbf{x})$ or the solution may not be unique. However, it has been shown that when it is possible to provide a collection of $\mathcal{M} = \{\Delta\mathbf{x}_1, \ldots, \Delta\mathbf{x}_m\}$ with $\Delta\mathbf{x}_i \in \mathcal{U}$ and $\Delta\mathbf{x}^* = \sum_{i \mid \Delta\mathbf{x}_i \in \mathcal{M}} \lambda_i \Delta\mathbf{x}_i$ for some $\lambda_i \geq 0$, then $\mathbf{d}$ is a descent direction for $g(\mathbf{x}; \mathbf{d})$, if $\mathbf{d}^\top \Delta\mathbf{x}_i < 0$ $\forall \Delta\mathbf{x}_i \in \mathcal{M}$ (Bertsimas et al. 2010b). Furthermore, such a descent direction points away from all the worst implementation errors in $\mathcal{U}$, as shown by the following theorem:

Theorem 1. *Suppose that $f(\mathbf{x})$ is continuously differentiable, the uncertainty set $\mathcal{U}$ is defined as in (4), and $\mathcal{U}^*(\mathbf{x}) := \left\{\Delta\mathbf{x}^* \mid \Delta\mathbf{x}^* \in \arg\max_{\Delta\mathbf{x} \in \mathcal{U}} f(\mathbf{x} + \Delta\mathbf{x})\right\}$. Then, $\mathbf{d} \in \mathbb{R}^n$ is a descent direction for the worst-case cost function $g(\mathbf{x})$ at $\mathbf{x} = \hat{\mathbf{x}}$ if and only if for all $\Delta\mathbf{x}^* \in \mathcal{U}^*(\hat{\mathbf{x}})$,*

$$\mathbf{d}^\top \Delta\mathbf{x}^* < 0,$$
$$\nabla_{\mathbf{x}} f(\mathbf{x} = \hat{\mathbf{x}} + \Delta\mathbf{x}^*) \neq \mathbf{0}.$$

Note that all descent directions $\mathbf{d}$ reside in the strict interior of a cone, which is normal to the cone spanned by all the vectors $\Delta\mathbf{x}^* \in \mathcal{U}^*(\hat{\mathbf{x}})$. Consequently, the worst-case cost at $\hat{\mathbf{x}}$ can be strictly decreased, if a sufficiently small step is taken along any directions within this cone, leading to solutions that are more robust. All worst solutions, $\hat{\mathbf{x}} + \Delta\mathbf{x}^*$, would also lie outside the neighborhood of the updated solution. Therefore, $\mathbf{x}^*$ can be considered a robust local minimum, if there exists no descent direction for the robust problem at $\mathbf{x} = \mathbf{x}^*$. The proof of Theorem 1 as well as empirical evidence of the robust optimization method's behavior is discussed in Bertsimas et al. (2010b).

In summary, if we can compute the directional derivative of the inner function of a robust optimization problem, the above method can efficiently provide robust solutions. We now extend this approach to compute robust estimators.

2.2.2. Local Optimal Estimator. Let $\phi(\theta)$ be the solution to the inner minimization problem (5) as

$$\phi(\theta) \equiv \min_{\Delta\mathbf{X} \in \mathcal{U}} \psi(\theta; \mathbf{X}^{\text{obs}} - \Delta\mathbf{X}) \tag{8}$$

with

$$\Delta\mathbf{X}^*(\theta) \equiv \arg\min_{\Delta\mathbf{X} \in \mathcal{U}} \psi(\theta; \mathbf{X}^{\text{obs}} - \Delta\mathbf{X}). \tag{9}$$

By applying Danskin's (1966) theorem, we have

$$\nabla_\theta \phi(\theta)\Big|_{\theta=\theta_0} = \nabla_\theta \psi(\theta; \mathbf{X}^{\text{obs}} - \Delta\mathbf{X}^*(\theta_0))\Big|_{\theta=\theta_0}. \tag{10}$$

Note that we do not need to calculate the gradient at $(\mathbf{X}^{\text{obs}} - \Delta\mathbf{X}^*(\theta))$. Given the ability to calculate Equation (10), we can construct a gradient descent algorithm with diminishing step size, which has been shown to converge to a local minimum (Bertsimas et al. 2010b).

2.2.3. The Case of Multivariate Normal Distribution. To demonstrate the performance of this approach, some specifications on the underlying distribution of the data are necessary. We use the example of the multivariate normal distribution of the observed data. In particular, the framework for computing AREs is employed to estimate the mean $\boldsymbol{\mu} \in \mathbb{R}^m$ and the covariance matrix $\boldsymbol{\Sigma} \in S_m^+$, where S_m^+ is the set of $m \times m$ symmetric and positive semidefinite matrices. The probability density function for some observed data $\mathbf{x}^{\text{obs}}$ is

$$f(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{x}^{\text{obs}}) = \frac{1}{(2\pi)^{m/2} |\boldsymbol{\Sigma}|^{1/2}} \cdot \exp\left(-\frac{1}{2}(\mathbf{x}^{\text{obs}} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x}^{\text{obs}} - \boldsymbol{\mu})\right). \tag{11}$$

Using the uncertainty set defined in (4), problem (8) for the estimators $\theta = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ becomes

$$\phi(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{X}^{\text{obs}}) = \min_{\Delta\mathbf{X} \in \mathcal{U}} \psi(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{X}^{\text{obs}} - \Delta\mathbf{X}) = \tag{12}$$
$$\min_{\|\Delta\mathbf{x}_i\|_2 \leq \Gamma} -\frac{nm}{2}\log(2\pi) - \frac{n}{2}\log|\boldsymbol{\Sigma}| + \sum_{i=1}^{n} -\frac{1}{2}(\mathbf{x}_i^{\text{obs}} - \Delta\mathbf{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i^{\text{obs}} - \Delta\mathbf{x}_i - \boldsymbol{\mu}).$$

Note, the objective function and constraints are separable in $\Delta\mathbf{x}_i$. Thus, it suffices to solve

$$\min_{\|\Delta\mathbf{x}_i\|_2 \leq \Gamma} -\frac{1}{2}(\mathbf{x}_i^{\text{obs}} - \Delta\mathbf{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i^{\text{obs}} - \Delta\mathbf{x}_i - \boldsymbol{\mu}) \tag{13}$$

for each $i = 1, 2, \ldots, n$. The objective function of problem (13) can be written as

$$-\frac{1}{2}(\mathbf{x}_i^{\text{obs}} - \Delta\mathbf{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i^{\text{obs}} - \Delta\mathbf{x}_i - \boldsymbol{\mu}) =$$
$$-\frac{1}{2}\Delta\mathbf{x}_i^\top \boldsymbol{\Sigma}^{-1}\Delta\mathbf{x}_i + [\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i^{\text{obs}} - \boldsymbol{\mu})]^\top \Delta\mathbf{x}_i$$
$$-\frac{1}{2}(\mathbf{x}_i^{\text{obs}} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i^{\text{obs}} - \boldsymbol{\mu}).$$

Problem (13) is a trust region problem that has been solved in the literature, as by Boyd and Vandenberghe (2004) and Rendl and Wolkowicz (1997). In another work, Ye (1992) demonstrated that for such problems, a hybrid algorithm that combines Newton's method and a binary search can solve the problem in $O(\log(\log(1/\epsilon)))$ iterations, with error ϵ. In the online supplement, Section 1, we briefly describe the steps for computing this inner problem. Overall, the algorithm to compute the robust and normal distributed estimators, as defined in problem (5), can be summarized as follows:

1. Initialize with some estimators $\theta = [\boldsymbol{\mu}, \boldsymbol{\Sigma}]^\top$.
2. Solve problem (13) to obtain $\Delta\mathbf{x}_i^*(\theta)$ for each $i = 1, 2, \ldots, n$.
3. Use the worse-case errors $\Delta\mathbf{X}^*(\theta) = [\Delta\mathbf{x}_1^*(\theta), \Delta\mathbf{x}_2^*(\theta), \ldots, \Delta\mathbf{x}_n^*(\theta)]^\top$ to calculate $\phi(\theta) = \psi(\theta; \mathbf{X}^{\text{obs}} - \Delta\mathbf{X}^*(\theta))$ and Equation (10) to compute its derivative $\nabla\phi$.
4. Construct a Q such that $Q \cdot \theta = 0$.
5. Compute $\widehat{\nabla\phi}$ as the projection of $\nabla\phi$ onto the subspace $Q \cdot \theta = 0$ using the kernel of Q.
6. Update θ using the descent direction given by $\widehat{\nabla\phi}$ (preserves $\boldsymbol{\Sigma} \in S^+$).
7. Stop when the norm of the derivative is smaller than some tolerance parameter ϵ; otherwise, iterate back to Step 2.

Following the result of Theorem 1, this algorithm provides the local robust maximum likelihood estimators. Furthermore, the convergence of the algorithm is linear, because it is a first-order method using gradient descent. Second-order methods are computationally inefficient, because the inner derivative $\nabla_\theta(\Delta\mathbf{x}_i^*(\theta))$ complicates the calculation of the second derivate of ϕ. We now discuss an alternative method to obtain the global optimal estimators in the presence of errors in the input data.

2.2.4. Global Optimal Estimator. The robust normal distribution estimators can also be calculated using a global search method. The global search method is based on the robust simulated annealing algorithm introduced by Bertsimas and Nohadani (2010). To follow the original robust simulated annealing methods, we recast the maximization problem $\max \phi(\theta; \mathbf{X}^{\text{obs}})$ as a minimization problem, $\min g$, with $g \equiv -\phi$. Starting from the nominal optimum, this iterative algorithm lowers the worst-case performance g successively. At each step, g is computed for the corresponding estimates within the uncertainty set, and the inverse temperature is determined. The Boltzmann weight assigned to the current estimates are then compared for a trial estimate. If the trial estimates lead to a lower g, they will become the estimates of the next step. Otherwise, they will most likely be rejected and new trial estimates will be generated and compared. We refer interested readers to Bertsimas and Nohadani (2010) for further details. For completeness, however, we summarize the steps of the algorithm along with the notation in Section 2 of the online supplement.

2.3. Distributional Robust Maximum Likelihood Estimators

In some cases, the errors $\Delta\mathbf{x}_i$ are independent of the samples and among each other, and they follow the same distribution P. For example, when there is a fixed error caused by misalignments of the equipment, we can consider them as persistent errors that lend themselves to a distributional description. For this, let $f_P(\Delta\mathbf{x})$ be the probability density function for P. Then, $\mathbf{x}_i^{\text{obs}}$ follows a distribution, whose density is the convolution

$$\int f(\theta; \mathbf{x}_i^{\text{obs}} - \Delta\mathbf{x}) f_P(\Delta\mathbf{x}) d\Delta\mathbf{x} = \int f(\theta; \mathbf{x}_i^{\text{obs}} - \Delta\mathbf{x}) dP(\Delta\mathbf{x})$$
$$= E_{\Delta\mathbf{x}\sim P}[f(\theta; \mathbf{x}_i^{\text{obs}} - \Delta\mathbf{x})]. \tag{14}$$

Following the paradigm of distributional robust optimization (Delage and Ye 2010), we consider all possible distributions to have a bounded support $\mathcal{S}$ and the DRE to be the solution to

$$\max_\theta \min_{P:\, supp(P)=\mathcal{S}} \sum_{i=1}^n \log E_{\Delta\mathbf{x}_i\sim P}[f(\theta; \mathbf{x}_i^{\text{obs}} - \Delta\mathbf{x}_i)]. \tag{15}$$

Because all distributions share $\mathcal{S}$, problem (15) can be reformulated as

$$\max_\theta \min_{\Delta\mathbf{x}\in\mathcal{S}} \sum_{i=1}^n \log f(\theta; \mathbf{x}_i - \Delta\mathbf{x}), \tag{16}$$

which has the same structure as problem (5), and thus can be solved in a similar fashion as that for the ARE, using a gradient descent algorithm.

2.3.1. The Case of Multivariate Normal Distribution. Analogous to AREs, we demonstrate the performance of DREs by the specific assumption of a multivariate

normal distribution. The inner minimization of problem (16) is

$$\begin{aligned}
&\min_{\|\Delta\mathbf{x}\|_2\le\Gamma} \psi(\boldsymbol{\mu},\boldsymbol{\Sigma};\mathbf{X}^{\text{obs}}-\mathbf{1}_n\Delta\mathbf{x}^\top)\\
&= \min_{\|\Delta\mathbf{x}\|_2\le\Gamma} -\frac{nm}{2}\log(2\pi)-\frac{n}{2}\log|\boldsymbol{\Sigma}|\\
&\quad+\sum_{i=1}^{n}-\frac{1}{2}(\mathbf{x}_i^{\text{obs}}-\Delta\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i^{\text{obs}}-\Delta\mathbf{x}-\boldsymbol{\mu})\\
&= -\frac{nm}{2}\log(2\pi)-\frac{n}{2}\log|\boldsymbol{\Sigma}|\\
&\quad+\min_{\|\Delta\mathbf{x}\|_2\le\Gamma}\left(\Delta\mathbf{x}^\top\left(-\frac{n}{2}\boldsymbol{\Sigma}^{-1}\right)\Delta\mathbf{x}+\left[\boldsymbol{\Sigma}^{-1}\left(\sum_{i=1}^{n}\mathbf{x}_i^{\text{obs}}-n\boldsymbol{\mu}\right)\right]^\top\Delta\mathbf{x}\right)\\
&\quad+\sum_{i=1}^{n}-\frac{1}{2}(\mathbf{x}_i^{\text{obs}}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i^{\text{obs}}-\boldsymbol{\mu}),
\end{aligned}\tag{17}$$

where $\mathbf{1}_n$ is a size n vector of ones. Note that problem (17) is a trust region problem (online supplement, Section 1.2). In the maximization of problem (16), the descent direction is projected into the space of positive semidefinite matrices. Furthermore, problem (17) is a special case of problem (12) in that all $\Delta\mathbf{x}_i$ are restricted to be the same. This implies that DREs are less conservative than AREs.

2.4. Discussion

When using data for statistical inference, overfitting has been a central challenge. A variety of machine learning algorithms have been developed to overcome these issues, particularly via regularization techniques. The paradigm of robust optimization provides a unifying justification for the success of many of these methods. The robustness of a result can be associated with problem attributes such as consistency and sparsity (see Sra et al. 2012 and references within). In this vein, our presented robust maximum likelihood estimators are also *robust* against overfitting, as the inner minimization problem inherently reduces the complexity and improves the predicative power of the estimator. Both of the following numerical experiments support this observation, as the resulting estimators are broadly insensitive to error size and sample size.

3. Computational Results with Simulation Data

To evaluate the robust estimators, we conduct experiments using computer-generated random data. The purpose of using simulated data is that we have accurate information about the true data that can serve as a reference, allowing us to directly measure the quality of robust estimators on observed data. We generate samples following a multivariate normal distribution. As our estimators are designed to deal with errors in the samples, we generate errors following both a normal and a uniform distribution, and use them to contaminate our samples.

More specifically, the experiments are conducted in the following fashion. A number of $n = 400$ samples in $\mathbb{R}^4$ are generated randomly, following the multivariate normal distribution with some random mean and some random covariance matrix. Let $\mathbf{X}^{\text{true}}$ be the 400×4 matrix containing the samples $\mathbf{x}_i^{\text{true}}$, $i = 1, 2, \ldots, 400$, in its rows. The vectors $\mathbf{x}_i^{\text{true}}$ are the true samples and are not affected by errors.

Furthermore, we generate errors on the samples in the following way: $\Delta\mathbf{X}_k$, $k = 1, 2, \ldots, 40$, is a 400×4 matrix containing errors corresponding to the samples in the 400×4 matrix $\mathbf{X}^{\text{true}}$. The errors in $\Delta\mathbf{X}_k$ follow a normal distribution with mean $\mathbf{0}$ and covariance matrix $\mathbf{I}_4$, where $\mathbf{0}$ is the zero vector in $\mathbb{R}^4$, and $\mathbf{I}_4$ is the 4×4 identity matrix. In the experiments, we will use the parameter ρ to scale the magnitude of contamination. Correspondingly, we also employ ρ to tune the uncertainty set size. In this context, Γ is used for a constant set size, as will be discussed in Section 3.2.

For the uncertainty set that contains the simulated errors, we evaluate the performance of the estimators using the worst-case and average values of the probability density, as well as their distance from the value of the nominal estimator on the true data. Initially, we use the normally distributed errors. Later, we compare the results with the case of uniformly distributed errors. In each case, we consider both AREs and DREs.

The experimental section is organized as follows. In Section 3.1, we evaluate the estimators based on the worst-case and average values of the probability density. In Section 3.2, we evaluate the estimators based on their distance from the nominal estimator on the true data. In Section 3.3, we compare the robust estimators to the nominal one, using the local and the global search methods for the ARE and DRE cases. In Section 3.4, we discuss the effects of different distributions for the errors, in particular, when we have uniformly distributed errors.

3.1. Worst-Case and Average Probability Density

To probe the efficiency of the proposed robust estimators, we will check the worst-case and average values of the probability density, as we add properly scaled errors from the set of errors $\Delta\mathbf{X}_k$ to the true values of the data.

In particular, we calculate the AREs for the true data $\mathbf{X}^{\text{true}}$, for varying sizes of the assumed uncertainty set, as defined in (4), between $\rho = 0$ and 3 with a step size of 0.1. We denote them by $\hat{\boldsymbol{\mu}}_{\text{rob.a.}}(\mathbf{X}^{\text{true}},\rho)$ and $\hat{\boldsymbol{\Sigma}}_{\text{rob.a.}}(\mathbf{X}^{\text{true}},\rho)$. Moreover, we calculate the DREs in the same cases and denote them by $\hat{\boldsymbol{\mu}}_{\text{rob.d.}}(\mathbf{X}^{\text{true}},\rho)$ and $\hat{\boldsymbol{\Sigma}}_{\text{rob.d.}}(\mathbf{X}^{\text{true}},\rho)$. For $\rho=0$, we have the nominal estimates, which coincide with the *true estimators*, denoted by $\hat{\boldsymbol{\mu}}_{\text{true}}(\mathbf{X}^{\text{true}})$ and $\hat{\boldsymbol{\Sigma}}_{\text{true}}(\mathbf{X}^{\text{true}})$. To calculate the robust estimators,

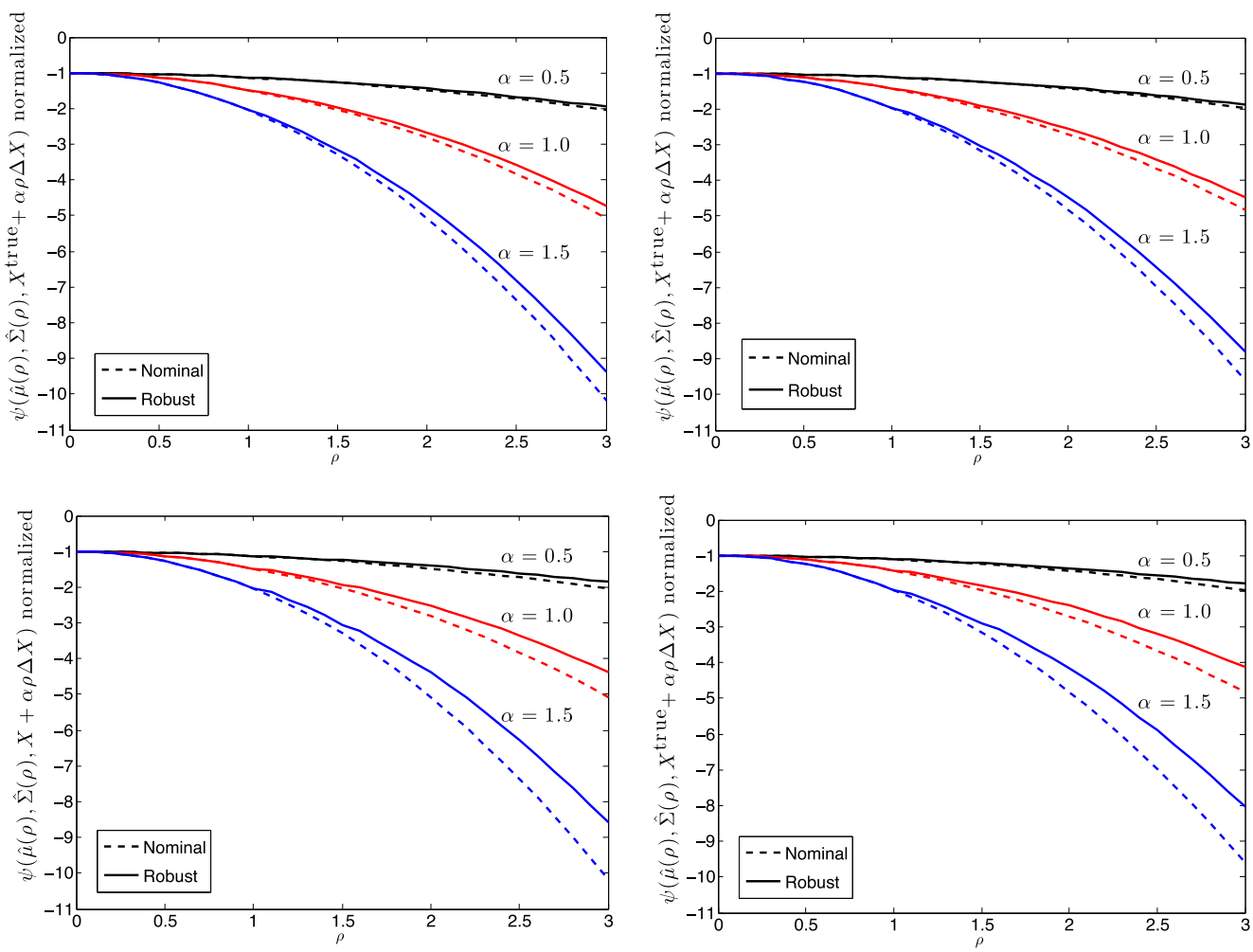
we use a first-order gradient descent method. The initial point is the robust estimate for the previous value of ρ, within the considered ρ sequence. For each computed estimate $\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}$, we determine the log-likelihood density of the observed samples $\psi(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}, \mathbf{X}^{\text{true}} + \alpha\rho\boldsymbol{\Delta}\mathbf{X}_k)$, $k = 1, 2, \ldots, 40$. We record the worst-case value as well as the average value over the set of errors indexed by k to rule out data-specific artifacts in our observations. We consider the cases $\alpha = 0.5$, $\alpha = 1.0$, and $\alpha = 1.5$.

3.1.1. ARE Case. Figure 1 (top row) shows the results in the ARE case. The worst-case and average values of the log-likelihood density function are plotted over the size of the perturbation ρ. For better comparison, these values are normalized to the corresponding unperturbed values. We observe that for small values of ρ, the nominal and robust estimators depict the same performance. As ρ grows, however, the difference between them increases, with the robust always showing a better performance than the nominal. This observation holds for both the worst-case value and the average value of the probability density. This is because the robust estimator always protects against the worst-case scenario; thus, it does not degrade as fast as the nominal estimators for increasing error sizes.

3.1.2. DRE Case. Figure 1 (bottom row) shows the performance of log-likelihood density function ψ as a function of the error size in the DRE case. We observe a similar behavior as in the ARE case. The difference between the DRE and the nominal grows at a higher rate than the difference between the ARE and the nominal. In all cases, the superiority of the robust estimator is detected for values of ρ greater than or equal to 1. The ARE is better than the nominal up to a factor of 10%, and the DRE is better than the nominal up to a factor of 15%. As α increases, both nominal and robust performances deteriorate at a higher rate.

Note the smooth transition between the nominal (i.e., $\rho = 0$ in $\psi(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}, \mathbf{X}^{\text{true}} + \alpha\rho\boldsymbol{\Delta}\mathbf{X})$) and the robust

Figure 1. (Color online) Comparison: (Left) Worst-Case and (Right) Average Log-Likelihood ψ for Changing Perturbation Size ρ for the (Top) ARE and (Bottom) DRE Cases



log-likelihood ($\rho > 0$) for both the ARE and DRE cases, suggesting the identifiability of estimators.

3.2. Distance From Nominal Estimators

To evaluate how robust estimators perform in the presence of errors, we compute both the nominal and the robust estimators on contaminated data with errors of different sizes and compare them to nominal estimators computed on the true data.

More specifically, we compute the AREs $\hat{\boldsymbol{\mu}}_{\text{rob.a.}}(\mathbf{X}^{\text{true}} + \delta\boldsymbol{\Delta}\mathbf{X}_k, \rho)$, $\hat{\boldsymbol{\Sigma}}_{\text{rob.a.}}(\mathbf{X}^{\text{true}} + \delta\boldsymbol{\Delta}\mathbf{X}_k, \rho)$ on the contaminated data for error sets $k = 1, 2, \ldots, 40$, for the values of $\delta = [0, 1]$ with 0.05 steps, and for the value of $\rho = [0, 1]$ with 0.1 steps. We also compute the DREs in the same fashion. For $\rho = 0$, we have the nominal estimators. For each estimate, Figure 2 shows the calculated distances

$$\text{Error}_{\mu} = \left\|\hat{\boldsymbol{\mu}}_{\text{rob}}(\mathbf{X}^{\text{true}} + \delta\boldsymbol{\Delta}\mathbf{X}_k, \rho) - \hat{\boldsymbol{\mu}}_{\text{true}}(\mathbf{X}^{\text{true}})\right\|_2, \quad (18)$$

$$\text{Error}_{\boldsymbol{\Sigma}} = \left\|\hat{\boldsymbol{\Sigma}}_{\text{rob}}(\mathbf{X}^{\text{true}} + \delta\boldsymbol{\Delta}\mathbf{X}_k, \rho) - \hat{\boldsymbol{\Sigma}}_{\text{true}}(\mathbf{X}^{\text{true}})\right\|_{\text{fro}}. \quad (19)$$

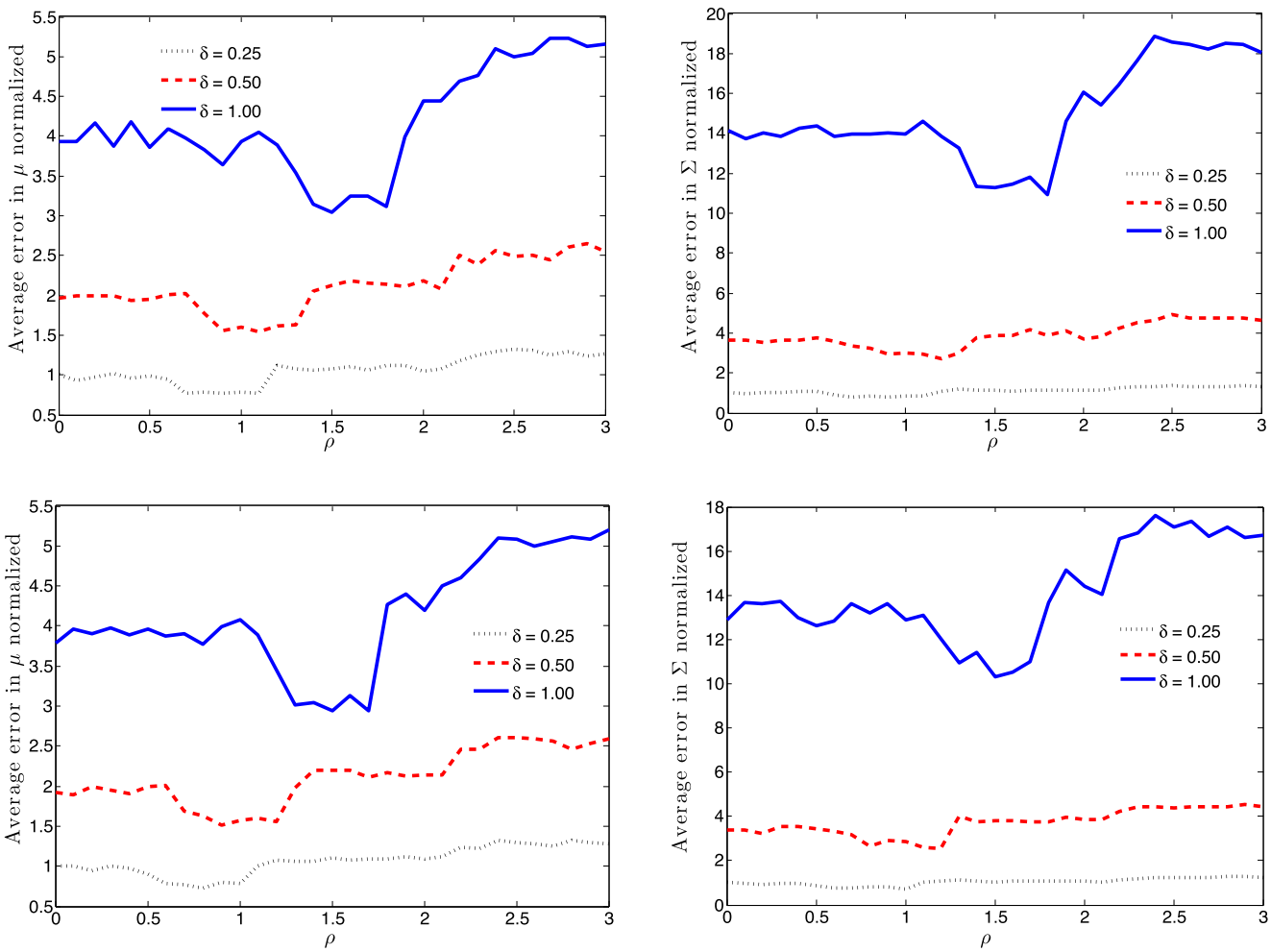
Note that $\|A\|_{\text{fro}}$ is the Frobenius norm of an $n \times m$ matrix A defined by

$$\|A\|_{\text{fro}} = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m} A_{i,j}^2},$$

where $A_{i,j}$ is the (i, j) element of matrix A. We average the calculated distances over the error sets k, $k = 1, 2, \ldots, 40$. We use the Frobenius norm because it takes into consideration the differences of the variances of the variables as well as their cross correlation terms.

The range of uncertainty size (ρ) where the robust estimator outperforms the nominal (true) depends on δ, the size of the errors, illustrated in Figure 2. When encountered errors are small sized, the robust solutions coincide with their nominal counterparts; that is, the difference between the robust and true estimators is independent of ρ. For a range of error sizes, the observable dip demonstrates that robust estimators clearly

Figure 2. (Color online) Range of Effectiveness of Robust Estimators as the Distance to True Estimators: (Left) Equation (18) for μ and (Right) Equation (19) for Σ for the (Top) ARE and (Bottom) DRE Cases



improve in accuracy. As δ increases, the interval with improved performance moves to higher ρ. This is explained by the fact that the robust estimator is secured against errors with the norm up to some ρ, and thus it cannot deal with higher errors. In data-driven applications, the analysis of the dip can be used reversely to motivate the appropriate size of the uncertainty set $\Gamma = \rho^*$.

3.3. Comparison of the Local and Global Estimators

In this section, we compare our gradient-based method with the global search algorithm. Both methods are employed to evaluate the ARE, as defined in problem (5). Because the inner problem is the same, we continue using the same algorithm for the inner problem, as described in Section 2.2, to enable a quantifiable comparison.

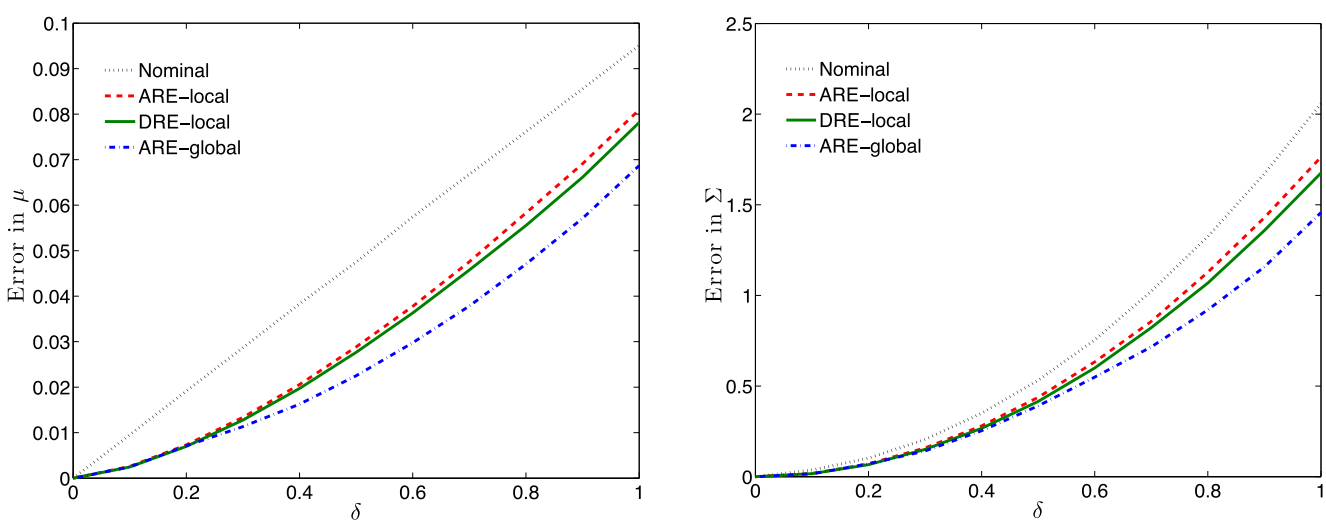
Using the same data set as before, we compare the corresponding robust estimators to the nominal (true) counterparts for various values of δ, analogous to the previous experiment. We evaluate both estimators on data contaminated with errors that are scaled with δ. For values of δ ranging between 0 and 1 with a step of 0.1, we compute the nominal (true) and the robust estimates on the contaminated data ($\mathbf{X}^{\text{true}} + \delta\mathbf{\Delta X}_k$). From the region of best performance of the robust estimator, as illustrated in Figure 2, we extract the best parameter value ρ^* that yields the lowest robust estimate errors for a particular δ, as defined in Equations (18) and (19). This ρ^* is used in the experiments to evaluate the performance.

To come close to a global optimum, we conduct the gradient-based local search from 100 different initial estimates. These are chosen as nodes of a uniform grid that covers the entire parameter space, belonging to the data set at hand. For each run, the performance is again evaluated by measuring the distance to the nominal estimators on the true data. Furthermore, for each initial point, all performances are averaged over the set of the errors used.

Figure 3 depicts the comparison between the local and the global search algorithms in the ARE case. For completeness, we also show the DRE case, which follows the same trend. These results show that as δ grows, the performance of all estimators deteriorates, but the deterioration has a smaller rate in the case of the robust estimators. The DREs show a slightly better performance than the AREs. This is because the errors for the samples in the ARE case are not correlated, whereas all sample errors in the DRE case follow the same perturbed distribution and, thus, are less conservative. For small size errors, both the local and the global as well as the ARE and DRE cases perform similarly. On the other hand, the robust simulated annealing algorithm outperforms the gradient-based local search algorithm for $\delta > 0.3$, despite the 100 initial points for the local search.

However, the global estimators come at a higher computational expense. Whereas the gradient ascent method requires approximately 100 seconds on a Intel Xeon 3.4 GHz desktop machine, the simulated annealing method terminated after approximately 215 seconds for one particular instance of the samples. Therefore, for small-sized errors, it is more meaningful to employ the gradient ascent method to evaluate local robust estimators. On the other hand, when errors are larger, using the more costly simulated annealing method leads to better-performing global robust estimators.

Figure 3. (Color online) Comparison of the Errors in (Left) $\boldsymbol{\mu}$ and (Right) $\boldsymbol{\Sigma}$ in the ARE and DRE Cases Using the Local and Global Robust Algorithms for Different Levels of Contamination δ



Note. All errors follow a normal distribution.

3.4. Comparison Between the Error Distributions

In this section, we investigate whether the above observations and, thus, the quality of the proposed robust estimators depend on the source of errors. For this, we conduct the same experiments using a different error distribution, namely, when errors are uniformly distributed. Now, $\Delta\mathbf{X}_k$, $k = 1, 2, \ldots, 40$, is a 400×4 matrix, where each of its rows follows the uniform distribution in a ball with radius 1.

When measuring the distances to the true estimators, as in (18) and (19), we observe that for the same δ, the region where the robust estimator is superior is shifted to smaller values of ρ (not shown here, but comparable to Figure 2). This is because the samples from a uniform distribution are contained in the ball with radius δ, whereas normally distributed samples can reside outside this region. Overall, we observe similar trends for the ARE and DRE cases as in the case of the normally distributed errors, as exemplified in Figure 4. Therefore, we can conclude that robust estimators are also robust to the source of uncertainty. Because we do not expect any additional insight by a comparison with global estimates, we omit the discussion here.

3.5. Comparison with Factor Analysis

In this section, we compare the proposed method with an existing method. We select an extreme case, where the distribution of the data uncertainty is fully accounted for. Consider the observed data $X^{\text{obs}} = (x_1, x_2)$ to be generated by the following factor analysis model:

$$x_1 = x^{\text{true}} + \Delta x_1 \quad \text{and} \quad x_2 = b \cdot x^{\text{true}} + \Delta x_2,$$

where x^{true} is normal distributed $\mathcal{N}(0, 1)$. The uncertain Δx_1 and Δx_2 are independent of each other and of x^{true} and are also normal distributed but with variances < 1, respectively. This means $\boldsymbol{X}^{\text{obs}}$ follows a bivariate normal distribution with zero mean, variances of 1 and b^2, and covariances equal to zero. For this strictly parametrized setting, we estimate the parameters by the proposed robust MLE method and by conventional factor analysis. In this experiment, we scale the results in terms of b for a direct comparison.

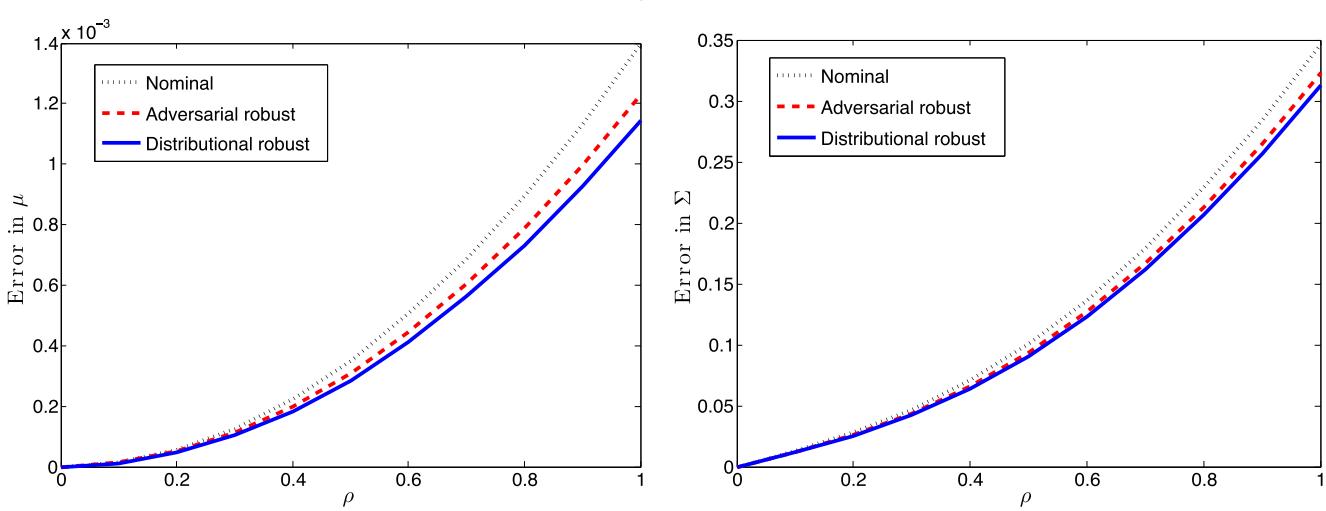
We observe that both methods comparably estimate the variances to an accuracy of $\pm b/10$. In fact, the estimator provided by robust maximum likelihood estimation improves slightly over factor analysis within an uncertainty size window, similar to the observations in Section 3.2 and displayed in Figure 2. On the other hand, factor analysis outperforms robust maximum likelihood estimation in estimating the covariances. The Frobenius norm of the difference between the estimated and generated covariance matrices is $0.6b$ smaller for factor analysis than robust maximum likelihood estimation, demonstrating the advantage of the matching parametric model. However, this advantage deteriorates when an incorrect parametric factor analysis model is used.

3.6. Summary of Simulation Results

In our computer simulations, we applied the local and the global RO methods to randomly generated data sets that were contaminated in two different ways: the ARE case and the DRE case. First, we investigated the performance of the log-likelihood density function values ψ as the error size increases. Using the local gradient-based robust method to estimate the parameters, we observed that ψ degrades at a much lower rate than when using the nominal estimators, demonstrating that the robust method protects against worst-case scenarios. This behavior was confirmed by both the worst-case performance and the average performance of ψ, as well as for both cases, the ARE and DRE cases.

Next, we compared the local and global robust algorithms by computing the distance of the robust estimates to their "true" and unperturbed counterparts.

Figure 4. (Color online) Comparison of the Errors in (Left) $\boldsymbol{\mu}$ and (Right) $\boldsymbol{\Sigma}$ for Uniformly Distributed Errors



Obviously, as the size of error increases, this distance grows for all methods. However, because this distance grows linearly for the nominal method, we observe a clear advantage using the robust methods. Because the DRE case is less conservative, it shows a slightly better performance than the ARE. As we encounter larger errors, the robust simulated annealing method finds the global robust optima and, thus, outperforms all other local methods. This comes at a higher computational cost. Independent of approach, we also studied the range of effectiveness for robust estimators, because for small-sized errors, the robust and nominal estimates coincide, and for too large errors, any method fails.

We also probed the sensitivity of the estimators with respect to the sources of errors. Using the local robust method, we compared two DRE cases: (a) when the errors follow a normal distribution and (b) when the errors follow a uniform distribution. We show that the proposed estimators remain robust even when the errors follow a different distribution than the one we prepared for.

Last, we compared the robust MLE method against a fully parametric model of factor analysis. We observe that the semiparametric robust MLE method compares well with factor analysis for estimating variances but underperforms for covariances. This demonstrates that although robust maximum likelihood estimation offers advantages to estimates of uncertain input data, it underperforms when the data-generating process is known and can be leveraged.

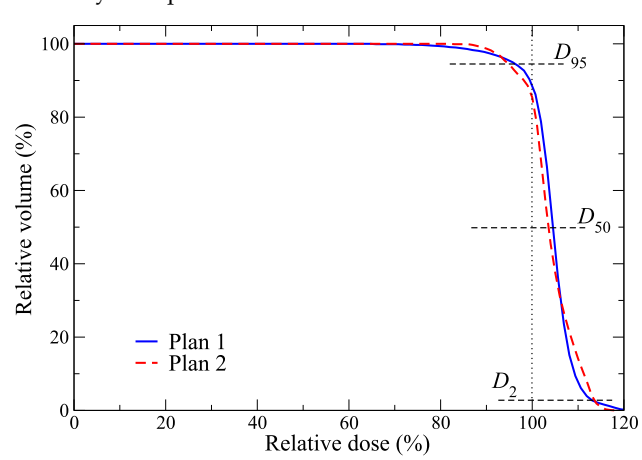
4. Cancer Radiation Therapy Plan Evaluation

Intensity-modulated radiation therapy has the advantage of delivering highly conformal dose distributions to tumors of geometrically complex shapes. The treatment-planning process involves iterative interactions between a commercial software product and a group of planners, dosimetrists and medical physicists. Therefore, the final product is the unpredictable outcome of a series of trial-and-error attempts at meeting competing *clinical objectives*. To guide the decision making and to assure plan quality, institutional and international recommendations are followed (International Commission on Radiation Units and Measurements 2010). Even though these constraints are rigorously imposed, substantial deviations can occur (Das et al. 2008, Roy et al. 2013). These often occur because some of the guidelines cannot be followed because some of them are competing or infeasible in certain cases. In practice, these deviations are statistically analyzed for both reporting and process control purposes. For this, conventional statistical estimators (sample mean and covariances) are used. However, these estimators are sensitive to sample quality and sample size, rendering the conclusions less dependable. Our goal is to demonstrate that robust estimators are more accurate in the presence of uncertainties. They can produce more reliable guidelines that can be followed in practical settings and prevent undesirable deviations.

In this section, we focus on radiation dose to tumor structures and defer the discussion on other organs to a more dosimetric study. In clinical practice, spatial dose distributions are evaluated with a cumulative dose–volume histogram (DVH). It measures the portion of the volume (e.g., of tumor) that receives a certain fraction of the prescribed dose. Figure 5 illustrates two acceptable tumor DVHs. The ideal distribution is to deliver 100% of the prescribed dose to the entire volume (dotted step function). Note that the DVH, by design, normalizes over anatomical sites and volume differences, enabling direct comparisons across patients. Guidelines are typically imposed as DVH control points, for example, for the dose D_x to a tumor. The quantity D_x measures the percentage of the prescribed dose that was received by x% of the volume and is typically implemented as a soft constraint during plan optimization. Figure 5 also shows three such control points. Despite international and institutional recommendations, the delivered values often differ from protocols (Nohadani et al. 2015).

As input data, we employ D_x control points of 491 treatment plans that have already been delivered. Specifically, the treated tumors are in the abdomen, brain, bladder, breast, eyelid, esophagus, head and neck, liver, lung, pancreas, parotid gland, pelvis, prostate, rectum, thyroid, tongue, tonsil, and vagina. This range serves to limit site-specific bias, and the use of the DVH enables comparability. Note that all treatments followed the same clinical protocols and were optimized with the same commercial software

Figure 5. (Color online) Dose–Volume Histogram of Two Clinically Acceptable Treatment Plans for Tumor



Note. The dotted line is a guide for the eye for the ideal plan, and the dashed lines mark three dose control points D_x.

(Pinnacle 2012). Therefore, the remaining sources of uncertainty are in the variations among the patients and the path of decision making that clinicians have taken to arrive at these final plans. We compute the corresponding AREs for these plans.

4.1. Summary of Results

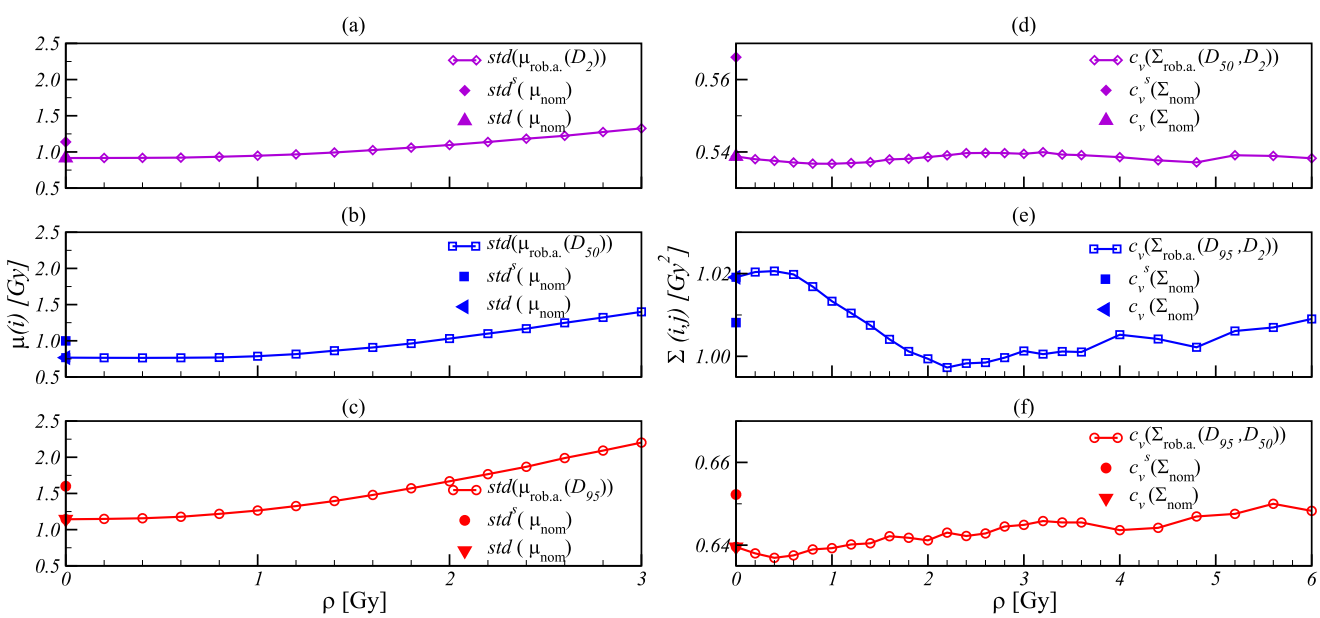
We evaluate DVH dose points D_{100}, D_{98}, D_{95}, D_{50}, and D_2 on the treated tumors. Specifically, we simulate a typical institutional analysis, where means and covariances of D_x are reported. A report on a protocol (not data) is considered reliable when the estimator values remain stable for differing sets of data. In other words, insensitivity to sample and sample size is given, when estimators exhibit a narrow spread over the sampled data (and size). In the following, we describe two experiments, probing the dependence on samples and on sample size.

To simulate semiannual report scenarios, we randomly divide the 491 five-dimensional plan data into a subset of 50. For each set, we evaluate estimators for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. First, we compute the corresponding AREs of 3,000 such samples. Next, the standard (nominal) estimators on the same samples are calculated for comparison. In the first experiment, an estimator is considered superior when its value is not sizably affected by the choice of the sample, that is, the estimator spread is narrower. In the second experiment, only 80 randomly chosen samples (each having 50 plans) are considered, and the results are denoted by superscript "s." Therefore, the difference from the subsampled estimator displays sample dependence. In both experiments, the independence among patients justifies the assumption of a normal distribution for the deviations from protocol. Therefore, the corresponding uncertainty set can be modeled as in (4).

Figure 6 (left) shows that for the three key clinical control points, the spread of $\boldsymbol{\mu}_{\text{rob.a.}}$ is stable and at the same level as the nominal estimator (at $\rho = 0$) up to $\rho \leq 2$ Gy. Beyond this point, the performance deteriorates (also observed in Section 3.2). However, with regard to sample-size dependence, $\boldsymbol{\mu}_{\text{nom}}$ exhibits a sizable sensitivity (seen by the difference $\Delta^s = |std(\boldsymbol{\mu}_{\text{nom}}) - std^s(\boldsymbol{\mu}_{\text{nom}})| > 0$), as opposed to $\boldsymbol{\mu}_{\text{rob.a.}}$, which remains unchanged (hence, not plotted). This observation applies to all DVH control points D_x. Furthermore, a deviation beyond 2 Gy is typically considered clinically unacceptable during planning. Therefore, it is justified to state that the stability of the robust estimator is superior over the clinically relevant range.

The true advantage of the robust method becomes apparent when comparing higher estimators. Figure 6 (right panels) illustrates the coefficient of variation ($c_v = \frac{\sigma}{\mu}$) that measures the normalized spread of the covariances. With regard to the sample dependence, the spread of $\boldsymbol{\Sigma}_{\text{rob.a.}}(D_{50}, D_2)$ and $\boldsymbol{\Sigma}_{\text{rob.a.}}(D_{95}, D_{50})$ is comparable to $\boldsymbol{\Sigma}_{\text{nom}}$ over all samples (see Figure 6, (d) and (f)). However, $\boldsymbol{\Sigma}_{\text{rob.a.}}(D_{95}, D_2)$, which measures the relation across the full DVH range and is clinically most meaningful, exhibits a narrower spread over the entire region, as shown in Figure 6(e)), demonstrating the superiority of $\boldsymbol{\Sigma}_{\text{rob.a.}}$ for this key metric. Note that only

Figure 6. (Color online) Stability of Estimators: Dependence on the Size of Uncertainty Set ρ



Note. The figure shows the standard deviation of the mean (left) and coefficient of variation ($c_v = \frac{\sigma}{\mu}$) of the covariance matrix elements over the samples (right).

$\mathbf{\Sigma}_{\text{rob.a.}}(D_{95}, D_2)$ (Figure 6(e)) is inferior for $\rho \leq 0.8$, which is expected, as the performance of any robust quantity sets in beyond a specific uncertainty size, also demonstrated in Figure 2 with computer-simulated data. When probing the sample-size dependence, a sizable variation is observed for the nominal estimator. Furthermore, all matrix elements of $\mathbf{\Sigma}_{\text{rob.a.}}$ display a near constant spread over an extended range of uncertainty size ρ. Also in this application, we observe a smooth transition between the nominal and the robust estimators, supporting the identifiability of estimators.

4.2. Clinical Implications

The mean of dose points D_x is often used to track both planners and the institutional performance. It is also used to analyze outcome data (survival, toxicity, and pain levels), both for monitoring and for reporting. Here, we presented three dose points for the sake of exposition (the other two revealed qualitatively comparable results). The covariance between dose points is often recorded and analyzed to track the efficacy of the guidelines. In fact, in a recent study using nominal estimators on 100 head-and-neck cases, D_{95} was found to be negatively correlated to D_{50} (Nohadani et al. 2015, Roy et al. 2016). This means that these two criteria, although recommended, cannot be satisfied concurrently. Robust estimators are suited to providing more reliable recommendations.

In medical research and practice, statistical estimators are arguably among the most prevalent quantitative tools. These results show that taking errors into account with a robust approach can significantly advance the reliability of conclusions. Furthermore, in clinical trials, large samples are required to control errors. The presented robust approach can significantly mediate the sample size and cost while ensuring the same, if not a better, level of reliability of results.

5. Conclusions

In this work, we extend the method of maximum likelihood estimation to also account for errors in input data, using the paradigm of robust optimization. We introduce two types of robust estimators to cope with adversarial and distributional errors. We show that adversarial robust estimators can be efficiently computed using a directional gradient–based algorithm. In the distributional case, we show that the inner infinite dimensional optimization problem can be solved via a finite dimensional problem, constituting a special case of the adversarial estimators. For multivariate normally distributed errors, arising in many practical cases, we develop local and global search algorithms to efficiently calculate the robust estimators.

We demonstrate the performance of the robust estimators in two types of data sets. Computer-simulated data serve for a controlled experiment. We observe that the robust estimators are significantly more protected against errors than nominal ones. For small errors, the local and global robust estimators are comparable. However, for larger errors, the global estimators outperform the local ones. Moreover, we show that the proposed estimators remain stable even when errors follow a different distribution than assumed.

In the second application, a large data set of cancer radiation therapy plans that have been delivered serve to probe the estimators in clinical decision making. We show that whereas conventional estimators lead to heavily sample-dependent conclusions, the robust estimators exhibit a narrow spread across samples. This sample independence allows for reliable decisions. Given the independence of possible data structure and the generic error models, we believe that our proposed methods are directly applicable to a wide variety of real-world maximum likelihood settings.

Acknowledgments

The authors thank Apostolos Fertis for insightful discussions and Indra Das for generously providing the clinical data.

References

Ben-Tal A, El Ghaoui L, Nemirovski A (2009) *Robust Optimization* (Princeton University Press, Princeton, NJ).

Berkson J (1950) Are there two regressions? *J. Amer. Statist. Assoc.* 45(250):164–180.

Bertsimas D, Nohadani O (2010) Robust optimization with simulated annealing. *J. Global Optim.* 48(2):323–334.

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

Bertsimas D, Nohadani O, Teo K (2010a) Nonconvex robust optimization for problems with constraints. *INFORMS J. Comput.* 22(1):44–58.

Bertsimas D, Nohadani O, Teo K (2010b) Robust optimization for unconstrained simulation-based problems. *Oper. Res.* 58(1): 161–178.

Boyd S, Vandenberghe L (2004) *Convex Optimization* (Cambridge University Press, Cambridge, UK).

Buonaccorsi JP (2010) *Measurement Error; Models, Methods and Applications* (CRC Press, Boca Raton, FL).

Calafiore G, El Ghaoui L (2001) Robust maximum likelihood estimation in the linear model. *Automatica* 37(4):573–580.

Cheng CL, Van Ness JW, et al. (1999) *Statistical Regression with Measurement Error* (Oxford University Press, New York).

Danskin JM (1966) The theory of max-min, with applications. *SIAM J. Appl. Math.* 14(4):641–664.

Das I, Desrosiers C, Srivastava S, Chopra K, Khadivi K, Taylor M, Zhao Q, Johnstone P (2008) Dosimetric variability in lung cancer IMRT/SBRT among institutions with identical treatment planning systems. *Internat. J. Radiation Oncology Biol. Phys.* 72(1): 604–605.

Delage E, Ye Y (2010) Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Oper. Res.* 58(3):595–612.

El Ghaoui L, Lebret H (1997) Robust solutions to least-squares problems with uncertain data. *SIAM J. Matrix Anal. Appl.* 18(4): 1035–1064.

Fuller WA (2009) *Measurement Error Models,* vol. 305 (John Wiley & Sons, Hoboken, NJ).

Hampel FR (1974) The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.* 69(346):383–393.

Huber PJ (1964) Robust estimation of a location parameter. *Ann. Math. Statist.* 35(1):73–101.

Huber PJ (1980) *Robust Statistics* (John Wiley & Sons, Cambridge, MA).

Huber PJ (1996) *Robust Statistical Procedures* (SIAM).

International Commission on Radiation Units and Measurements (2010) Prescribing, recording and reporting photon-beam intensity-modulated radiation therapy (IMRT). ICRU Report-83, International Commission on Radiation Units and Measurements 10(1).

Jaeckel LA (1971) Robust estimates of location: Symmetry and asymmetric contamination. *Ann. Math. Statist.* 42(3):1020–1034.

Nohadani O, Roy A, Das I (2015) Large-scale DVH quality study: Correlated aims lead relaxations. *Medical Phys.* 42(6):3457–3457.

Pfanzagl J (1994) *Parametric Statistical Theory* (Walter de Gruyter, Berlin).

Rendl F, Wolkowicz H (1997) A semidefinite framework for trust region subproblems with applications to large scale minimization. *Math. Programming* 77(1):273–299.

Roy A, Das I, Srivastava S, Nohadani O (2013) Analysis of planner dependence in IMRT. *Medical Phys.* 40(6):260.

Roy A, Das IJ, Nohadani O (2016) On correlations in IMRT planning aims. *J. Appl. Clinical Medical Phys.* 17(6):44–59.

Sra S, Nowozin S, Wright S (2012) *Optimization for Machine Learning* (MIT Press, Cambridge, MA).

Ye Y (1992) A new complexity result on minimization of a quadratic function with a sphere constraint. Floudas C, Pardalos P, eds. *Recent Advances in Global Optimization* (Princeton University Press, Princeton, NJ), 19–31.