Open access • Journal Article • DOI:10.1137/16M1075910

Distributionally Robust Optimization with Principal Component Analysis

— Source link

Jianqiang Cheng, Richard Li-Yang Chen, Habib N. Najm, Ali Pinar ...+2 more authors

Published on: 14 Jun 2018 - Siam Journal on Optimization (Society for Industrial and Applied Mathematics)

Topics: Semidefinite programming, Robust optimization, Stochastic programming, Moment problem and Duality (optimization)

Distributionally Robust Optimization with Principal Component Analysis

SIAM J. Optim.
Vol. 28, No. 2, pp. 1817–1841

DISTRIBUTIONALLY ROBUST OPTIMIZATION WITH PRINCIPAL COMPONENT ANALYSIS*

JIANQIANG CHENG†, RICHARD LI-YANG CHEN‡, HABIB N. NAJM‡, ALI PINAR‡, COSMIN SAFTA‡, AND JEAN-PAUL WATSON§

Abstract. Distributionally robust optimization (DRO) is widely used because it offers a way to overcome the conservativeness of robust optimization without requiring the specificity of stochastic programming. On the computational side, many practical DRO instances can be equivalently (or approximately) formulated as semidefinite programming (SDP) problems via conic duality of the moment problem. However, despite being theoretically solvable in polynomial time, SDP problems in practice are computationally challenging and quickly become intractable with increasing problem sizes. We propose a new approximation method to solve DRO problems with moment-based ambiguity sets. Our approximation method relies on principal component analysis (PCA) for optimal lower dimensional representation of variability in random samples. We show that the PCA approximation yields a relaxation of the original problem and derive theoretical bounds on the gap between the original problem and its PCA approximation. Furthermore, an extensive numerical study shows the strength of the proposed approximation method in terms of solution quality and runtime. As examples, for distributionally robust conditional value-at-risk and risk-averse production-transportation problems the proposed PCA approximation using only 50% of the principal components yields near-optimal solutions (within 1%) with a one to two order of magnitude reduction in computation time.

Key words. stochastic programming, distributionally robust optimization, principal component analysis, semidefinite programming

AMS subject classifications. 90C15, 90C22, 90C59

DOI. 10.1137/16M1075910

1. Introduction.

The ability of stochastic programming (SP) to incorporate uncertainty within an optimization framework is driving its increasing popularity as it caters to the needs of modeling uncertain real-world problems. There are many applications of SP in energy, transportation, and finance [5, 9, 34]. We refer the reader to Prékopa [33] and Shapiro, Dentcheva, and Ruszczński [39] for details on the theory and applications of SP. In SP, the fundamental assumption is that the probability distributions of uncertain parameters are either known or can be estimated with some degree of accuracy. However, in many real-world applications this assumption is not necessarily realistic since the probability distribution is typically unknown as it may only be indirectly observable or estimated through limited samples. And due to the limited number of samples, estimation of the uncertainty space may be biased such that the solution may be suboptimal and perform poorly out-of-sample.

*Received by the editors May 18, 2016; accepted for publication (in revised form) March 21, 2018; published electronically June 14, 2018.

http://www.siam.org/journals/siopt/28-2/M107951.html

Funding: This work was funded by the Laboratory Directed Research and Development (LDRD) program of the Sandia National Laboratories. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA-0003525.

†Department of Systems and Industrial Engineering, University of Arizona, Tucson, AZ 85721 (jqcheng@email.arizona.edu).

‡Sandia National Laboratories, Livermore, CA 94551 (rlchen@sandia.gov, hnnajm@sandia.gov, apinar@sandia.gov, csafta@sandia.gov).

§Sandia National Laboratories, Albuquerque, NM 87185 (jwatson@sandia.gov).

A practical alternative to SP is robust optimization (RO), where only the support of uncertain parameters is assumed to be available. RO ensures that solutions are feasible for all realizations of the uncertain parameters, and are therefore robust against the full variability of the uncertain parameters. RO has been applied in a broad range of areas, such as control theory and energy systems [4, 25]. For a comprehensive overview of RO theory and applications, we refer the reader to [2, 3] and references therein. However, aside from the support, RO models do not exploit additional distributional details, such as any moment information, which may be available even when the probability distribution is not fully known. It is beneficial to employ advanced RO methods that leverage available distributional information, aside from the support, to reduce conservativeness and improve solution quality.

Distributionally robust optimization (DRO), which dates back to the 1950s [35], was developed to fill this gap. In DRO, the probability distributions of uncertain parameters are assumed to belong to an ambiguity set, a family of distributions that share common properties. Since the introduction of DRO by Scarf [35], several ambiguity sets have been proposed and analyzed. Among these, three types have received significant attention: moment-based ambiguity sets, structural ambiguity sets, and metric-based ambiguity sets. In moment-based ambiguity sets, it is assumed that all distributions in the distribution family share the same moment information [8, 10, 29, 32, 43, 44]. In structural ambiguity sets, distributions share the same structural properties, such as symmetry, unimodality, and monotonicity [20, 31, 40, 41]. Metric-based ambiguity sets are created by requiring that all distributions are close to a reference (or nominal) distribution within a prespecified probability distance. The reference distribution is usually estimated using sampled data. Several types of probability distance functions have been proposed: the Prohorov metric [12], the ϕ-divergence [1, 17, 24, 26], and the Wasserstein metric [13, 14, 15]. We observe that DRO with the Wasserstein metric gained popularity recently as Wasserstein ambiguity sets provide powerful out-of-sample performance guarantees and enable decision makers to control the degree of conservativeness of the underlying optimization problem [13]. We refer the reader to [7, 13, 14, 15, 16, 36, 37] for more information on Wasserstein ambiguity sets as well as some established results on the equivalence between regularization and the corresponding DRO problem.

Leveraging conic duality for moment problems [23, 38] and developments in interior point algorithms for solving semidefinite programming (SDP) problems, many DRO problems with moment-based ambiguity sets can be reformulated equivalently (or approximately) as SDP problems. Although SDP problems can be solved theoretically in polynomial time, significant computational hurdles remain in practice for large-scale instances. Consequently, recent research has focused on reducing the size of SDP reformulations by exploiting DRO problem structure [8, 28, 29]. Natarajan, Sim, and Uichanco [29] proposed a computationally efficient second-order cone programming (SOCP) exact reformulation for a class of robust expected utility models under known mean and covariance matrix. Cheng, Delage, and Lisser [8] developed an SOCP reformulation for distributionally robust chance constrained problems by integrating independence information. Recently, Natarajan and Teo [28] developed an exact SDP reformulation that is smaller in size than existing SDP formulations for specially structured DRO problems with second-order moment information.

Another promising alternative, which we explore in this work, is the development of approximation methods that provide efficient trade-offs between solution quality and computational tractability, using smaller-size SDP formulations for solving DRO problems. The size of matrices in the SDP formulations of DRO problems increases quadratically with the number of random variables involved. Thus, one potential approximation approach would employ dimensionality reduction approaches to reduce the number of random variables under consideration. In this context, principal component analysis (PCA) is an effective technique. PCA employs a linear combination of orthogonal eigenmodes to provide an optimal representation of the variability in the data. For more information on PCA, we refer the reader to [42].

In this paper, we study DRO problems with moment-based ambiguity sets, which account for information about the support, mean, and covariance of random variables. Although the Wasserstein ambiguity set offers powerful out-of-sample performance guarantees for decision makers, the Wasserstein ambiguity set and the moment-based ambiguity set are complementary to each other, each with distinct advantages. For instance, [16] shows that DRO with moment-based ambiguity set performs better than DRO with Wasserstein ambiguity set in high-correlation regimes, while DRO with Wasserstein ambiguity set performs better in medium- and low-correlation regimes. Thus, efficiently solving DRO problems with moment-based ambiguity sets remains an important and challenging problem. We present a dimensionality reduction scheme for DRO based on PCA, which allows for direct control of the trade-offs between solution quality and computation time. However, the PCA approximation technique proposed later can not be directly applied to DRO with Wasserstein ambiguity sets.

The contributions of this paper can be summarized as follows.

1. We propose a new approximation method based on PCA to reduce the dimensionality of DRO problems with moment-based ambiguity sets. This proposed approximation framework can also be extended to more general ambiguity sets.
2. We show that the PCA approximation yields a relaxation of the original problem and quantify the impact of the number of principal components on solution quality by deriving theoretical bounds on the gap between the original problem and its PCA approximation. Moreover, we prove that the PCA approximation is exact when all the principal components are considered.
3. We demonstrate the efficacy of the proposed PCA approximation for solving large-scale problems and verify the theoretical results through a comprehensive numerical study on a distributionally robust conditional value-at-risk (CVaR) problem and a risk-averse production-transportation problem. Numerical results show that the proposed PCA approximations with only half of the principal components yield near-optimal solutions (within 1%) with a one to two order of magnitude reduction in computation time.

The remainder of this paper is organized as follows. In section 2, we present the DRO problem, develop our PCA approximation framework, and derive theoretical results. In section 3, we extend the approximation method to distributionally robust chance constrained programs. In section 4, a comprehensive numerical study on a CVaR problem and a risk-averse production-transportation problem is conducted to demonstrate the strengths of the proposed approximation method. Finally, section 5 discusses future work and concludes the paper.

2. PCA approximation for DRO problems.

In this section, we describe the reformulation strategies of [10] to transform DRO problems with moment-based ambiguity sets into equivalent SDP problems. We then describe low-rank approximations for matrices and apply one such method, principal component analysis (PCA), to effectively approximate the SDP reformulation of the DRO problem. Finally, we present theoretical bounds on the quality of the proposed PCA approximation.

2.1. SDP reformulation for DRO problems. We consider a stochastic optimization problem with the following form:

$$\underset{\boldsymbol{x}\in X}{\text{minimize}}\ \mathbf{E}_F f(\boldsymbol{x},\xi), \tag{2.1}$$

where $\boldsymbol{x} \in \mathbb{R}^n$, $X \subset \mathbb{R}^n$ is a convex set, $\xi \in \mathbb{R}^m$ is random vector with a distribution F, and $f(\boldsymbol{x},\xi)$ is a cost function that is convex in $\boldsymbol{x}$ for a given ξ. In stochastic optimization, it is typically assumed that distribution F is known exactly. However, this assumption is overly restrictive in many cases. In many practical settings it is difficult to infer the exact distribution given limited data samples. In such cases, it may be necessary to work with only partial information on the distribution F, (e.g., its support and some moments). In other words, distribution F belongs to some ambiguity set $\mathcal{D}$ that encompasses the partial information. Under a robust optimization framework, we can consider the worst-case result of the stochastic optimization problem, namely the distributionally robust optimization problem, as follows:

$$\underset{\boldsymbol{x}\in X}{\text{minimize}}\ \underset{F\in\mathcal{D}}{\text{maximize}}\,\mathbf{E}_F f(\boldsymbol{x},\xi). \tag{2.2}$$

Here (2.2) can be interpreted as a risk-averse (conservative) approximation of problem (2.1). In the remainder of the paper, we focus on the moment-based ambiguity set, where information on the support, mean, and covariance of ξ is known explicitly.

Assumption 1. The distributional ambiguity set, $\mathcal{D}(\mathcal{S},\mu,\Sigma)$, accounts for information about the convex support $\mathcal{S}$, mean μ in the strict interior of $\mathcal{S}$, and an upper bound $\Sigma \succ 0$ (positive definite) on the covariance matrix of the random vector ξ, i.e.,

$$\mathcal{D}(\mathcal{S},\mu,\Sigma) = \left\{ F \;\middle|\; \begin{array}{l} \mathbf{P}(\xi\in\mathcal{S}) = 1, \\ \mathbf{E}_F[\xi] = \mu, \\ \mathbf{E}_F[(\xi-\mu)(\xi-\mu)^T] \preceq \Sigma \end{array} \right\}. \tag{2.3}$$

Remark 1. An extension to a more general ambiguity set—for instance, the mean of ξ lies in an ellipsoid with center μ (see [10] for details)—is straightforward and is omitted to simplify the introduction of the proposed approximation method. If Σ is not positive definite, that is $\text{rank}(\Sigma) < m$, then ξ can be represented by a linear combination of a subset of ξ with size $\text{rank}(\Sigma)$ [27]. Thus, we replace ξ by a lower dimensional vector of the problem under consideration and the assumption is satisfied.

Using ambiguity set $\mathcal{D}(\mathcal{S},\mu,\Sigma)$, Delage and Ye [10] showed that problem (2.2) can be reformulated as a semidefinite programming problem because of the conic duality for moment problems.

THEOREM 2.1. *Under Assumption* 1*, if $f(\boldsymbol{x},\xi)$ is F-integrable for any $F \in \mathcal{D}$, then problem* (2.2) *has the same optimal value as the following problem:*

$$\underset{\boldsymbol{x},s,\boldsymbol{q},\boldsymbol{Q}}{\text{minimize}}\quad s + \mu^T\boldsymbol{q} + (\Sigma + \mu\mu^T)\bullet\boldsymbol{Q} \tag{2.4a}$$

$$\textit{subject to}\quad s + \xi^T\boldsymbol{q} + \xi^T\boldsymbol{Q}\xi \geq f(\boldsymbol{x},\xi)\ \forall\xi\in\mathcal{S}, \tag{2.4b}$$

$$\boldsymbol{Q} \succeq 0,\ \boldsymbol{x}\in X, \tag{2.4c}$$

where $s \in \mathbb{R}$, $\boldsymbol{q} \in \mathbb{R}^m$, $\boldsymbol{Q} \in \mathbb{R}^{m\times m}$, and "$\bullet$" is the inner product defined by $A \bullet B = \sum_{i,j} A_{ij}B_{ij}$, where A and B are two conformal matrices.

Proof. The result can be deduced from Lemma 1 in [10] . ◻

COROLLARY 2.2. *When the support of ξ is polyhedral with at least one interior point, i.e., $\mathcal{S} = \{\xi | A\xi \le b\} \neq \emptyset$ with $A \in \mathbb{R}^{n_1 \times m}$ and $b \in \mathbb{R}^{n_1}$, if $f(\boldsymbol{x}, \xi)$ is a piecewise linear convex function in ξ (precisely, $f(\boldsymbol{x}, \xi) = \max_{k=1}^{K} y_k^0(\boldsymbol{x}) + \boldsymbol{y}_k(\boldsymbol{x})^T \xi$, where $\boldsymbol{y}_k(\boldsymbol{x}) = [y_k^1(\boldsymbol{x}), \ldots, y_k^m(\boldsymbol{x})]^T$ as well as $y_k^0(\boldsymbol{x})$ are affine in $\boldsymbol{x}$ for $k = 1, \ldots, K$), problem (2.4) is reduced to the following problem:*

$$\text{(2.5a)} \qquad \underset{\boldsymbol{x}, s, \boldsymbol{q}, \boldsymbol{\lambda}, \boldsymbol{Q}}{\text{minimize}} \quad s + \mu^T \boldsymbol{q} + (\Sigma + \mu\mu^T) \bullet \boldsymbol{Q}$$

$$\text{(2.5b)} \qquad \textit{subject to} \quad \begin{bmatrix} s - y_k^0(\boldsymbol{x}) - \boldsymbol{\lambda}_k^T b & \frac{(\boldsymbol{q} - \boldsymbol{y}_k(\boldsymbol{x}) + A^T \boldsymbol{\lambda}_k)^T}{2} \\ \frac{\boldsymbol{q} - \boldsymbol{y}_k(\boldsymbol{x}) + A^T \boldsymbol{\lambda}_k}{2} & \boldsymbol{Q} \end{bmatrix} \succeq 0 \ \forall k \in \{1, \ldots, K\},$$

$$\text{(2.5c)} \qquad \boldsymbol{Q} \succeq 0, \ \boldsymbol{\lambda}_k \ge 0 \ \forall k \in \{1, \ldots, K\}, \ \boldsymbol{x} \in X,$$

where $\boldsymbol{\lambda}_k \in \mathbb{R}^{n_1}, \ k = 1, \ldots, K.$

Proof. The basic idea of the proof is to apply the strong duality theorem to constraint (2.4b). Let

$$\mathbf{Z}_k = \begin{bmatrix} s - y_k^0(\boldsymbol{x}) - \boldsymbol{\lambda}_k^T b & \frac{1}{2}(\boldsymbol{q} - \boldsymbol{y}_k(\boldsymbol{x}) + A^T \boldsymbol{\lambda}_k)^T \\ \frac{1}{2}(\boldsymbol{q} - \boldsymbol{y}_k(\boldsymbol{x}) + A^T \boldsymbol{\lambda}_k) & \boldsymbol{Q} \end{bmatrix}.$$

As $f(\boldsymbol{x}, \xi)$ is a piecewise linear convex function, constraint (2.4b) is reformulated as

$$\text{(2.6)} \qquad s + \xi^T \boldsymbol{q} + \xi^T \boldsymbol{Q} \xi \ge y_k^0(\boldsymbol{x}) + \boldsymbol{y}_k(\boldsymbol{x})^T \xi \quad \forall \xi \in \mathcal{S}, \ \forall k \in \{1, 2, \ldots, K\}.$$

Let $g_k(\xi) := s + \xi^T \boldsymbol{q} + \xi^T \boldsymbol{Q} \xi - y_k^0(\boldsymbol{x}) - \boldsymbol{y}_k(\boldsymbol{x})^T \xi$. Then constraint (2.6) is equivalent to $\text{minimize}_{A\xi \le b, \xi \in \mathbb{R}^m} \, g_k(\xi) \ge 0$. Further, we consider the Lagrange dual problem of $\text{minimize}_{A\xi \le b} \, g_k(\xi)$, i.e., $\text{maximize}_{\boldsymbol{\lambda}_k \ge 0} \inf_{\xi} g_k(\xi) + \boldsymbol{\lambda}_k^T (A\xi - b)$, where $\boldsymbol{\lambda}_k \in \mathbb{R}^{n_1}$. Since function $g_k(\xi)$ is convex in ξ, and together with the assumption that there exists an interior point for the primal problem, we have that constraint (2.6) is equivalent to the following one:

$$\text{(2.7)} \qquad \underset{\boldsymbol{\lambda}_k \ge 0}{\text{maximize}} \ \inf_{\xi \in \mathbb{R}^m} g_k(\xi) + \boldsymbol{\lambda}_k^T (A\xi - b) \ge 0 \quad \forall k \in \{1, 2, \ldots, K\}.$$

Further, constraint (2.7) is equivalent to the following:

$$\exists \boldsymbol{\lambda}_k \ge 0, \ s + \xi^T \boldsymbol{q} + \xi^T \boldsymbol{Q} \xi - y_k^0(\boldsymbol{x}) - \boldsymbol{y}_k(\boldsymbol{x})^T \xi + \boldsymbol{\lambda}_k^T (A\xi - b) \ge 0 \ \forall \xi \in \mathbb{R}^m, \ \forall k$$

$$\text{(2.8)} \qquad \Leftrightarrow \exists \boldsymbol{\lambda}_k \ge 0, \ \begin{bmatrix} 1 & \xi^T \end{bmatrix} \mathbf{Z}_k \begin{bmatrix} 1 \\ \xi \end{bmatrix} \ge 0 \ \forall \xi \in \mathbb{R}^m, \ \forall k$$

$$\text{(2.9)} \qquad \Leftrightarrow \exists \boldsymbol{\lambda}_k \ge 0, \ \mathbf{Z}_k \succeq 0 \ \forall k,$$

where the first equivalence is direct from the definition of $\mathbf{Z}_k$ and we next prove the latter equivalence. First, $\Leftarrow$ follows directly from the definition of positive semidefiniteness of a matrix. Then we prove $\Rightarrow$. For any $[\eta_0; \eta] \in \mathbb{R}^{m+1}$, where $\eta_0 \in \mathbb{R}$ and $\eta \in \mathbb{R}^m$, there are two cases: $\eta_0 = 0$ and $\eta_0 \neq 0$. When $\eta_0 = 0$, we have

$$\begin{bmatrix} \eta_0 & \eta^T \end{bmatrix} \mathbf{Z}_k \begin{bmatrix} \eta_0 \\ \eta \end{bmatrix} = \eta^T \boldsymbol{Q} \eta \ge 0,$$

where the inequality results from the positive semidefiniteness of $\boldsymbol{Q}$. When $\eta_0 \neq 0$, we have

$$\left[\begin{array}{cc} \eta_0 & \eta^T \end{array}\right] \mathbf{Z}_k \left[\begin{array}{c} \eta_0 \\ \eta \end{array}\right] = \eta_0^2 \left[\begin{array}{cc} 1 & \frac{\eta^T}{\eta_0} \end{array}\right] \mathbf{Z}_k \left[\begin{array}{c} 1 \\ \frac{\eta^T}{\eta_0} \end{array}\right] \geq 0,$$

where the inequality is due to (2.8). Therefore, the conclusion of equivalence follows.□

To distinguish (2.5) from subsequent approximation schemes, hereafter we will refer to it as the original reformulation. Problem (2.5) is an SDP problem, and is thus theoretically solvable in polynomial time. In practice, however, computational challenges remain when the problem size is large. In order to reduce the size of the moment inequality constraint, we employ principal component analysis (PCA) and retain only the most important principal components (PCs). In what follows, we first present the PCA approach from an optimization perspective.

2.2. Low-rank approximation for matrices. At the heart of our proposed methods is exploiting the lower dimensional structure inherent in matrices in practical applications. From a practical perspective, our goal is to approximate a large matrix by a lower dimensional matrix to reduce problem sizes. The literature on approximating a matrix by a lower dimensional matrix dates back to the seminal paper by Eckart and Young [11]. Eckart and Young showed that given an $m \times n$ matrix A, the problem

$$\min_{\widehat{A}} \quad \|A - \widehat{A}\|_{\mathrm{F}} \quad \text{subject to} \quad \operatorname{rank}(\widehat{A}) \leq r \leq n$$

can be solved using singular value decomposition of the matrix A. Let $A = U\Lambda V^T$ be the singular value decomposition of A such that Λ is a diagonal matrix with entries λ_i and $\lambda_1 \geq \lambda_2, \geq \cdots \geq \lambda_m$. Let u_i and v_i correspond to the columns of U and V, respectively. Define $A_r = \sum_{i=1}^r u_i \lambda_i v_i^T$. A_r will be an optimal approximation for A for the Frobenius and spectral norms. That is, A_r minimizes $\|A - \widehat{A}\|_F$ and $\|A - \widehat{A}\|_2$.

In the remainder of this paper, we will use singular value decompositions (or, specifically for our case, eigendecompositions) to approximate matrices in lower dimensions. As stated earlier, low-rank approximations are an active area of research and many other approximations have been proposed based on alternative objective functions and constraints. Our proposed approach is generalizable to these alternative low-rank approximation techniques. Our goal in this paper is to demonstrate how low-rank approximations can be adopted in distributionally robust optimization, and thus in this paper we will restrict our discussions to the most commonly used technique for low-rank matrix approximations.

2.3. Low-rank approximation for DRO. The eigendecomposition of the positive definite matrix Σ can be expressed as follows:

$$\Sigma = U\Lambda U^T = U\Lambda^{1/2}(U\Lambda^{1/2})^T,$$

where $U \in \mathbb{R}^{m\times m}$, $\Lambda \in \mathbb{R}^{m\times m}$ is a diagonal matrix, and $\Lambda^{1/2}$ replaces diagonal entries of Λ with their square roots. Without loss of generality, it is assumed that the diagonal elements of Λ are arranged in decreasing order. We introduce another random vector $\xi_I \in \mathbb{R}^m$ whose ambiguity set is

$$\mathcal{D}_I(\mathcal{S}_I, \mu_I, \Sigma_I) = \left\{ F_I \;\middle|\; \begin{array}{l} \mathbf{P}(\xi_I \in \mathcal{S}_I) = 1, \\ \mathbf{E}_{F_I}[\xi_I] = \mathbf{0}_m, \\ \mathbf{E}_{F_I}[(\xi_I)(\xi_I)^T] \preceq \mathbf{I}_m \end{array} \right\},$$

where $\mathcal{S}_I := \{\xi_I \in \mathbb{R}^m : U\Lambda^{1/2}\xi_I + \mu \in \mathcal{S}\}$, $\mathbf{0}_m \in \mathbb{R}^m$ is a vector of zeros, and $\mathbf{I}_m$ is the identity matrix of size m.

Lemma 2.3. *If $f(\boldsymbol{x}, U\Lambda^{1/2}\xi_I + \mu)$ is F_I-integrable for any $F_I \in \mathcal{D}_I$, then the original problem* (2.2) *has the same optimal value as*

$$\underset{\boldsymbol{x}\in X}{\text{minimize}}\ \underset{F_I\in\mathcal{D}_I}{\text{maximize}}\, \mathbf{E}_{F_I} f(\boldsymbol{x}, U\Lambda^{1/2}\xi_I + \mu). \tag{2.10}$$

Proof. Letting $\bar{\xi} := U\Lambda^{1/2}\xi_I + \mu$, we denote the distribution of $\bar{\xi}$ by $\bar{F}$. First, we prove that $\bar{F} \in \mathcal{D}$ for any $\xi_I \sim F_I \in \mathcal{D}_I$. Because $F_I \in \mathcal{D}_I$, we have $\mathbf{E}_{F_I}[\bar{\xi}] = \mu$ and $\mathbf{E}_{F_I}[(\bar{\xi}-\mu)(\bar{\xi}-\mu)^T] \preceq U\Lambda U^T = \Sigma$. Moreover, as $\mathcal{S}_I = \{\xi_I \in \mathbb{R}^m : U\Lambda^{1/2}\xi_I + \mu \in \mathcal{S}\}$, $\xi_I \in \mathcal{S}_I$ implies that $\bar{\xi} \in \mathcal{S}$ and thus $\mathbf{P}(\bar{\xi} \in \mathcal{S}) \geq \mathbf{P}(\xi_I \in \mathcal{S}_I) = 1$. As a consequence, we have $\mathbf{P}(\bar{\xi} \in \mathcal{S}) = 1$. Therefore $\bar{F} \in \mathcal{D}$.

Second, for any $\xi \sim F \in \mathcal{D}$, we prove that there exists a random vector $\xi_I \sim F_I \in \mathcal{D}_I$ such that $\xi = U\Lambda^{1/2}\xi_I + \mu$. For any $\xi \sim F \in \mathcal{D}$, we first construct a vector $\xi_I = (U\Lambda^{1/2})^{-1}(\xi - \mu)$. It is straightforward to have $\mathbf{E}_{F_I}[\xi_I] = \mathbf{0}_m$ and

$$\begin{aligned}\mathbf{E}_{F_I}[(\xi_I)(\xi_I)^T] &= \mathbf{E}_F[(U\Lambda^{1/2})^{-1}(\xi-\mu)(\xi-\mu)^T((U\Lambda^{1/2})^{-1})^T] \\ &\preceq (U\Lambda^{1/2})^{-1}\Sigma((U\Lambda^{1/2})^{-1})^T \\ &\preceq \mathbf{I}_m.\end{aligned}$$

Moreover, $\xi_I = (U\Lambda^{1/2})^{-1}(\xi - \mu)$ implies that $\xi = U\Lambda^{1/2}\xi_I + \mu$. As $\xi \sim F \in \mathcal{D}$, $\mathbf{P}(U\Lambda^{1/2}\xi_I + \mu \in \mathcal{S}) = \mathbf{P}(\xi \in \mathcal{S}) = 1$. Following the definition of $\mathcal{S}_I$, $U\Lambda^{1/2}\xi_I + \mu \in \mathcal{S}$ implies that $\xi_I \in \mathcal{S}_I$ and thus $\mathbf{P}(\xi_I \in \mathcal{S}_I) \geq \mathbf{P}(U\Lambda^{1/2}\xi_I + \mu \in \mathcal{S}) = 1$. Accordingly, we have $\mathbf{P}(\xi_I \in \mathcal{S}_I) = 1$. Therefore, we conclude that $\xi_I \sim F_I \in \mathcal{D}_I$ and $\xi = U\Lambda^{1/2}\xi_I + \mu$.

Altogether, the proof is complete. □

By Lemma 2.3, problem (2.10) is also equivalent to the original problem (2.2). We now introduce an approximation of problem (2.10). Relying on the ideas behind PCA, we use the leading m_1 ($m_1 \leq m$) random variables of ξ_I, capturing the dominant variability of $U\Lambda^{1/2}\xi_I$ in ξ_I, to approximate ξ. Specifically, $\xi \approx U\Lambda^{1/2}[\xi_r; \mathbf{0}_{m-m_1}] + \mu = U_{m\times m_1}\Lambda_{m_1}^{1/2}\xi_r + \mu$, where $\xi_r \in \mathbb{R}^{m_1}$ is the m_1-dimensional subvector of ξ_I, $\mathbf{0}_{m-m_1} \in \mathbb{R}^{m-m_1}$ is a vector whose elements are zero, $U_{m\times m_1} \in \mathbb{R}^{m\times m_1}$ is the $m \times m_1$ upper-left submatrix of U, and $\Lambda_{m_1}^{1/2}$ is the $m_1 \times m_1$ upper-left submatrix of $\Lambda^{1/2}$. Then we have the following approximation to (2.10):

$$\underset{\boldsymbol{x}\in X}{\text{minimize}}\ \underset{F_r\in\mathcal{D}_r}{\text{maximize}}\, \mathbf{E}_{F_r} f(\boldsymbol{x}, U_{m\times m_1}\Lambda_{m_1}^{1/2}\xi_r + \mu), \tag{2.11}$$

where

$$\mathcal{D}_r(\mathcal{S}_r, \mu_r, \Sigma_r) = \left\{ F_r \;\middle|\; \begin{array}{l} \mathbf{P}(\xi_r \in \mathcal{S}_r) = 1, \\ \mathbf{E}_{F_r}[\xi_r] = \mathbf{0}_{m_1}, \\ \mathbf{E}_{F_r}[(\xi_r)(\xi_r)^T] \preceq \mathbf{I}_{m_1} \end{array} \right\} \tag{2.12}$$

and

$$\mathcal{S}_r := \{\xi_r \in \mathbb{R}^{m_1} : U_{m\times m_1}\Lambda_{m_1}^{1/2}\xi_r + \mu \in \mathcal{S}\}. \tag{2.13}$$

THEOREM 2.4. *If $f(\boldsymbol{x}, U_{m\times m_1}\Lambda_{m_1}^{1/2}\xi_r+\mu)$ is F_r-integrable for any $F_r\in\mathcal{D}_r$, then problem* (2.11) *has the same optimal value as the following problem:*

$$\begin{aligned} &(2.14)\\ &Z^*(m_1) := \underset{\boldsymbol{x},s,\boldsymbol{q}_r,\boldsymbol{Q}_r}{\text{minimize}} \quad s+\mathbf{I}_{m_1}\bullet\boldsymbol{Q}_r\\ &\qquad\qquad\quad \textit{subject to} \quad s+\xi_r^T\boldsymbol{q}+\xi_r^T\boldsymbol{Q}_r\xi_r \ge f(\boldsymbol{x}, U_{m\times m_1}\Lambda_{m_1}^{1/2}\xi_r+\mu)\ \forall\xi_r\in\mathcal{S}_r,\\ &\qquad\qquad\qquad\qquad\quad \boldsymbol{Q}_r\succeq 0,\ \boldsymbol{x}\in X,\end{aligned}$$

where $s\in\mathbb{R}$, $\boldsymbol{q}_r\in\mathbb{R}^{m_1}$, and $\boldsymbol{Q}_r\in\mathbb{R}^{m_1\times m_1}$. We also have the following.

1. *The optimal value of problem* (2.14) *is a lower bound for that of problem* (2.4). *In other words, problem* (2.11) *(the PCA approximation) is a relaxation of problem* (2.2).
2. *The optimal value $Z^*(m_1)$ is a nondecreasing function of m_1, i.e., $Z^*(m_1)\le Z^*(m_2)$ if $m_2\ge m_1$;*
3. *If $m_1=m$, the two problems have the same optimal values. In other words, the PCA approximation provides an exact reformulation of problem* (2.2).

Proof. We start with the deterministic reformulation, whose proof is the same as that of Theorem 2.1. Let $\zeta=U_{m\times m_1}\Lambda_{m_1}^{1/2}\xi_r+\mu$ and let $\mathcal{S}_\zeta$ and $\mathcal{D}_\zeta$ denote its support and ambiguity set, respectively. As $\mathcal{S}_r=\{\xi_r\in\mathbb{R}^{m_1}: U_{m\times m_1}\Lambda_{m_1}^{1/2}\xi_r+\mu\in\mathcal{S}\}$ and $\mathcal{S}_\zeta=\{\zeta\in\mathbb{R}^m:\zeta=U_{m\times m_1}\Lambda_{m_1}^{1/2}\xi_r+\mu,\xi_r\in\mathcal{S}_r\}$, we can deduce $\mathcal{S}_\zeta\subset\mathcal{S}$. We also have $\mathbf{E}[\zeta]=\mu$ and

$$\begin{aligned}\mathbf{E}[(\zeta-\mu)(\zeta-\mu)^T] &\preceq U_{m\times m_1}\Lambda_{m_1}U_{m\times m_1}^T\\ &= U\begin{bmatrix}\Lambda_{m_1} & \mathbf{0}_{m_1,m-m_1}\\ \mathbf{0}_{m-m_1,m_1} & \mathbf{0}_{m-m_1,m-m_1}\end{bmatrix}U^T\preceq U\Lambda U^T=\Sigma,\end{aligned}$$

where $\mathbf{0}_{n,m}$ is a zero matrix (all of whose elements are zero) of size $n\times m$. Thus the ambiguity set of ζ lies in $\mathcal{D}$ (the ambiguity set of ξ). Thus, we have $\mathcal{D}_\zeta\subset\mathcal{D}$ and, further,

$$\underset{F_\zeta\in\mathcal{D}_\zeta}{\text{maximize}}\,\mathbf{E}_{F_\zeta}f(\boldsymbol{x},\zeta)\le\underset{F\in\mathcal{D}}{\text{maximize}}\,\mathbf{E}_F f(\boldsymbol{x},\xi).$$

Therefore the optimal value of problem (2.11) is a lower bound for that of problem (2.2).

Secondly, let $\zeta_1=U_{m\times m_1}\Lambda_{m_1}^{1/2}\xi_{r_1}+\mu$ and $\zeta_2=U_{m\times m_2}\Lambda_{m_2}^{1/2}\xi_{r_2}+\mu$, where $\xi_{r_1}\in\mathbb{R}^{m_1}$, $\xi_{r_2}\in\mathbb{R}^{m_2}$ for $m_2>m_1$. We denote the ambiguity set and support of ξ_{r_i}, $i=1,2$, by $\mathcal{D}_{r_i}$ (defined as in (2.12)) and $\mathcal{S}_{r_i}$ (defined as in (2.13)), respectively, while accordingly we denote the ambiguity set of ζ_i, $i=1,2$, by $\mathcal{D}_{\zeta_i}$, i.e.,

$$\mathcal{D}_{\zeta_i}=\{F_{\zeta_i}|\zeta_i\sim F_{\zeta_i},\zeta_i=U_{m\times m_i}\Lambda_{m_i}^{1/2}\xi_{r_i}+\mu,\xi_{r_i}\sim F_{r_i}\in\mathcal{D}_{r_i}\}.$$

For any $\zeta_1\sim F_{\zeta_1}\in\mathcal{D}_{\zeta_1}$, there exists a $\xi_{r_1}\sim F_{r_1}\in\mathcal{D}_{r_1}$ such that $\zeta_1=U_{m\times m_1}\Lambda_{m_1}^{1/2}\xi_{r_1}+\mu=U_{m\times m_2}\Lambda_{m_2}^{1/2}[\xi_{r_1};\mathbf{0}_{m_2-m_1}]+\mu$. Let $\bar{\xi}_{r_2}=[\xi_{r_1};\mathbf{0}_{m_2-m_1}]\in\mathbb{R}^{m_2}$.

Following the definition of $\mathcal{S}_{r_1}$, we have

$$\mathbf{P}\{\xi_{r_1}\in\mathcal{S}_{r_1}\}=\mathbf{P}\{U_{m\times m_1}\Lambda_{m_1}^{1/2}\xi_{r_1}+\mu\in\mathcal{S}\ \}=1.$$

Due to $U_{m\times m_1}\Lambda_{m_1}^{1/2}\xi_{r_1}=U_{m\times m_2}\Lambda_{m_2}^{1/2}\bar{\xi}_{r_2}$, we have $\mathbf{P}\{U_{m\times m_2}\Lambda_{m_2}^{1/2}\bar{\xi}_{r_2}+\mu\in\mathcal{S}\ \}=1$, which implies that $\mathbf{P}\{\bar{\xi}_{r_2}\in\mathcal{S}_{r_2}\}=1$ by the definition of $\mathcal{S}_{r_2}$. Moreover, we note that

$\mathbf{E}[\bar{\xi}_{r_2}] = \mathbf{E}[\xi_{r_1}; \mathbf{0}_{m_2-m_1}] = \mathbf{0}_{m_2}$ and

$$\mathbf{E}[\bar{\xi}_{r_2}(\bar{\xi}_{r_2})^T] = \mathbf{E}[\xi_{r_1}; \mathbf{0}_{m_2-m_1}]([\xi_{r_1}; \mathbf{0}_{m_2-m_1}])^T = \begin{bmatrix} \mathbf{E}[\xi_{r_1}(\xi_{r_1})^T] & \mathbf{0}_{m_1,m_2-m_1} \\ \mathbf{0}_{m_2-m_1,m_1} & \mathbf{0}_{m_2-m_1,m_2-m_1} \end{bmatrix} \preceq \mathbf{I}_{m_2}.$$

Consequently, the distribution of $\bar{\xi}_{r_2}$ belongs to the ambiguity set $\mathcal{D}_{r_2}$ and thus $F_{\zeta_1} \in \mathcal{D}_{\zeta_2}$. Therefore, we have $\mathcal{D}_{\zeta_1} \subset \mathcal{D}_{\zeta_2}$ and

$$\underset{F_{\zeta_1} \in \mathcal{D}_{\zeta_1}}{\text{maximize}}\, \mathbf{E}_{F_{\zeta_1}} f(\boldsymbol{x}, \zeta_1) \leq \underset{F_{\zeta_2} \in \mathcal{D}_{\zeta_2}}{\text{maximize}}\, \mathbf{E}_{F_{\zeta_2}} f(\boldsymbol{x}, \zeta_2).$$

Finally, when $m_1 = m$, problem (2.11) is exactly the same as problem (2.10). Then by Lemma 2.3, the PCA approximation yields an exact reformulation of problem (2.2). ☐

Like Corollary 2.2, for some cases of $f(\boldsymbol{x}, \xi)$ and the support, we have a simpler deterministic version.

COROLLARY 2.5. *When the support of ξ is polyhedral with at least one interior point, i.e., $\mathcal{S} = \{\xi | A\xi \leq b\}$ with $A \in \mathbb{R}^{n_1 \times m}$ and $b \in \mathbb{R}^{n_1}$, if $f(\boldsymbol{x}, \xi)$ is a piecewise linear convex function (precisely, $f(\boldsymbol{x}, \xi) = \max_{k=1}^{K} y_k^0(\boldsymbol{x}) + \boldsymbol{y}_k(\boldsymbol{x})^T \xi$, where $\boldsymbol{y}_k(\boldsymbol{x}) = [y_k^1(\boldsymbol{x}), \ldots, y_k^m(\boldsymbol{x})]$ as well as $y_k^0(\boldsymbol{x})$ are affine in $\boldsymbol{x}$ for $k = 1, \ldots, K$), problem* (2.14) *is reduced to the following problem:*

$$Z^*(m_1) = \underset{\boldsymbol{x}, s, \boldsymbol{q}_r, \boldsymbol{\lambda}, \boldsymbol{Q}_r}{\text{minimize}}\; s + \mathbf{I}_{m_1} \bullet \boldsymbol{Q}_r \tag{2.15a}$$

(2.15b)
subject to

$$\begin{bmatrix} s - y_k^0(\boldsymbol{x}) - \boldsymbol{\lambda}_k^T b - \boldsymbol{y}_k(\boldsymbol{x})^T \mu + \boldsymbol{\lambda}_k^T A \mu & \frac{(\boldsymbol{q}_r + (U_{m \times m_1} \Lambda_{m_1}^{1/2})^T (A^T \boldsymbol{\lambda}_k - \boldsymbol{y}_k(\boldsymbol{x})))^T}{2} \\ \frac{\boldsymbol{q}_r + (U_{m \times m_1} \Lambda_{m_1}^{1/2})^T (A^T \boldsymbol{\lambda}_k - \boldsymbol{y}_k(\boldsymbol{x}))}{2} & \boldsymbol{Q}_r \end{bmatrix} \succeq 0$$

$\forall k \in \{1, 2, \ldots, K\}$,

$$\boldsymbol{Q}_r \succeq 0,\; \boldsymbol{\lambda}_k \geq 0,\; \forall k \in \{1, \ldots, K\},\; \boldsymbol{x} \in X, \tag{2.15c}$$

where $\boldsymbol{\lambda}_k \in \mathbb{R}^{n_1},\; k = 1, \ldots, K$.

Proof. The idea of the proof is the same as that of Corollary 2.2. ☐

Remark 2. Corollary 2.5 can be extended to more general supports, such as those defined through semidefinite representable inequalities (e.g., an ellipsoid support [8]).

2.4. Quality of PCA solutions. By Theorem 2.4, the PCA approximation yields a relaxation to the original problem. In other words, the optimal value of the PCA approximation is less than or equal to that of the original problem. However, Theorem 2.4 does not quantify the gap between the optimal solution value and its PCA approximation. The following proposition presents bounds on this gap. The importance of our result is that the bound only depends on the input parameters, and thus it can guide the number of principal vectors for a specified error bound.

PROPOSITION 2.6. *When the support of ξ is polyhedral with at least one interior point, i.e., $\mathcal{S} = \{\xi | A\xi \leq b\}$ with $A \in \mathbb{R}^{n_1 \times m}$ and $b \in \mathbb{R}^{n_1}$, if $f(\boldsymbol{x}, \xi)$ is a piecewise linear convex function (precisely, $f(\boldsymbol{x}, \xi) = \max_{k=1}^{K} y_k^0(\boldsymbol{x}) + \boldsymbol{y}_k(\boldsymbol{x})^T \xi$, where $\boldsymbol{y}_k(\boldsymbol{x}) =$*

$[y_k^1(\boldsymbol{x}), \ldots, y_k^m(\boldsymbol{x})]$ as well as $y_k^0(\boldsymbol{x})$ are affine in $\boldsymbol{x}$ for $k = 1, \ldots, K$), then

$$0 \le Z^*(m) - Z^*(m_1) \le \sum_{k=1}^{K} \sqrt{\sum_{i=m_1+1}^{m} \Lambda_{i,i}[(A^T \boldsymbol{\lambda}_k^* - \boldsymbol{y}_k(\boldsymbol{x}^*))^T U_i]^2},$$

where $\boldsymbol{x}^*$ and $\boldsymbol{\lambda}_k^*, k = 1, \ldots, K$ are optimal solutions of the PCA approximation (2.15) *and $Z^*(\cdot)$ is defined as in* (2.14)*.*

Proof. Theorem 2.4 implies that $Z^*(m) - Z^*(m_1) \ge 0$. According to Corollary 2.5, we have

$$Z^*(m_1) = \underset{\boldsymbol{x}, s, \boldsymbol{q}_r, \boldsymbol{\lambda}, \boldsymbol{Q}_r}{\text{minimize}} \; s + \mathbf{I}_{m_1} \bullet \boldsymbol{Q}_r$$

(2.16)

subject to

$$\begin{bmatrix} s - y_k^0(\boldsymbol{x}) - \boldsymbol{\lambda}_k^T b - \boldsymbol{y}_k(\boldsymbol{x})^T \mu + \boldsymbol{\lambda}_k^T A \mu & \frac{(\boldsymbol{q}_r + (U_{m \times m_1} \Lambda_{m_1}^{1/2})^T (A^T \boldsymbol{\lambda}_k - \boldsymbol{y}_k(\boldsymbol{x})))^T}{2} \\ \frac{\boldsymbol{q}_r + (U_{m \times m_1} \Lambda_{m_1}^{1/2})^T (A^T \boldsymbol{\lambda}_k - \boldsymbol{y}_k(\boldsymbol{x}))}{2} & \boldsymbol{Q}_r \end{bmatrix} \succeq 0$$

$\forall k \in \{1, 2, \ldots, K\}$,

$$\boldsymbol{Q}_r \succeq 0, \; \boldsymbol{\lambda}_k \ge 0, \; k = 1, \ldots, K, \; \boldsymbol{x} \in X,$$

while

$$Z^*(m) = \underset{\boldsymbol{x}, s, \boldsymbol{q}, \boldsymbol{\lambda}, \boldsymbol{Q}}{\text{minimize}} \; s + \mathbf{I}_m \bullet \boldsymbol{Q}$$

subject to

$$\begin{bmatrix} s - y_k^0(\boldsymbol{x}) - \boldsymbol{\lambda}_k^T b - \boldsymbol{y}_k(\boldsymbol{x})^T \mu + \boldsymbol{\lambda}_k^T A \mu & \frac{(\boldsymbol{q} + (U \Lambda^{1/2})^T (A^T \boldsymbol{\lambda}_k - \boldsymbol{y}_k(\boldsymbol{x})))^T}{2} \\ \frac{\boldsymbol{q} + (U \Lambda^{1/2})^T (A^T \boldsymbol{\lambda}_k - \boldsymbol{y}_k(\boldsymbol{x}))}{2} & \boldsymbol{Q} \end{bmatrix} \succeq 0$$

(2.17) $\forall k \in \{1, 2, \ldots, K\}$,

$$\boldsymbol{Q} \succeq 0, \; \boldsymbol{\lambda}_k \ge 0, \; k = 1, \ldots, K, \; \boldsymbol{x} \in X,$$

Let $(\boldsymbol{x}^*, s^*, \boldsymbol{q}_r^*, \boldsymbol{\lambda}^*, Q_r^*)$ be the optimal solution of problem (2.16). For clarity, we define

$$\boldsymbol{q}_{m-m_1}^k := [U_{m \times m - m_1} (\bar{\Lambda}^{m-m_1})^{1/2}]^T (A^T \boldsymbol{\lambda}_k^* - \boldsymbol{y}_k(\boldsymbol{x}^*)),$$

where $\bar{\Lambda}^{m-m_1}$ is the $m - m_1 \times m - m_1$ lower-right submatrix of Λ. Then we set

$$\begin{aligned} \boldsymbol{q} &= [\boldsymbol{q}_r^*; \mathbf{0}_{m-m_1,1}], \\ s &= s^* + \sum_{k=1}^{K} \frac{\sqrt{(\boldsymbol{q}_{m-m_1}^k)^T \boldsymbol{q}_{m-m_1}^k}}{2} \\ &= s^* + \sum_{k=1}^{K} \frac{\sqrt{\sum_{i=m_1+1}^{m} \Lambda_{i,i}[(A^T \boldsymbol{\lambda}_k^* - \boldsymbol{y}_k(\boldsymbol{x}^*))^T U_i]^2}}{2}, \\ \boldsymbol{Q} &= \begin{bmatrix} \boldsymbol{Q}_r^* & \mathbf{0}_{m_1, m-m_1} \\ \mathbf{0}_{m-m_1, m_1} & \sum_{k=1}^{K} \frac{\boldsymbol{q}_{m-m_1}^k (\boldsymbol{q}_{m-m_1}^k)^T}{2\sqrt{(\boldsymbol{q}_{m-m_1}^k)^T \boldsymbol{q}_{m-m_1}^k}} \end{bmatrix}. \end{aligned}$$

Since

$$\begin{bmatrix} \frac{\sqrt{(\boldsymbol{q}^k_{m-m_1})^T\boldsymbol{q}^k_{m-m_1}}}{2} & \mathbf{0}_{1,m_1} & \frac{(\boldsymbol{q}^k_{m-m_1})^T}{2} \\ \mathbf{0}_{m_1,1} & \mathbf{0}_{m_1,m_1} & \mathbf{0}_{m_1,m-m_1} \\ \frac{\boldsymbol{q}^k_{m-m_1}}{2} & \mathbf{0}_{m-m_1,m_1} & \frac{\boldsymbol{q}^k_{m-m_1}(\boldsymbol{q}^k_{m-m_1})^T}{2\sqrt{(\boldsymbol{q}^k_{m-m_1})^T\boldsymbol{q}^k_{m-m_1}}} \end{bmatrix} \succeq 0,$$

$(\boldsymbol{x}^*, s, \boldsymbol{q}, \boldsymbol{\lambda}^*, \boldsymbol{Q})$ is a feasible solution of problem (2.17). Thus

$$\begin{aligned} Z^*(m) &\leq s + \mathbf{I}_m \bullet \boldsymbol{Q} \\ &= Z^*(m_1) + \sum_{k=1}^{K} \frac{\sqrt{(\boldsymbol{q}^k_{m-m_1})^T\boldsymbol{q}^k_{m-m_1}}}{2} + \operatorname{trace}\left(\sum_{k=1}^{K} \frac{\boldsymbol{q}^k_{m-m_1}(\boldsymbol{q}^k_{m-m_1})^T}{2\sqrt{(\boldsymbol{q}^k_{m-m_1})^T\boldsymbol{q}^k_{m-m_1}}}\right) \\ &= Z^*(m_1) + \sum_{k=1}^{K} \sqrt{\sum_{i=m_1+1}^{m} \Lambda_{i,i}[(A^T\boldsymbol{\lambda}^*_k - \boldsymbol{y}_k(\boldsymbol{x}^*))^T U_i]^2}. \end{aligned}$$

Therefore, we have

$$Z^*(m) - Z^*(m_1) \leq \sum_{k=1}^{K} \sqrt{\sum_{i=m_1+1}^{m} \Lambda_{i,i}[(A^T\boldsymbol{\lambda}^*_k - \boldsymbol{y}_k(\boldsymbol{x}^*))^T U_i]^2}. \qquad \square$$

In Proposition 2.6, the tightness of the bound is restricted (e.g., some elements of $\boldsymbol{q}$ are treated as 0). However, a tighter bound can be obtained as follows:

$$\begin{aligned} 0 \leq Z^*(m) - Z^*(m_1) \leq Z^1_{gap}(m_1) \leq Z^2_{gap}(m_1) \\ \leq \sum_{k=1}^{K} \sqrt{\sum_{i=m_1+1}^{m} \Lambda_{i,i}[(A^T\boldsymbol{\lambda}^*_k - \boldsymbol{y}_k(\boldsymbol{x}^*))^T U_i]^2}, \end{aligned}$$

where $Z^1{}_{gap}(m_1)$ is the optimal value of the optimization problem

$$\begin{aligned} Z^1_{gap}(m_1) = \underset{s,\boldsymbol{q},\bar{\boldsymbol{Q}}}{\text{minimize}} \quad & s + \mathbf{I}_{m-m_1} \bullet \bar{\boldsymbol{Q}} \\ \text{subject to} \quad & \begin{bmatrix} s & \frac{(\boldsymbol{q}^k_{m-m_1} - \boldsymbol{q})^T}{2} \\ \frac{\boldsymbol{q}^k_{m-m_1} - \boldsymbol{q}}{2} & \bar{\boldsymbol{Q}} \end{bmatrix} \succeq 0 \ \forall k \in \{1, 2, \ldots, K\}, \\ & \boldsymbol{q} \in \mathbb{R}^{m-m_1}, \ \bar{\boldsymbol{Q}} \in \mathbb{R}^{m-m_1 \times m-m_1}. \end{aligned} \tag{2.18}$$

Note that even if m_1 (number of components used) is small, problem (2.18) as an SDP problem is still computationally challenging to solve. Alternatively, a weaker bound $Z^2_{gap}(m_1)$ can be computed as

$$\begin{aligned} Z^2_{gap}(m_1) = \underset{s,\boldsymbol{q}}{\text{minimize}} \quad & \sum_{k=1}^{K} s + \frac{(\boldsymbol{q}^k_{m-m_1} - \boldsymbol{q})^T(\boldsymbol{q}^k_{m-m_1} - \boldsymbol{q})}{4s} \\ \text{subject to} \quad & s \geq 0, \ \boldsymbol{q} \in \mathbb{R}^{m-m_1}, \end{aligned} \tag{2.19}$$

which is a conservative approximation to (2.18) since

$$\bar{\boldsymbol{Q}} = \sum_{k=1}^{K} \frac{(\boldsymbol{q}^k_{m-m_1} - \boldsymbol{q})(\boldsymbol{q}^k_{m-m_1} - \boldsymbol{q})^T}{4s}.$$

Problem (2.19) can be formulated as a second-order conic programming problem and thus is significantly more tractable.

3. Low-rank approximations for chance constrained optimization. In section 2, it was assumed that the uncertainty lies in the objective function. We now turn our attention to the case in which the uncertainty lies in the feasible set X; specifically, the constraint function involves uncertainty as follows:

$$X = \Big\{ \boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{x} \in X_0, \quad \underset{F \in \mathcal{D}}{\text{minimize}}\, \mathbf{P}_F \{ h_l^0(\boldsymbol{x}) + \boldsymbol{h}_l(\boldsymbol{x})^T \xi < 0 \} \geq 1 - \epsilon_l \ \forall l = 1, \dots, L \Big\}, \tag{3.1}$$

where $X_0 \subset \mathbb{R}^n$ is a convex closed set that can be represented by semidefinite constraints, $\xi \in \mathbb{R}^m$ is a random vector with a distribution F, $\boldsymbol{h}_l(\boldsymbol{x}) = [h_l^1(\boldsymbol{x}), \dots, h_l^m(\boldsymbol{x})]$, and $h_l^0(\boldsymbol{x})$ are affine in $\boldsymbol{x}$ for $l = 1, \dots, L$. In addition, $\mathbf{P}$ is a probability measure on $\mathbb{R}^m$ induced by ξ. Constraint (3.1) is called an individual chance constraint, where ϵ_l, $l = 1, \dots, L$, are confidence parameters chosen by the decision maker, typically close to zero (e.g., 0.05, 0.10).

THEOREM 3.1. *When the support of ξ is polyhedral with at least one interior point, i.e., $\mathcal{S} = \{\xi | A\xi \leq b\}$ with $A \in \mathbb{R}^{n_1 \times m}$ and $b \in \mathbb{R}^{n_1}$, set X is equivalent to the following set:*

$$\bar{X} := \Bigg\{ \boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{x} \in X_0,\ \exists t_l \geq 0,\ s_l \in \mathbb{R},\ \boldsymbol{q}_l \in \mathbb{R}^m,\ 0 \preceq \boldsymbol{Q}_l \in \mathbb{R}^{m \times m},\ \boldsymbol{\lambda}_l, \bar{\boldsymbol{\lambda}}_l \in \mathbb{R}_+^{n_1}$$

$$\textit{subject to} \quad s_l + \mu^T \boldsymbol{q}_l + (\Sigma + \mu\mu^T) \bullet \boldsymbol{Q}_l \leq \epsilon_l t_l, \tag{3.2a}$$

$$\begin{bmatrix} s_l - h_l^0(\boldsymbol{x}) - t_l - b^T \bar{\boldsymbol{\lambda}}_l & \frac{(\boldsymbol{q}_l - \boldsymbol{h}_l(\boldsymbol{x}) + A^T \bar{\boldsymbol{\lambda}}_l)^T}{2} \\ \frac{\boldsymbol{q}_l - \boldsymbol{h}_l(\boldsymbol{x}) + A^T \bar{\boldsymbol{\lambda}}_l}{2} & \boldsymbol{Q}_l \end{bmatrix} \succeq 0, \tag{3.2b}$$

$$\begin{bmatrix} s_l - b^T \boldsymbol{\lambda}_l & \frac{(\boldsymbol{q}_l + A^T \boldsymbol{\lambda}_l)^T}{2} \\ \frac{\boldsymbol{q}_l + A^T \boldsymbol{\lambda}_l}{2} & \boldsymbol{Q}_l \end{bmatrix} \succeq 0, \ l = 1, \dots, L \Bigg\}. \tag{3.2c}$$

Constraints (3.2b) and (3.2c) are linear matrix inequalities. The PCA approximation can be applied to chance constraints in order to reduce the size of the matrix inequality constraints. Akin to section 2.3, we replace ξ by $U_{m \times m_1} \Lambda_{m_1}^{1/2} \xi_r + \mu$ in constraint (3.1), where the ambiguity set of ξ_r is still $\mathcal{D}_r$. Accordingly, the PCA approximation of set $\bar{X}$ is given as follows:

$$\bar{X}_r(m_1) := \Bigg\{ \boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{x} \in X_0,\ \exists t_l \geq 0,\ s_l \in \mathbb{R},\ \boldsymbol{q}_l \in \mathbb{R}^{m_1},\ 0 \preceq \boldsymbol{Q}_l \in \mathbb{R}^{m_1 \times m_1},\ \boldsymbol{\lambda}_l, \bar{\boldsymbol{\lambda}}_l \in \mathbb{R}_+^{n_1}$$

$$\text{subject to} \quad s_l + \mathbf{I}_{m_1} \bullet \boldsymbol{Q}_l \leq \epsilon_l t_l,$$

$$\begin{bmatrix} s_l - h_l^0(\boldsymbol{x}) - \boldsymbol{h}_l(\boldsymbol{x})^T \mu - t_l - (b - A\mu)^T \bar{\boldsymbol{\lambda}}_l & \frac{(\boldsymbol{q}_l - (U_{m \times m_1} \Lambda_{m_1}^{1/2})^T \boldsymbol{h}_l(\boldsymbol{x}) + (AU_{m \times m_1} \Lambda_{m_1}^{1/2})^T \bar{\boldsymbol{\lambda}}_l)^T}{2} \\ \frac{(\boldsymbol{q}_l - (U_{m \times m_1} \Lambda_{m_1}^{1/2})^T \boldsymbol{h}_l(\boldsymbol{x}) + (AU_{m \times m_1} \Lambda_{m_1}^{1/2})^T \bar{\boldsymbol{\lambda}}_l)}{2} & \boldsymbol{Q}_l \end{bmatrix}$$

$$\succeq 0,$$

$$\begin{bmatrix} s_l - (b - A\mu)^T \boldsymbol{\lambda}_l & \frac{(\boldsymbol{q}_l + (AU_{m \times m_1} \Lambda_{m_1}^{1/2})^T \boldsymbol{\lambda}_l)^T}{2} \\ \frac{(\boldsymbol{q}_l + (AU_{m \times m_1} \Lambda_{m_1}^{1/2})^T \boldsymbol{\lambda}_l)}{2} & \boldsymbol{Q}_l \end{bmatrix} \succeq 0, \ l = 1, \dots, L \Bigg\}.$$

PROPOSITION 3.2. *The feasible set $\bar{X}$ is a subset of $\bar{X}_r(m_1)$, i.e., $\bar{X} \subset \bar{X}_r(m_1)$. At the same time, $\bar{X}_r(m_2) \subset \bar{X}_r(m_1)$ if $m_1 \leq m_2$. Furthermore, if $m_1 = m$, $\bar{X}_r(m_1) = \bar{X}$.*

Proof. The idea of proof is similar to that of Theorem 2.4. □

These results show that the proposed PCA approach can be extended to solving distributionally robust chance constrained problems. We now present comprehensive computational results to prove the efficacy of our approach in terms of runtime and solution quality.

4. Experimental results.

In this section, we compare the performances of the proposed PCA approximation with that of the original formulation on a distributionally robust conditional value-at-risk (CVaR) application and a risk-averse production-transportation application. All algorithms are implemented in MATLAB using the modeling language CVX [18, 19] and the corresponding SDP instances are solved using Mosek with default parameters on a machine with an Intel Core i7 2.8 GHz processor and 16GB RAM.

4.1. Distributionally robust CVaR.

We consider a distributionally robust version of CVaR problems. CVaR, as an approximation of value-at-risk (VaR), has been extensively studied due to desirable properties like subadditivity and convexity [34]. Additionally, in chance constrained programming, the CVaR approximation is the least conservative convex approximation of the chance constraints [30]. For more details on CVaR, we refer the reader to Rockafellar and Uryasev [34].

Rockafellar and Uryasev [34] proved that the $\text{CVaR}_{1-\alpha}$ of a cost function $g(\boldsymbol{x}, \xi)$ can be formulated as the following optimization problem:

$$\underset{t \in \mathbb{R}}{\text{minimize}} \; t + \frac{1}{\alpha} \mathbf{E}_F [g(\boldsymbol{x}, \xi) - t]^+ \tag{4.1}$$

where $\alpha \in (0,1)$ is a risk tolerance level, F is the probability distribution of ξ, and $[\cdot]^+ := \max\{0, \cdot\}$. When the exact distribution F is not available but information about the distribution family $\mathcal{D}$ is available, we can consider the following distributionally robust variant of the CVaR problem:

$$\underset{\boldsymbol{x} \in X}{\text{minimize}} \; \underset{F \in \mathcal{D}}{\text{maximize}} \; \underset{t \in \mathbb{R}}{\text{minimize}} \; t + \frac{1}{\alpha} \mathbf{E}_F [g(\boldsymbol{x}, \xi) - t]^+. \tag{4.2}$$

In what follows, we assume that $g(\boldsymbol{x}, \xi) = \boldsymbol{x}^T \xi$ and $X = \{\boldsymbol{x} \in \mathbb{R}^n_+ | \sum_{i=1}^n x_i = 1\}$ and that the distribution family $\mathcal{D}$ satisfies Assumption 1. In this case, problem (4.2) is equivalent to the following mini-max problem:

$$\underset{\boldsymbol{x} \in X, t \in \mathbb{R}}{\text{minimize}} \; \underset{F \in \mathcal{D}}{\text{maximize}} \; t + \frac{1}{\alpha} \mathbf{E}_F [\boldsymbol{x}^T \xi - t]^+. \tag{4.3}$$

The equivalence of (4.2) and (4.3) follows directly from the application of the minimax theorem [10] to $\text{maximize}_{F \in \mathcal{D}} \, \text{minimize}_{t \in \mathbb{R}}$. Observe that $t + \frac{1}{\alpha}[\boldsymbol{x}^T \xi - t]^+ := \max\{t, t + \frac{1}{\alpha}(\boldsymbol{x}^T \xi - t)\}$ satisfies Corollary 2.2. Thus, we can solve the CVaR problem exactly using the results of Corollary 2.2 or approximately using the proposed PCA approximation scheme.

Table 4.1

Performance of the PCA approximation on a 200*-dimensional CVaR application for varying ambiguity set support and number of principal components.*

CVAR $m = 200$ Support	Orig. time (s)	PCA ($m_1 = 200$)			PCA ($m_1 = 150$)			PCA ($m_1 = 100$)			PCA ($m_1 = 50$)			PCA ($m_1 = 20$)		
		time (s)	Gap (%)	Gap2 (%)	time (s)	Gap (%)	Gap2 (%)	time (s)	Gap (%)	Gap2 (%)	time (s)	Gap (%)	Gap2 (%)	time (s)	Gap (%)	Gap2 (%)
$[-2\sigma, 2\sigma]$	1019.5	654.5	0.00	0.00	219.4	0.26	8.37	41.1	1.55	9.10	3.1	3.57	12.93	2.0	5.24	18.45
$[-3\sigma, 3\sigma]$	1290.9	1078.3	0.00	0.00	334.2	2.46	7.40	40.8	4.20	9.93	2.7	6.45	14.85	1.1	8.49	19.50
$[-4\sigma, 4\sigma]$	1309.2	1362.0	0.00	0.00	324.1	3.06	7.42	42.9	5.49	10.19	3.1	8.37	14.18	1.7	10.56	19.13

4.1.1. Experimental setup: Approximation quality vs. number of principal components.

In this section, we focus on the effects of the number of principal components of the PCA approximation on solution quality and runtime. Computational results are presented for both randomly generated instances and instances based on historical financial market data. For randomly generated instances, we set $m = n = 200$ and $\alpha = 0.05$. The mean μ is picked uniformly at random from the interval $[-5, 5]$, the standard deviation of ξ is picked uniformly from the interval $[0, 2]$, and the correlation matrix is generated randomly using the MATLAB function "`gallery('randcorr',n)`." We consider three different supports for the ambiguity set, specifically $\mathcal{S} \in \{[-2\sigma, 2\sigma], [-3\sigma, 3\sigma], [-4\sigma, 4\sigma]\}$, where σ is the randomly generated standard deviation of ξ. For the PCA approximations, the number of principal components is $m_1 \in \{200, 150, 100, 50, 20\}$, which correspond to 100%, 75%, 50%, 25%, and 10% of the size of ξ, respectively. In the second part of the computational experiments, all means and covariances are estimated using historical market data (obtained from Yahoo Finance) and the tests are conducted using different values for α and different numbers of principal components.

4.1.2. Results on randomly generated covariance matrices.

To randomly generate Σ, we apply the MATLAB function "`gallery('randcorr',n)`," which generates a full-rank matrix. For each different support, 10 instances are generated and solved. The results are presented in Figures 4.1–4.3 (see color figures in the online version) and the average statistics over 10 instances are summarized in Table 4.1. Figures 4.1–4.3 show the runtime and the relative gap of the optimal value using different numbers of principal components for 2σ, 3σ, and 4σ supports. We define the relative gap of the optimal value between the PCA approximation and the original reformulation as $|(Z^*(m) - Z^*(m_1))/Z^*(m)| \times 100\%$, where $Z^*(m)$ and $Z^*(m_1)$ are the optimal values of the original reformulation and the PCA approximation, respectively.

In Figures 4.1–4.3, we display statistics for the relative gaps (GAP) over 10 randomly generated instances for three different supports $\mathcal{S}$ (2σ, 3σ, and 4σ). For each boxplot, the minimum, 9th percentile, median, 91st percentile, and maximum are given in order (bottom to top). We also show the average runtime over the 10 instances. In Table 4.1, column 1 shows the support of the ambiguity set and column 2 shows the runtime for solving the original reformulation. Each set of three columns that follows shows computational performance for the PCA approximation with varying number of principal components ($m_1 = 200,\ 150,\ 100,\ 50$, and 20). The metrics presented are runtime (in seconds), relative gap (Gap), and theoretical relative gap (Gap2) derived from Proposition 2.6.

Results in Figures 4.1–4.3 and Table 4.1 first show that as the number of principal components m_1 increases the numerical and theoretical gaps decrease and the runtime

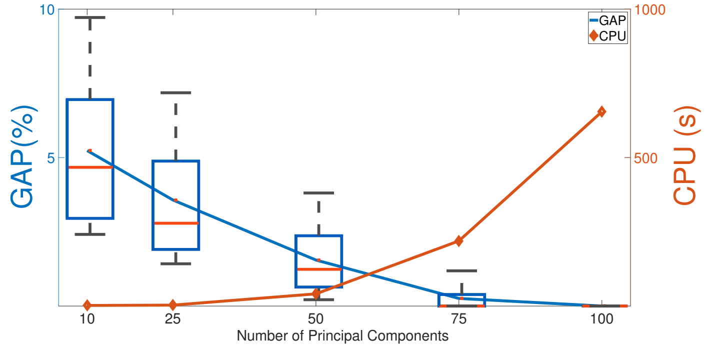

FIG. 4.1. *Performance of the PCA approximation when the support is* $[-2\sigma, 2\sigma]$ *with* 20, 50, 100, 150, 200 *components. The red line shows the runtime in seconds, while the blue line shows the optimality gap (%). Each boxplot displays the minimum, the* 9*th percentile, the median, the* 91*st percentile, and the maximum.*

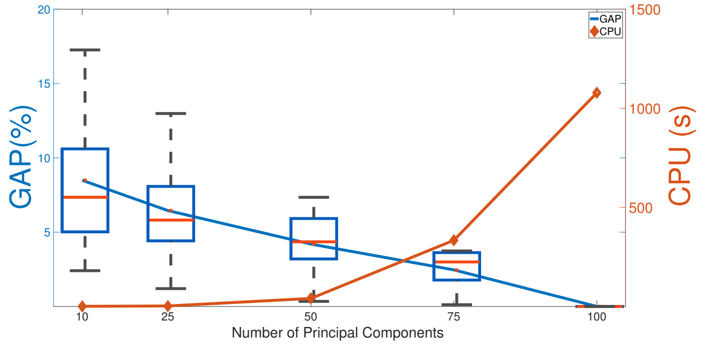

FIG. 4.2. *Performance of the PCA approximation when the support is* $[-3\sigma, 3\sigma]$ *with* 20, 50, 100, 150, 200 *components.*

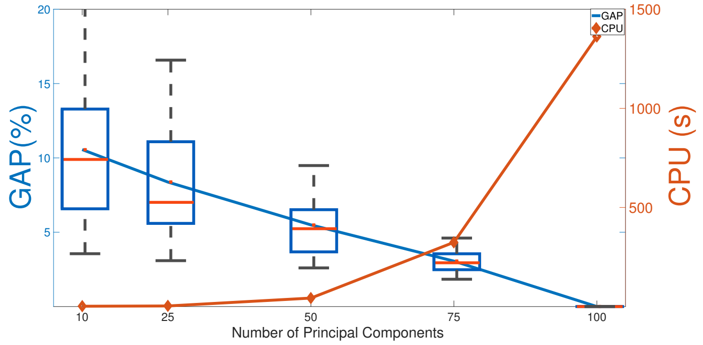

FIG. 4.3. *Performance of the PCA approximation when the support is* $[-4\sigma, 4\sigma]$ *with* 20, 50, 100, 150, 200 *components.*

Table 4.2

Performance of the PCA approximation on a 300*-dimensional problem for varying number of principal components.*

Size	Orig. time (h)	PCA ($m_1 = 300$)		PCA ($m_1 = 225$)		PCA ($m_1 = 150$)		PCA ($m_1 = 75$)		PCA ($m_1 = 30$)	
		time (h)	Gap (%)	time (h)	Gap (%)	time (h)	Gap (%)	time (h)	Gap (%)	time (h)	Gap (%)
$m = 300$	9.416	8.605	0.00	0.867	1.56	0.088	3.71	0.004	5.89	0.000	7.55

increases. Not surprisingly, the theoretical gap obtained via Proposition 2.6 is larger than the observed gap. The difference between Gap2 and Gap suggests that the theoretical gap may potentially be improved. Secondly, when we use all the principal components (i.e., $m_1 = m$), the PCA approximation yields an exact reformulation of the original problem, which is consistent with Theorem 2.4. Moreover, the PCA approximations with only a subset of the components provide lower bounds to the original problem. Finally, it is clear that the proposed PCA approximation allows for practical trade-offs between solution quality and computation time, higher solution quality comes at the cost of increased runtime. For instance, with $n = 200$ and support $[-2\sigma, 2\sigma]$, solving the original (exact) reformulation requires 1019 seconds; however, a high-quality solution within a 1.55% optimality gap can be computed within 42 seconds.

In addition, we also consider a larger CVaR problem with $m = 300$ and supports $\mathcal{S} = [-3\sigma, 3\sigma]$. For the PCA approximations, the number of the principal components is $m_1 \in \{300, 225, 150, 70, 30\}$, which correspond to 100%, 75%, 50%, 25%, and 10% of the size of ξ, respectively. Similarly, 10 instances are generated and solved, and the average performance is shown in Table 4.2. The first column shows the value of m and the second column shows the runtime for solving the original reformulation. Each pair of columns that follows shows computational performance, runtime (in hours), and relative gap (Gap) for the PCA approximation with varying number of principal components.

From Table 4.2, we can draw similar conclusions for the PCA approximations. Moreover, the runtime reductions of the PCA approximations for the larger-size problems are even more substantial. For instance, it took more than nine hours to solve the original reformulation. This is in sharp contrast to the runtime of the PCA approximation with 75% of the principal components, which completed in under one hour and with only a 1.56% optimality gap.

4.1.3. Specially structured covariance matrices. In this section, we consider the specially structured covariance matrix Σ. All the parameters are randomly generated using the same procedure as the previous experiment. For conciseness, we only consider one support for the ambiguity set with $m = 200$, specifically $\mathcal{S} = [-3\sigma, 3\sigma]$. With the randomly generated covariance matrix Σ, we replace its ith largest eigenvalue by three different kinds of generating functions: the first one is constant, i.e., 1, for $i = 1, \ldots, m$; the second one is linear, i.e., $1 - 0.5\frac{i-1}{m-1}$, $i = 1, \ldots, m$; the third one is exponential, i.e., $1 - (e^{\frac{-i+m+1}{m}\gamma} - 1)/(e^{\gamma} - 1)$, $i = 1, \ldots, m$, with slope γ. Here we consider four different slopes, $\gamma \in \{0.1, 1, 5, 15\}$. We display the six generating-eigenvalue functions in Figure 4.4 (see color figures in the online version). Similar to the previous experiment, 10 test instances are solved for each generating

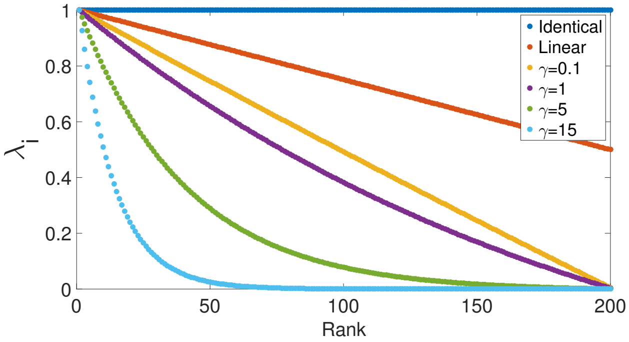

FIG. 4.4. *Eigenvalue generating functions used in the experiments.*

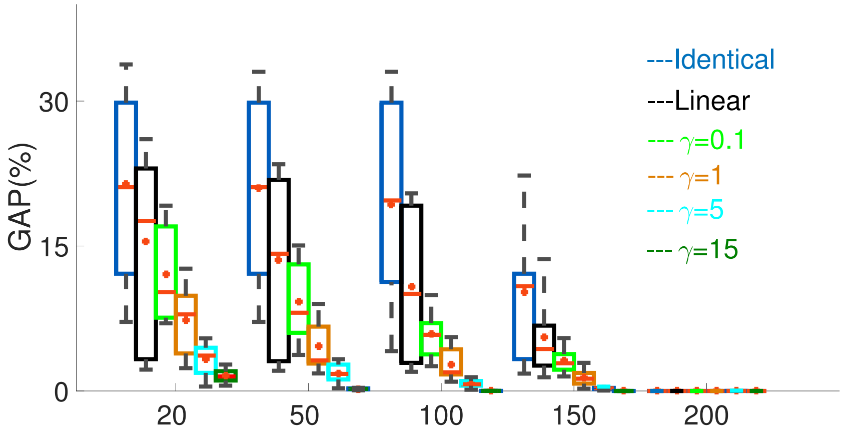

FIG. 4.5. *Performance of the PCA approximation with* 20, 50, 100, 150, 200 *components for six different eigenvalue generating functions.*

function. The main results are presented in boxplot form in Figure 4.5 and the average statistics are listed in Table 4.3.

In Figure 4.4, we consider three cases for eigenvalue decay: no decay, linear decay, and exponential decay. Note that for the exponential case, when γ increases the eigenvalues decay rapidly. From Figure 4.5 and Table 4.3, we can deduce similar conclusions about the solution quality and runtime improvements for instances with randomly generated Σs. First, the Gap and the runtime are inversely proportional, and as the Gap decreases computation time increases. Second, the faster the eigenvalues decay, the smaller the relative gap between the PCA approximation and the original reformulation. For instance, when the number of principal components is 100, the gap decreases from 19.24% for the constant eigenvalues case to 10.82% for the case of linearly decaying eigenvalues. Further, in the case of exponentially decreasing eigenvalues, the gap decreases from 5.88% to 0.01% when the parameter γ increases from 0.1 to 15. In addition, when γ is 15, only 50% of the principal components are needed to obtain a high-quality, near-optimal solution (Gap is 0.01% and runtime

Table 4.3

Performance of the PCA approximation on a 200*-dimensional problem for varying number of principal components and decay parameters for the eigenvalues. Metrics presented are runtime (in seconds) and relative gap % (Gap).*

Slope	Orig. time (s)	PCA ($m_1 = 200$)		PCA ($m_1 = 150$)		PCA ($m_1 = 100$)		PCA ($m_1 = 50$)		PCA ($m_1 = 20$)	
		time (s)	Gap (%)	time (s)	Gap (%)	time (s)	Gap (%)	time (s)	Gap (%)	time (s)	Gap (%)
Identical	1234.2	1036.9	0.00	148.8	10.24	21.6	19.29	2.2	21.02	2.0	21.40
Linear	1344.8	1326.5	0.00	296.8	5.58	41.7	10.82	3.0	13.58	2.0	15.47
0.1	1401.0	1561.2	0.00	337.9	3.12	42.4	5.88	3.1	9.24	2.0	12.06
1	1643.7	1800.1	0.00	340.0	1.38	51.1	2.70	2.7	4.62	1.0	7.31
5	1731.4	1560.0	0.00	346.5	0.26	45.4	0.75	2.8	1.83	1.0	3.26
15	1503.1	1624.7	0.00	325.2	0.00	42.3	0.01	2.6	0.21	1.1	1.59

is 42.3 seconds). This computation time is drastically less than the 1503.1 seconds runtime required for solving the original reformulation. These results suggest that the PCA approximation performs well with fewer principal components and can be used to obtain high-quality, near-optimal solutions when the eigenvalues of the Σs drop rapidly.

In the next set of experiments, we will showcase how our techniques can be adopted to solve problems that are too large to be solved by existing methods. For these experiments, we consider a larger CVaR problem with specially structured Σ when $m = 1000$ in the case of exponential generating functions with different slopes. For the PCA approximations, the number of the principal components is $m_1 \in \{200, 150, 100, 50\}$, which correspond to 20%, 15%, 10%, and 5% of the size of ξ, respectively. Similarly, 10 instances are generated and solved and the average performance is shown in Table 4.4. When $m = 1000$, the original reformulation problem and the PCA approximation problem using even 50% of principal components are too large to solve due to memory limits. Thus, we present an upper bound on the relative gap of the optimal value between the PCA approximation and the original reformulation as $|Z^2_{gap}(m_1)/Z^*_{lb}| \times 100$, where $Z^2_{gap}(m_1)$ is a theoretical gap derived from (2.19) and Z^*_{lb} is a lower bound of the absolute value of the optimal value of the original reformulation. Here $Z^*_{lb} = \min\{|Z^*(m_1) + Z^2_{gap}(m_1)|, |Z^*(m_1)|\}$, where $Z^*(m_1)$ is the optimal value of the PCA approximation with $m_1 = 200$.

From Table 4.4, we can draw similar conclusions for the PCA approximations in the case in which $m = 200$. Moreover, the runtime reductions obtained by using PCA approximations for the larger-size problems are even more substantial. For example, in this setup the original reformulation is intractable due to memory limitations. However, the PCA approximation with 20% ($m_1 = 200$) of the principal components solved in less than half an hour and obtained a solution within a 1.91% optimality gap when $\gamma = 15$. We want to stress that gaps presented here are theoretical gaps that provide upper bounds, and actual gaps are tighter. Nevertheless, the results show that our methods can be used to compute provably accurate approximations to problems that are too large to be solved exactly in practice.

Table 4.4
Performance of the PCA approximation on a 1000*-dimensional problem for varying number of principal components and decay parameters for the eigenvalues. "–" indicates that no solution was found and "∗" indicates an upper bound for the relative gap rather than the actual gap.*

Slope	Orig. time (s)	PCA ($m_1 = 200$)		PCA ($m_1 = 150$)		PCA ($m_1 = 100$)		PCA ($m_1 = 50$)	
		time (s)	Gap (%)	time (s)	Gap (%)	time (s)	Gap (%)	time (s)	Gap (%)
0.1	–	1947.5	17.25*	427.4	18.81*	61.8	20.01*	4.6	26.14*
1	–	1781.1	16.50*	408.5	17.22*	50.7	19.42*	4.7	25.02*
5	–	1817.9	8.72*	381.0	9.92*	54.2	11.44*	4.3	14.33*
15	–	1661.1	1.91*	421.6	2.61*	48.5	4.35*	4.3	6.62*

Table 4.5
Tickers of 123 *Assets in the Yahoo financial market data.*

Energy	CMLP CLMT	CELP NGL	NTI GLOP	KNOP GLP	DLNG TLP	USAC WPT	MMLP TCP	GMLP DKL	EXLP	DPM
Basic Materials	UAN	ARLP	AHGP	NRP	HCLP	OCIR	TNH	VALE	BBL	RIO
Consumer Staples	CHSCO	CHSCP	CHSCN	TIS	GTY	PG	SON	KMB	UL	CLX
Consumer Discretionary	CLCT MO	MAT CRWS	MHG RMCF	FUN NTRI	PM	BGS	EBF	BTI	GM	TUP
Financials	RSO OSBHF EPR	TICC GAIN	NYMT STWD	PMT SLRC	PSEC IEP	HTS HPT	NLY MAIN	CMO BX	MFA CLNY	NEWT OTCM
Healthcare	PDLI ARE	SNH LLY	SBRAP	SBRA	HCN	LTC	OHI	AZN	PETS	PMD
Industrials	TAL	SSW	TGH	FLY	AIRI	CVA	CTT	PLOW	AYR	GE
Utilities	APU WGL	CPL	EGAS	NGG	BIP	ED	WR	HE	PEG	EXC
Technology	MNDO	WILN	CCUR	AREEP	IRM	DFT	DLR	EVOL	GRMN	CPSI
Telecoms	CTL	PHI	TLSYY	T	BCE	VOD	VZ			

4.1.4. Covariance matrices based on financial market data. In this section, we evaluate the proposed PCA approximation on CVaR problems using historical market data in 2014/2015 for 123 assets (obtained from Yahoo Finance). The top assets of 10 industry sectors [21] (see Table 4.5) are considered for the portfolio. The mean and variance of 123 returns are estimated using 2014/2015 historical data, and historical ranges are used as support. We perform the numerical tests using different percentages of principal components (100% to 10%) and several values for the confidence parameter α (0.01 to 0.1). The corresponding results are reported in Table 4.6. The first column shows the percentages of principal components used. Columns 2 to 4 (and corresponding sets of subcolumns) show the optimal value, the relative gap, and the runtime, respectively, for $\alpha \in \{0.01, 0.05, 0.10\}$.

From Table 4.6, similar conclusions to randomly generated Σ and structured Σ computational experiments can be drawn. First, as the number of components increases the relative gap decreases and runtime increases. Second, the runtime for solving the PCA approximations with less than full components is significantly less than the runtime for solving the original reformulation, albeit at the expense of suboptimality. However, with $\alpha = 0.01$ the PCA approximation with 50% of components

TABLE 4.6

Performance of the PCA approximations for Yahoo financial market data. The first column shows the percentages of principal components used. Columns 2 *to* 4 *(and corresponding sets of subcolumns) show the optimal value, the relative gap, and the runtime, respectively, for the confidence parameter* $\alpha \in \{0.01, 0.05, 0.10\}$.

	$\alpha = 0.01$			$\alpha = 0.05$			$\alpha = 0.1$		
	Value (%)	Gap (%)	time (s)	Value (%)	Gap (%)	time (s)	Value (%)	Gap (%)	time (s)
Original	2.21	–	98.6	1.45	–	160.2	0.99	–	153.4
100%	2.21	0.00	98.8	1.45	0.00	162.2	0.99	0.00	155.9
95%	2.21	0.00	82.0	1.44	0.69	135.3	0.98	1.01	125.5
90%	2.21	0.00	65.7	1.43	1.38	113.5	0.98	1.01	103.5
80%	2.21	0.00	38.9	1.39	4.14	50.5	0.95	4.04	62.8
70%	2.21	0.00	21.5	1.36	6.22	34.8	0.93	6.06	29.5
50%	2.21	0.00	5.8	1.17	19.31	8.1	0.80	19.19	8.1
25%	1.40	36.65	1.2	0.83	42.76	1.3	0.57	42.42	1.3
10%	0.89	59.72	1.2	0.70	51.72	1.2	0.48	51.51	1.2

finds the optimal solution (optimality gap 0.00%) within less than one tenth of the time it took to solve the original reformulation. Results again suggest that the PCA approximation yields good solution quality and runtime trade-offs under a range of conditions.

4.2. Risk-averse production-transportation problem. In this section, we illustrate the use of the PCA approximation for the solution of DRO applied to a risk-averse production-transportation problem with random transportation cost (see Bertsimas et al. [6]). In the production-transportation problem, there are m facilities (supply points) and n customer locations (demand points). It is assumed that there is a normalized capacity for each facility. Let $x_i \geq 0$ be the amount that is produced at each facility i and let y_{ij} be the amount that is shipped from facility i to customer location j. We denote the unit production cost at facility i by c_i, the demand at customer location j by d_j, and the unit transportation cost from facility i to customer location j by ξ_{ij}. We further assume that demand d_j is known in advance and $\sum_j d_j \leq m$, which means the total demand does not exceed the total production capacity. If ξ_{ij} is deterministic, then the deterministic production-transportation problem can be formulated as follows:

$$\begin{aligned} \underset{\boldsymbol{x},\boldsymbol{y}}{\text{minimize}} \quad & \sum_{i=1}^{m} c_i x_i + \sum_{i=1}^{m} \sum_{j=1}^{n} \xi_{ij} y_{ij} \\ \text{subject to} \quad & \sum_{i=1}^{m} y_{ij} = d_j, \ j = 1, \ldots, n, \\ & \sum_{j=1}^{n} y_{ij} = x_i, \ i = 1, \ldots, m, \\ & 0 \leq x_i \leq 1, \ i = 1, \ldots, m, \\ & y_{ij} \geq 0, \ i = 1, \ldots, m, \ j = 1, \ldots, n. \end{aligned} \tag{4.4}$$

When the transportation cost ξ_{ij} is random, we have a two-stage version of the problem where production decision x_i should be made now whereas transportation decision

y_{ij} will be made after the realization of the random transportation cost ξ_{ij}. Under these assumptions, the distributionally robust variant of risk-averse production-transportation problem can be formulated as follows:

$$\begin{aligned} &\underset{\boldsymbol{x}}{\text{minimize}} && \sum_{i=1}^{m} c_i x_i + \underset{F\in\mathcal{D}}{\text{maximize}}\, \mathbf{E}_F[\mathcal{U}(\mathcal{Q}(\boldsymbol{x},\xi))] \\ &\text{subject to} && 0 \le x_i \le 1,\ i = 1,\dots,m, \end{aligned} \tag{4.5}$$

where $\mathcal{Q}(\boldsymbol{x},\xi)$ is the optimal value of the second-stage problem, defined as

$$\begin{aligned} &\underset{\boldsymbol{y}}{\text{minimize}} && \sum_{i=1}^{m}\sum_{j=1}^{n} \xi_{ij} y_{ij} \\ &\text{subject to} && \sum_{i=1}^{m} y_{ij} = d_j,\ j = 1,\dots,n, \\ &&& \sum_{j=1}^{n} y_{ij} = x_i,\ i = 1,\dots,m, \\ &&& y_{ij} \ge 0,\ i = 1,\dots,m,\ j = 1,\dots,n, \end{aligned}$$

and the function $\mathcal{U}(\cdot)$ is a convex nondecreasing disutility function that captures risk aversion with respect to the total achieved cost. The definition of disutility function $\mathcal{U}(\cdot)$ follows Bertsimas et al. [6] and is given as

$$\mathcal{U}(\mathcal{Q}(\boldsymbol{x},\xi)) = \max_{k\in\{1,2,\dots,K\}} a_k \mathcal{Q}(\boldsymbol{x},\xi) + b_k, \tag{4.6}$$

with nonnegative coefficients, i.e., $a_k \ge 0$ for all k.

For clarity, we also assume that the support of the ambiguity set satisfies Assumption 1 with $\mathcal{S} = \mathbb{R}^{nm}$. Then the original formulation of the risk-averse production-transportation problem is as follows (see [6]):

$$\begin{aligned} &\underset{\boldsymbol{x},\boldsymbol{y},s,\boldsymbol{q},\boldsymbol{Q}}{\text{minimize}} && c^T\boldsymbol{x} + s + \mu^T\boldsymbol{q} + (\Sigma + \mu\mu^T)\bullet\boldsymbol{Q} \\ &\text{subject to} && \begin{bmatrix} s - b_k & \frac{(\boldsymbol{q}-a_k\boldsymbol{y}_k)^T}{2} \\ \frac{(\boldsymbol{q}-a_k\boldsymbol{y}_k)}{2} & \boldsymbol{Q} \end{bmatrix} \succeq 0,\ k = 1,\dots,K, \\ &&& \sum_{i=1}^{m} y_{ijk} = d_j,\ j = 1,\dots,n,\ k = 1,\dots,K, \\ &&& \sum_{j=1}^{n} y_{ijk} = x_i,\ i = 1,\dots,m,\ k = 1,\dots,K, \\ &&& 0 \le x_i \le 1,\ i = 1,\dots,m, \\ &&& y_{ijk} \ge 0,\ i = 1,\dots,m,\ j = 1,\dots,n,\ k = 1,\dots,K, \end{aligned} \tag{4.7}$$

where $s \in \mathbb{R}$, $\boldsymbol{q} \in \mathbb{R}^{mn}$, $\boldsymbol{Q} \in \mathbb{R}^{mn\times mn}$, $c = [c_1;\dots;c_m]$, and $\boldsymbol{y}_k \in \mathbb{R}^{mn}$ is a vector whose $(i-1)*m+j$th element is y_{ijk}.

The PCA approximation with m_1 principal components for the risk-averse production-transportation problem is given as

$$
\begin{aligned}
&\underset{\boldsymbol{x},\boldsymbol{y},s,\boldsymbol{q}_r,\boldsymbol{Q}_r}{\text{minimize}} \quad c^T\boldsymbol{x} + s + \mathbf{I}_{m_1} \bullet \boldsymbol{Q}_r \\
&\text{subject to} \quad \begin{bmatrix} s - b_k - a_k\mu^T\boldsymbol{y}_k & \frac{(\boldsymbol{q}_r - a_k(U_{mn\times m_1}\Lambda_{m_1}^{1/2})^T\boldsymbol{y}_k)^T}{2} \\ \frac{(\boldsymbol{q}_r - a_k(U_{mn\times m_1}\Lambda_{m_1}^{1/2})^T\boldsymbol{y}_k)}{2} & \boldsymbol{Q}_r \end{bmatrix} \succeq 0 \\
&\qquad \forall k \in \{1,\ldots,K\}, \\
&\qquad \sum_{i=1}^m y_{ijk} = d_j,\ j = 1,\ldots,n,\ k = 1,\ldots,K, \\
&\qquad \sum_{j=1}^n y_{ijk} = x_i,\ i = 1,\ldots,m,\ k = 1,\ldots,K, \\
&\qquad 0 \leq x_i \leq 1,\ i = 1,\ldots,m, \\
&\qquad y_{ijk} \geq 0,\ i = 1,\ldots,m,\ j = 1,\ldots,n,\ k = 1,\ldots,K,
\end{aligned}
\tag{4.8}
$$

where $s \in \mathbb{R}$, $\boldsymbol{q}_r \in \mathbb{R}^{m_1}$, and $\boldsymbol{Q}_r \in \mathbb{R}^{m_1\times m_1}$.

4.2.1. Effect of the number of principal components. In this section, we conduct numerical experiments on the risk-averse production-transportation problem to demonstrate the efficacy of the PCA approximation. Following the setup in Bertsimas et al. [6], we randomly generate m facilities and n customer locations within a unit square. Let $\bar{\xi}_{ij}$ be the distance between facility i and customer location j for the randomly generated transportation network. The mean and covariance matrix of ξ are set to be the sample mean and sample covariance of 10,000 independent samples ξ_t generated from independent uniform distributions on intervals $[0.5\bar{\xi}_{ij}, 1.5\bar{\xi}_{ij}]$ for all i, j. The production cost c_i is uniformly generated on the interval $[0.5\bar{c}, 1.5\bar{c}]$, where $\bar{c}$ is the average transportation cost. The demand h_j is uniformly generated on the interval $[0.5\frac{m}{n}, \frac{m}{n}]$. We perform our tests on three different problem sizes characterized by the following parameters: number of facilities $m \in \{5, 8, 10\}$ and number of customer locations $n \in \{20, 25, 30\}$. We also consider the same disutility function $\mathcal{U}(x) = \gamma(e^{\delta x} - 1)$, where $\gamma, \delta > 0$. In the following numerical tests, γ and δ are set to be 0.25 and 2, respectively. We approximate the convex disutility function $\mathcal{U}(x)$ by using an equidistant linear approximation with $K = 5$ on the interval $[0, 1]$.

For each different size of the problem, 10 instances are generated and solved and the average statistics over the 10 instances are summarized in Table 4.7, where the first column lists the size of the problem and the second column shows the runtime (in seconds) for solving the original formulation. Each pair of columns that follow show computational performance for the PCA approximation with varying percentages of number of principal components (100%, 75%, 50%, 25%, and 10%). The metrics presented are runtime (in seconds) and relative gap (Gap). When $(m, n) = (10, 30)$, the original reformulation problem and the PCA approximation problem using all principal components are too large to solve for the computer due to memory limits. Thus, for this problem size, we present an upper bound of the relative gap of the optimal value between the PCA approximation and the original reformulation as $|Z^1_{gap}(m_1)/Z^*(m_1)| \times 100$, where $Z^1_{gap}(m_1)$ is a theoretical gap derived from (2.18) and $Z^*(m_1)$ is the optimal value of the PCA approximation.

Table 4.7

Performance of the PCA approximation for production-transportation problems of various sizes. m and n denote the numbers of facilities and customer locations, respectively. Time is in seconds and Gap measures the difference between the optimal solution and its PCA approximation. "–" indicates that no solution was found and "$*$" indicates an upper bound (computed based on (2.18)*) for the relative gap rather than the actual gap.*

(m, n)	Orig. time (s)	PCA (100%)		PCA (75%)		PCA (50%)		PCA (25%)		PCA (10%)	
		time (s)	Gap (%)	time (s)	Gap (%)	time (s)	Gap (%)	time (s)	Gap (%)	time (s)	Gap (%)
(5, 20)	91.4	88.2	0.00	27.4	0.25	7.7	0.57	2.2	0.93	1.7	0.94
(8, 25)	2574.5	2392.1	0.00	609.6	0.06	99.0	0.11	9.2	0.12	2.5	0.12
(10, 30)	–	–	–	4888.2	1.07*	705.2	1.44*	42.7	1.76*	5.3	2.35*

Results presented in Table 4.7 indicate that using a small number of principal components can produce high quality solutions. First, as the number of components increases, the relative gap decreases and runtime increases. Second, the PCA approximation using all principal components yields the same optimal solution as the original reformulation with similar runtimes. Overall the results show the efficacy of the PCA approximation for this class of problem. The solution quality is extremely high even when only 10% of the principal components are used. For example, the PCA approximation using only 10% of components provides a high-quality, near-optimal solution with a gap less than 2.35% (upper bound) in less than 6 seconds, a runtime speedup of several orders of magnitude.

5. Conclusions and future work.

We have proposed a computationally efficient approximation for distributionally robust optimization (DRO) problems with moment-based ambiguity sets. Previous results provide a way to reformulate DROs with moment-based ambiguity sets as equivalent semidefinite programs (SDPs), which can be solved in polynomial time. However, when the dimensionality of uncertainty is large, corresponding SDP instances become intractable in practice. Thus, new approximation methods that can trade off between solution quality and computation time are needed. For this purpose, we proposed an efficient approximation method based on principal component analysis (PCA), which reduces the dimensionality of the uncertainty space, with minimum loss of information. We showed that the proposed PCA approximation is a relaxation of the original problem when only a subset of the principal components are used. However, when all principal components are used the PCA approximation is exact. We also provided a theoretical bound on the differences between the optimal objective function values of the original problem and the proposed PCA approximation. This bound can serve as a guide to determine the number of principal components to use in practice and to allow for direct trade-offs between solution quality and runtime.

Finally, a comprehensive computational study using a distributionally robust CVaR problem with different covariance structures as well as a risk-averse production-transportation problem was conducted to show the strengths of the proposed PCA approximation. We have observed reductions in runtime as we used fewer principal components with reasonable deviations in accuracy. For instance, in many cases for the CVaR problem and the risk-averse production-transportation problem, using only 50% of the principal components provides a near-optimal solution with an optimality gap of less than 1% while reducing the runtimes by orders of magnitude. Our proposed approximation scheme provides decision makers with greater flexibility in dealing with the computational challenges of solving large-scale DRO problems, allowing for direct control of the trade-offs between solution quality and runtime.

Future work will consider extending our PCA-based approximation technique for DRO with moment-based ambiguity sets to the context of DRO with Wasserstein ambiguity sets. Further, there would be significant value in determining whether there is an equivalence between regularization and DRO with moment-based ambiguity sets.

Acknowledgments. The authors would like to thank the two anonymous referees and the Associate Editor Andrzej Ruszczynski for their very useful comments and suggestions, which helped to improve our paper significantly.

REFERENCES

[1] G. Bayraksan and D. K. Love, *Data-driven stochastic programming using phi-divergences*, in The Operations Research Revolution, Tutor. Oper. Res., D. M. Aleman, and A. C. Thiele, eds., INFORMS, 2015, pp. 1–19; available at https://pubsonline.informs.org/doi/book/10.1287/educ.2015.

[2] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski, *Robust Optimization*, Princeton University Press, Princeton, NJ, 2009.

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

[4] D. Bertsimas, E. Litvinov, X. A. Sun, J. Zhao, and T. Zheng, *Adaptive robust optimization for the security constrained unit commitment problem*, IEEE Trans. Power Systems, 28 (2013), pp. 52–63.

[5] D. Bienstock, M. Chertkov, and S. Harnett, *Chance-constrained optimal power flow: Risk-aware network control under uncertainty*, SIAM Rev., 56 (2014), pp. 461–495.

[6] D. Bertsimas, X. V. Doan, K. Natarajan, and C.-P. Teo, *Models for minimax stochastic linear optimization problems with risk aversion*, Math. Oper. Res., 35 (2010), pp. 580–602.

[7] J. Blanchet, Y. Kang, and K. Murthy, *Robust Wasserstein Profile Inference and Applications to Machine Learning*, preprint, arXiv:1610.05627, 2016.

[8] J. Cheng, E. Delage, and A. Lisser, *Distributionally robust stochastic knapsack problem*, SIAM J. Optim., 24 (2014), pp. 1485–1506.

[9] J. Cheng and A. Lisser, *Maximum probability shortest path problem*, Discrete Appl. Math., 192 (2015), pp. 40–48.

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

[11] C. Eckart and G. Young, *The approximation of one matrix by another of lower rank*, Psychometrika, 1 (1936), pp. 211–218.

[12] E. Erdoğan and G. Iyengar, *Ambiguous chance constrained problems and robust optimization*, Math. Program., 107 (2006), pp. 37–61.

[13] P. M. Esfahani and D. Kuhn, *Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations*, Math. Program. (2017), https://doi.org/10.1007/s10107-017-1172-1.

[14] R. Gao, X. Chen, and A. J. Kleywegt, *Distributional Robustness and Regularization in Statistical Learning*, preprint, arXiv:1712.06050, 2017.

[15] R. Gao and A. J. Kleywegt, *Distributionally Robust Stochastic Optimization with Wasserstein Distance*, preprint, arXiv:1604.02199, 2016.

[16] R. Gao and A. J. Kleywegt, *Distributionally Robust Stochastic Optimization with Dependence Structure*, preprint, arXiv:1701.04200, 2017.

[17] J. Gotoh, M. J. Kim, and A. Lim, *Robust Empirical Optimization Is Almost the Same as Mean-Variance Optimization*, preprint, https://ssrn.com/abstract=3073013, 2015.

[18] M. Grant and S. Boyd, *CVX: Matlab Software for Disciplined Convex Programming*, Version 2.0 Beta, http://cvxr.com/cvx, September 2013.

[19] M. Grant and S. Boyd, *Graph implementations for nonsmooth convex programs*, in Recent Advances in Learning and Control, Lect. Notes Control Inform. Sci. 371, V. D. Blondel, S. P. Boyd, and H. Kimura, eds., Springer, London, 2008, pp. 95–110.

[20] G. A. Hanasusanto, V. Roitch, D. Kuhn, and W. Wiesemann, *A distributionally robust perspective on uncertainty quantification and chance constrained programming*, Math. Program., 151 (2015), pp. 35–62.

[21] *High Dividend Stocks by Sector*, http://www.doubledividendstocks.com, August 26, 2015.

[22] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, 2012.

[23] K. Isii, *On sharpness of Tchebycheff-type inequalities*, Ann. Inst. Statist. Math., 14 (1962), pp. 185–197.

[24] R. Jiang and Y. Guan, *Data-driven chance constrained stochastic program*, Math. Program., 158 (2016), pp. 291–327.

[25] P. P. Khargonekar, I. R. Petersen, and K. Zhou, *Robust stabilization of uncertain linear systems: Quadratic stabilizability and H^∞ control theory*, IEEE Trans. Automat. Control, 35 (1990), pp. 356–361.

[26] H. Lam, *Robust sensitivity analysis for stochastic systems*, Math. Oper. Res., 41 (2016), pp. 1248–1275.

[27] B. Lindgren, *Statistical Theory*, 4th ed., Chapman & Hall/CRC Texts Statis. Sci. 22, CRC Press, New York, 1993.

[28] K. Natarajan and C.-P. Teo, *On reduced semidefinite programs for second order moment bounds with applications*, Math. Program., 161 (2017), pp. 487–518.

[29] K. Natarajan, M. Sim, and J. Uichanco, *Tractable robust expected utility and risk models for portfolio optimization*, Math. Finance, 20 (2010), pp. 695–731.

[30] A. Nemirovski and A. Shapiro, *Convex approximations of chance constrained programs*, SIAM J. Optim., 17 (2006), pp. 969–996.

[31] I. Popescu, *A semidefinite programming approach to optimal-moment bounds for convex classes of distributions*, Math. Oper. Res., 30 (2005), pp. 632–657.

[32] I. Popescu, *Robust mean-covariance solutions for stochastic optimization*, Oper. Res., 55 (2007), pp. 98–112.

[33] A. Prékopa, *Stochastic Programming*, Kluwer Academic Publishers, Dordrecht, 1995.

[34] R. T. Rockafellar and S. Uryasev, *Optimization of conditional value-at-risk*, J. Risk, 2 (2000), pp. 21–42.

[35] H. Scarf, *A min-max solution of an inventory problem*, in Studies in the Mathematical Theory of Inventory and Production, Stanford University Press, Redwood City, CA, 1958, pp. 201–209.

[36] S. Shafieezadeh-Abadeh, P. M. Esfahani, and D. Kuhn, *Distributionally robust logistic regression*, in Adv. Neural Inf. Process. Syst. 28, Curran Associates, Red Hook, NY, 2015, pp. 1576–1584.

[37] S. Shafieezadeh-Abadeh, D. Kuhn, and P. M. Esfahani, *Regularization via Mass Transportation*, preprint, arXiv:1710.10016, 2017.

[38] A. Shapiro, *On duality theory of conic linear problems*, in Semi-Infinite Programming, Nonconvex Optim. Appl. 57, Springer, New York, 2001, 135–165.

[39] A. Shapiro, D. Dentcheva, and A. Ruszczyński, *Lectures on stochastic programming: Modeling and theory*, MOS-SIAM Ser. Optim. 16, SIAM, Philadelphia, PA, 2014.

[40] B. P. G. Van Parys, P. J. Goulart, and D. Kuhn, *Generalized Gauss inequalities via semidefinite programming*, Math. Program., 156 (2016), pp. 271–302.

[41] B. P. G. Van Parys, P. J. Goulart, and M. Morari, *Distributionally robust expectation inequalities for structured distributions*, Math. Program. (2017), https://doi.org/10.1007/s10107-017-1220-x.

[42] S. Wold, K. Esbensen, and P. Geladi, *Principal component analysis*, Chemometrics and Intell. Laboratory Syst., 2 (1987), pp. 37–52.

[43] W. Wiesemann, D. Kuhn, and M. Sim, *Distributionally robust convex optimization*, Oper. Res., 62 (2014), pp. 1358-1376.

[44] S. Zymler, D. Kuhn, and B. Rustem, *Distributionally robust joint chance constraints with second-order moment information*, Math. Program., 137 (2013), pp. 167–198.

[45] A. Ruszczyński, *Probabilistic programming with discrete distributions and precedence constrained knapsack polyhedra*, Math. Program., 93 (2002), pp. 195–215.