Preprints, 10th IFAC International Symposium on
Advanced Control of Chemical Processes
Shenyang, Liaoning, China, July 25-27, 2018

Robust Optimization in High-Dimensional Data Space with Support Vector Clustering

Chao Shang * **Fengqi You** *

* *Robert Frederick Smith School of Chemical and Biomolecular Engineering, Cornell University, Ithaca, New York 14853, USA (E-mail: {chao.shang,fengqi.you}@cornell.edu).*

Abstract: Data-driven robust optimization has attracted immense attentions. In this work, we propose a data-driven uncertainty set for robust optimization under high-dimensional uncertainty. We propose to first decompose the high-dimensional data space into the principal subspace and the residual subspace by employing principal component analysis, and then adopt support vector clustering and classic polyhedral uncertainty set to describe the intricate geometry in the principal subspace and the tiny variations in the residual subspace, respectively, giving rise to a new data-driven uncertainty set. Similar to classic uncertainty sets, the proposed data-driven uncertainty set can also preserve the tractability of robust optimization problems. In addition, we establish the probabilistic guarantee theoretically by further calibrating the uncertainty set with an independent dataset, which ensures that the data-driven uncertainty set covers a portion of uncertainty with a given confidence level. Numerical results show the effectiveness of the proposed uncertainty set in reducing conservatism of robust optimization problems as well as the fidelity of the established probabilistic guarantee.

Keywords: Data-based decision-making, robust optimization, support vector clustering, principal component analysis, dimension reduction.

1. INTRODUCTION

Robust optimization (RO) has been a basic tool for decision-making under uncertainty (Ben-Tal et al. (2009); Gabrel et al. (2014)). In process systems engineering, RO has found extensive applications spanning across process network design (Gong et al. (2016); Gong and You (2017)), supply chain management (Tong et al. (2014); Yue and You (2016)), and process scheduling (Lappas and Gounaris (2016); Shi and You (2016)).

A key ingredient of RO is the uncertainty set that is used to describe the possible realization of uncertain parameters. Classic uncertainty sets that are widely adopted include the box set, the ellipsoidal set, the polyhedral uncertainty set, etc. The geometry and size of the uncertainty sets exert paramount influence on the quality of solutions to RO problems. On one hand, the uncertainty set shall cover all possible realizations of uncertainties to provide adequate safeguards. On the other hand, unnecessary coverage shall be removed to avoid over-conservative solutions. Therefore, a common criticism on generic uncertainty sets is that all of them have fixed geometric shapes, thereby lacking sufficient flexibility to describe the high probability region of the underlying distribution $\mathbb{P}$.

In practice, we commonly have no idea about the underlying distribution $\mathbb{P}$ of uncertainty. The proliferation of data nowadays provides more room to extract meaningful information about $\mathbb{P}$, which is in line with the spirit of big data analytics (Qin (2014)). Based on such a motivation, more and more research attentions have been paid to data-driven robust optimization recently. The key idea is to extract support information from data to develop a high-quality uncertainty set. Ferreira et al. (2012) propose to parameterize generic norm-induced uncertainty sets based on principal component analysis (PCA) and minimum power decomposition (MPD). Ning and You (2017) propose to adopt the Dirichlet process (DP) mixture model to extract the support information from data. By using the union of several basic norm-induced uncertainty sets, the resulting uncertainty set provides better representation capability. In virtue of support vector clustering (SVC), Shang et al. (2017) propose an effective kernel learning approach to construct a non-parametric uncertainty set. An attractive feature of this work is that the fraction of data coverage can be explicitly controlled in the modeling stage, which provides considerable convenience in practical use. In addition, only quadratic programs (QPs) need to be solved, which are computationally thrifty.

Under high-dimensional uncertainty, it is often the case that significant correlations exist. This phenomenon is typically termed as *data rich but information poor*. More concretely, there exists a low-dimensional subspace that explains most information within high-dimensional data. Under such circumstance, some data-driven approaches may lose effects due to the ignorance of the reduced-dimensional subspace. For example, the kernel learning approach proposed by Shang et al. (2017) is typically prone to the curse of dimensionality. To reasonably approximate the high probability region of high-dimensional

uncertainty, one needs to collect millions of data samples for modeling, which is rather unrealistic in practice.

Therefore, in this article, we develop an efficient data-driven approach tailored to RO problems under high-dimensional uncertainty. To construct the data-driven uncertainty set, we propose to first carry out dimension reduction by means of PCA. Then the SVC-based and the generic norm-induced uncertainty sets are adopted to characterize variations in the low-dimensional principal subspace (PS) and residual subspace (RS), respectively. The rationale lies in that, since most information is abstracted in the low-dimensional PS, it deserves more attention and its complicated geometry shall be subtly described. This is achieved by harnessing the modeling power of the SVC-based uncertainty set proposed by Shang et al. (2017). Meanwhile, generic polyhedral uncertainty sets suffice to capture the slight variations within the RS. In this way, a data-driven uncertainty set can be developed, which accurately delineates the shape of high-dimensional uncertainty data, and turns out to be helpful for deriving high-quality solutions and reducing conservativeness. The proposed uncertainty set leads to a tractable reformulation of RO problems with linear constraints, which brings computational benefits. We further establish the probabilistic guarantee in theory by calibrating the uncertainty set after the modeling phase, which ensures that the data-driven uncertainty set covers a certain portion of uncertainty with a given confidence level. Numerical examples are presented to demonstrate the effectiveness of the proposed approach in alleviating the curse of dimensionality brought by high-dimensional uncertainty in RO problems.

2. PRELIMINARIES

2.1 SVC-Based Uncertainty Set Constructions using the Weighted Generalized Intersection Kernel

SVC is a well-established machine learning approach for estimating the support of an unknown distribution from observation data (Ben-Hur et al. (2001)). The optimization problem of SVC aims at minimizing the radius R of a circle enclosing most data samples $\{\boldsymbol{\phi}(\mathbf{u}^{(i)})\}$ in the feature space while penalizing potential outliers residing outside the circle:

$$\begin{aligned}\min_{\mathbf{a},R,\boldsymbol{\xi}}\ & R^2+\frac{1}{N\nu}\sum_{i=1}^{N}\xi_i\\ \text{s.t. }& ||\boldsymbol{\phi}(\mathbf{u}^{(i)})-\mathbf{a}||^2\le R^2+\xi_i,\ i=1,\cdots,N\\ & \xi_i\ge 0,\ i=1,\cdots,N\end{aligned}\tag{1}$$

where $\mathbf{a}$ is the center of the circle, $\boldsymbol{\phi}(\cdot)$ denotes the feature mapping, and the slack variable ξ indicates whether $\boldsymbol{\phi}(\mathbf{u}^{(i)})$ lies outside the circle. ν is a regularization parameter used to balance between minimizing the volume of the circle and penalizing outliers. By introducing Lagrange multipliers $\boldsymbol{\alpha}$, we could arrive at the following equivalent dual problem, which is a QP:

$$\begin{aligned}\max_{\boldsymbol{\alpha}}\ & -\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j K(\mathbf{u}^{(i)},\mathbf{u}^{(j)})+\sum_{i=1}^{N}\alpha_i K(\mathbf{u}^{(i)},\mathbf{u}^{(i)})\\ \text{s.t. }& 0\le\alpha_i\le 1/N\nu,\ i=1,\cdots,N\\ & \sum_{i=1}^{N}\alpha_i=1\end{aligned}\tag{2}$$

where $K(\mathbf{u}^{(i)},\mathbf{u}^{(j)})=\boldsymbol{\phi}(\mathbf{u}^{(i)})^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{u}^{(j)})$ stands for the kernel function. The radial basis function (RBF) kernel $K(\mathbf{u},\mathbf{v})=\exp\{-||\mathbf{u}-\mathbf{v}||^2/2\sigma^2\}$ is mostly adopted in SVC. However, as pointed in Shang et al. (2017), the RBF kernel is unsuitable for uncertainty set construction because the computational tractability of RO problems will be disrupted. For this reason, Shang et al. (2017) further propose the following weighted generalized intersection kernel (WGIK):

$$K(\mathbf{u},\mathbf{v})=\sum_{k=1}^{n}l_k-||\mathbf{Q}(\mathbf{u}-\mathbf{v})||_1,\tag{3}$$

which leads to the following data-driven uncertainty set:

$$\begin{aligned}\mathcal{U}_{\mathrm{SVC}}(\nu,\mathcal{D})&=\left\{\mathbf{u}\,\middle|\,||\boldsymbol{\phi}(\mathbf{u})-\mathbf{p}||^2\le R^2\right\}\\&=\left\{\mathbf{u}\,\middle|\,\sum_{i\in\mathrm{SV}}\alpha_i||\mathbf{Q}(\mathbf{u}-\mathbf{u}^{(i)})||_1\le\theta\right\}\end{aligned}\tag{4}$$

where

$$\theta=\sum_{i\in\mathrm{SV}}\alpha_i||\mathbf{Q}(\mathbf{u}^{(i')}-\mathbf{u}^{(i)})||_1,\ i'\in\mathrm{BSV}.\tag{5}$$

It can be observed that the data-driven uncertainty set $\mathcal{U}_\nu(\mathcal{D})$ is essentially a *polyhedron*, defined based on SVs with $\alpha_i>0$, thereby bearing a non-parametric scheme. In this way, it can preserve the tractability of RO problems.

A desirable feature of $\mathcal{U}_{\mathrm{SVC}}(\nu,\mathcal{D})$ is that it approximately covers $(1-\nu)\times 100\%$ of N training data samples, which provides a convenient way to adjust the conservative of the uncertainty set.

2.2 PCA for Dimension Reduction

PCA is a basic method for dimension reductions. Assume that all available data samples have been stacked into a data matrix $\mathbf{X}\in\mathbb{R}^{N\times m}$, and each dimension has been scaled to zero mean and unit variance. Performing singular value decomposition (SVD) on the covariance matrix of data yields:

$$\mathbf{R}=\frac{1}{N-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}=\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathrm{T}},\tag{6}$$

where $\boldsymbol{\Lambda}=\mathrm{diag}\{\lambda_1,\cdots,\lambda_m\}$ is a diagonal matrix whose diagonal elements are arranged in a descending order. Denoting by $\mathbf{P}=\mathbf{U}(:,1:A)$ the matrix formed by eigenvectors in $\mathbf{U}$ associated with A largest eigenvalues, we could obtain the following PCA model for dimension reduction:

$$\mathbf{X}=\mathbf{T}\mathbf{P}^{\mathrm{T}}+\tilde{\mathbf{T}}\tilde{\mathbf{P}}^{\mathrm{T}}\tag{7}$$

$$\mathbf{T}=\mathbf{X}\mathbf{P}\tag{8}$$

$$\tilde{\mathbf{T}}=\mathbf{X}\tilde{\mathbf{P}}\tag{9}$$

where $\mathbf{T}\in\mathbb{R}^{N\times A}$ and $\tilde{\mathbf{T}}\in\mathbb{R}^{N\times(m-A)}$ are the score matrices for principal components (PCs) and residuals, respectively. $\mathbf{P}\in\mathbb{R}^{m\times A}$ and $\tilde{\mathbf{P}}\in\mathbb{R}^{m\times(m-A)}$ are loading

matrices for PCs and residuals. A is the number of PCs. $\hat{\mathbf{X}} = \mathbf{T}\mathbf{P}^{\mathrm{T}}$ explains most variations in high-dimensional PS, while $\tilde{\mathbf{X}} = \tilde{\mathbf{T}}\tilde{\mathbf{P}}^{\mathrm{T}}$ primarily reveals unimportant noise in RS.

3. A DATA-DRIVEN ROBUST OPTIMIZATION APPROACH UNDER HIGH-DIMENSIONAL UNCERTAINTIES

3.1 Data-Driven Construction of Uncertainty Sets

The classic PCA model (7)-(9) makes only use of second-order information within data. In one sense, it falls short of describing more intricate geometry of data space, especially *the variations in the low-dimensional PS*. The loss of high-order information may lead to suboptimal solutions to RO problems. Therefore, we propose to further describe the variations in PS in virtue of the modeling power of SVC. That is, to run the SVC algorithm with N score samples $\mathcal{T} = \{\mathbf{t}^{(i)}\}$ based on a pre-specified ν, where $\mathbf{t}^{(i)}$ is the ith column of $\mathbf{T}^{\mathrm{T}}$, and hence whitening matrix adopted in (3) is given by $\mathbf{Q} = \mathbf{\Lambda}^{-\frac{1}{2}}$. We denote this uncertainty set obtained as $\mathcal{U}_{\nu}(\mathcal{T})$.

To characterize tiny variations within the RS, we propose to employ traditional norm-based polyhedral set:

$$\mathcal{U}_{\mathrm{poly}}(\Gamma) = \left\{ \tilde{\mathbf{t}} \,\middle|\, \sum_{k} |\tilde{t}_k| \leq \Gamma \right\}, \tag{10}$$

where Γ is the budget parameter for controlling the size of $\mathcal{U}_{\mathrm{poly}}(\Gamma)$.

By delineating variations within two subspaces individually, we propose a new data-driven uncertainty set for high-dimensional uncertainty, formally expressed as:

$$\mathcal{U}_{\nu,A,\Gamma}(\mathcal{D}) = \left\{ \mathbf{u} \,\middle|\, \begin{array}{l} \mathbf{u} = \mathbf{P}\mathbf{t} + \tilde{\mathbf{P}}\tilde{\mathbf{t}} \\ \mathbf{t} \in \mathcal{U}_{\mathrm{SVC}}(\nu, \mathcal{T}) \\ \tilde{\mathbf{t}} \in \mathcal{U}_{\mathrm{poly}}(\Gamma) \end{array} \right\}. \tag{11}$$

Due to the orthogonal property of the loading matrix $[\mathbf{P}\ \tilde{\mathbf{P}}]$, we can further rewrite $\mathcal{U}_{\nu,A,\Gamma}(\mathcal{D})$ as the intersection of two basic uncertainty sets:

$$\begin{aligned} &\mathcal{U}_{\nu,A,\Gamma}(\mathcal{D}) \\ &= \left\{ \mathbf{u} \,\middle|\, \mathbf{P}^{\mathrm{T}}\mathbf{u} \in \mathcal{U}_{\mathrm{SVC}}(\nu, \mathcal{T}) \right\} \cap \left\{ \mathbf{u} \,\middle|\, \tilde{\mathbf{P}}^{\mathrm{T}}\mathbf{u} \in \mathcal{U}_{\mathrm{poly}}(\Gamma) \right\} \\ &\triangleq \mathcal{U}_{PS} \cap \mathcal{U}_{RS} \end{aligned} \tag{12}$$

Notice that the proposed data-driven uncertainty set is essentially parameterized by three parameters $\{A, \nu, \Gamma\}$. The selection of the number A of PCs can be made based on existing methods for PCA, such as the cumulative percent variance (CPV) criterion, the Akaike information criterion (AIC), the minimum description length (MDL) criterion, and so on. As with the determination of values of ν and Γ, it will be discussed in the sequel.

3.2 Tractable Reformulation

A key attribute of uncertainty sets in RO is that they must lead to tractable reformulations of optimization problems. It has been proved by Shang et al. (2017) that the SVC-based uncertainty set enables an LP reformulation of the worst-case robust linear constraint. In fact, for the proposed uncertainty set $\mathcal{U}_{\nu,A,\Gamma}(\mathcal{D})$, we could also establish a similar result based on the following theorem.

Theorem 1: The worst-case robust linear constraint

$$\max_{\boldsymbol{\xi} \in \mathcal{U}_{\nu,A,\Gamma}(\mathcal{D})} \boldsymbol{\xi}^{\mathrm{T}}\mathbf{x} \leq b \tag{13}$$

is equivalent to the following problem:

$$\left\{ \begin{array}{l} \displaystyle\sum_{i \in \mathrm{SV}} (\boldsymbol{\mu}_i - \boldsymbol{\lambda}_i)^{\mathrm{T}} \mathbf{Q}\mathbf{t}^{(i)} + \eta\theta + p\Gamma \leq b \\ \displaystyle\sum_{i \in \mathrm{SV}} \mathbf{Q}(\boldsymbol{\lambda}_i - \boldsymbol{\mu}_i) + \mathbf{P}^{\mathrm{T}}\mathbf{x} = \mathbf{0} \\ \boldsymbol{\lambda}_i + \boldsymbol{\mu}_i = \eta \cdot \alpha_i \cdot \mathbf{1},\ \boldsymbol{\lambda}_i, \boldsymbol{\mu}_i \in \mathbb{R}^A_+,\ \forall i \in \mathrm{SV} \\ \eta \geq 0 \\ p \geq |\tilde{\mathbf{p}}_k^{\mathrm{T}}\mathbf{x}|,\ \forall k = 1, \cdots, m - A \end{array} \right. \tag{14}$$

which is essentially an LP.

The proof is omitted here due to page limitations. It can be obtained by using the strong duality of LPs as well as the boundness of $\mathcal{U}_{PS}$ and $\mathcal{U}_{RS}$.

3.3 Data-Driven Calibration with Probabilistic Guarantee

In a data-driven context, the resulting uncertainty set is uncertain itself because data collected contain randomness. Therefore, a desirable data-driven uncertainty set should contain the high-density region of uncertainty with high confidence. Therefore, we focus on the following probabilistic guarantee (Hong et al. (2016)):

$$\mathbb{P}_{\mathcal{D}} \left\{ \mathbb{P}(\boldsymbol{\xi} \in \mathcal{U}(\mathcal{D})) \geq 1 - \alpha \right\} \geq 1 - \beta. \tag{15}$$

Here, $\mathbb{P}_{\mathcal{D}}\{\cdot\}$ is taken with respect to *data*, whose sampling involves uncertainty, while $\mathbb{P}\{\cdot\}$ is taken with respect to the uncertainty $\boldsymbol{\xi}$. The fraction of uncertainty coverage to be achieved is denoted as $1 - \alpha$, whereas the confidence level of the event $\mathbb{P}(\boldsymbol{\xi} \in \mathcal{U}(\mathcal{D})) \geq 1 - \alpha$ is $1 - \beta$. An efficient data-driven strategy for uncertainty set calibration is proposed by Hong et al. (2016), which ensures (15) to hold. The idea is to split all available data $\mathcal{D}$ into a training dataset $\mathcal{D}_1$ and a calibration dataset $\mathcal{D}_2$, which include N_1 and N_2 data samples, respectively. A basic requirement to employ this procedure is that $\mathcal{U}(\mathcal{D})$ can be expressed in a parametric form of

$$\mathcal{U}(\mathcal{D}) = \{ \boldsymbol{\xi} \,|h(\boldsymbol{\xi}) \leq 0 \}, \tag{16}$$

where $h(\boldsymbol{\xi})$ is a certain scalar function established using the training dataset $\mathcal{D}_1$ in the first step. After that, we compute the value of $h(\boldsymbol{\xi})$ on the calibration dataset $\mathcal{D}_2$ and obtain $\{h(\boldsymbol{\xi}_n), \boldsymbol{\xi}_n \in \mathcal{D}_2\}$, which can be further ordered as $h(\boldsymbol{\xi}_{(1)}) < h(\boldsymbol{\xi}_{(2)}) < \cdots < h(\boldsymbol{\xi}_{(N_2)})$. Computing the optimal index r^* as

$$r^* = \min \left\{ r : \sum_{k=0}^{r-1} \binom{N_2}{k} (1 - \alpha)^k \alpha^{N-k} \geq 1 - \beta \right\}, \tag{17}$$

and further calibrating the data-driven uncertainty set as

$$\mathcal{U}(\mathcal{D}) = \left\{ \boldsymbol{\xi} \,\middle|\, h(\boldsymbol{\xi}) \leq h(\boldsymbol{\xi}_{(r^*)}) \right\}, \tag{18}$$

the probabilistic guarantee (15) will hold for the uncertainty set (18) after calibrations (Hong et al. (2016)). In addition, due to the discrete nature of the minimizer r^*, the true confidence level $1 - \beta_{true}$ can be derived as

$$1 - \beta_{true} = \sum_{k=0}^{r^*-1} \binom{N_2}{k} (1 - \alpha)^k \alpha^{N-k} \tag{19}$$

which will be always larger than it prespecified value $1-\beta$.

Our goal is to devise a similar scheme that ensure the probabilistic guarantee for the proposed data-driven uncertainty set (12). However, the aforementioned result cannot be directly applied. Notice that (12) can be regarded as the intersection of two basic uncertainty sets having the form of (16). Based on such an observation, we establish the following theorem:

Theorem 2: *For the proposed uncertainty set (12), if*

$$\mathbb{P}_{\mathcal{D}}\left\{\mathbb{P}(\boldsymbol{\xi}\in\mathcal{U}_{PS})\geq 1-\alpha\cdot z_1^{\alpha}\right\}\geq 1-\beta\cdot z_1^{\beta}, \tag{20}$$

and

$$\mathbb{P}_{\mathcal{D}}\left\{\mathbb{P}(\boldsymbol{\xi}\in\mathcal{U}_{RS})\geq 1-\alpha\cdot z_2^{\alpha}\right\}\geq 1-\beta\cdot z_2^{\beta} \tag{21}$$

with

$$z_1^{\alpha}+z_2^{\alpha}=1,\ z_1^{\beta}+z_2^{\beta}=1, \tag{22}$$

then the following probabilistic guarantee holds

$$\mathbb{P}_{\mathcal{D}}\left\{\mathbb{P}(\boldsymbol{\xi}\in\mathcal{U}_{\nu,A,\Gamma}(\mathcal{D}))\geq 1-\alpha\right\}\geq 1-\beta. \tag{23}$$

Proof: We denote the events $\boldsymbol{\xi}\in\mathcal{U}_{PS}$, $\boldsymbol{\xi}\in\mathcal{U}_{RS}$ and $\boldsymbol{\xi}\in\mathcal{U}_{\nu,A,\Gamma}(\mathcal{D})$ as B_1, B_2 and B, respectively, and

$$\mathbb{P}(B_1)\geq 1-\alpha\cdot z_1^{\alpha} \tag{24}$$
$$\mathbb{P}(B_2)\geq 1-\alpha\cdot z_2^{\alpha} \tag{25}$$
$$\mathbb{P}(B)\geq 1-\alpha \tag{26}$$

as C_1, C_2 and C, respectively. Notice that $B=B_1\cap B_2$. Next we show that $C\supseteq C_1\cap C_2$. Assume that (21) and (22) hold, then we have:

$$\begin{aligned}\mathbb{P}(B)&=\mathbb{P}(B_1\cap B_2)\\&\geq\mathbb{P}(B_1)+\mathbb{P}(B_2)-1\\&=(1-\alpha\cdot z_1^{\alpha})+(1-\alpha\cdot z_2^{\alpha})-1\\&=1-\alpha\end{aligned} \tag{27}$$

where the second inequality arises from a well-known probability inequality. This indicates the event C must occur on the condition of $C_1\cap C_2$. Then we can proceed in the following way:

$$\begin{aligned}\mathbb{P}_{\mathcal{D}}\{C\}&\geq\mathbb{P}_{\mathcal{D}}\{C_1\cap C_2\}\\&\geq\mathbb{P}_{\mathcal{D}}\{C_1\}+\mathbb{P}_{\mathcal{D}}\{C_2\}-1\\&=(1-\beta\cdot z_1^{\beta})+(1-\beta\cdot z_2^{\beta})-1\\&=1-\beta\end{aligned} \tag{28}$$

This completes the proof.

The above theorem indicates that, we only need to establish probabilistic guarantees for two uncertainty sets, respectively. This can be easily done by directly adopting the procedure proposed by Hong et al. (2016) twice, since both $\mathcal{U}_{PS}$ and $\mathcal{U}_{RS}$ admit parametric expressions in the form of (16):

$$\mathcal{U}_{PS}=\left\{\mathbf{u}\,\middle|\,\sum_{i\in\mathrm{SV}}\alpha_i||\boldsymbol{\Lambda}^{-\frac{1}{2}}(\mathbf{P}^{\mathrm{T}}\mathbf{u}-\mathbf{t}^{(i)})||_1-\theta\leq 0\right\}, \tag{29}$$

$$\mathcal{U}_{RS}=\left\{\mathbf{u}\,\middle|\,||\tilde{\mathbf{P}}^{\mathrm{T}}\mathbf{u}||_1-\Gamma\leq 0\right\}. \tag{30}$$

To build the SVC-based uncertainty set in PS, a reasonable choice of ν is to set it as α in practice, because the SVC-based uncertainty set will approximately capture $(1-\nu)\times 100\%$ of data samples in $\mathcal{D}_1$, which is in line with the spirit of the probabilistic guarantee implying that $\mathbb{P}(\boldsymbol{\xi}\in\mathcal{U}_{\nu,A,\Gamma}(\mathcal{D}))\geq 1-\alpha$ holds with high proability. In addition, we suggest using $z_1^{\alpha}=z_2^{\alpha}=z_1^{\beta}=z_2^{\beta}=1/2$ as a trivial choice.

4. A NUMERICAL EXAMPLE

4.1 Numerical Experiment Setup

In this section, we verify the effectiveness of the proposed data-driven uncertainty set in RO based on a numerical example. We consider the following simple RO problem:

$$\begin{aligned}\max_{\mathbf{x}}\min_{\mathbf{u}}\ &\mathbf{c}^{\mathrm{T}}\mathbf{x}\\\text{s.t. }&\mathbf{a}^{\mathrm{T}}\mathbf{x}+\mathbf{u}^{\mathrm{T}}\mathbf{x}\leq b\\&\mathbf{D}\mathbf{x}\leq\mathbf{f}\end{aligned} \tag{31}$$

where $\mathbf{x}\in\mathbb{R}^{10}$ includes decision variables. $\mathbf{u}\in\mathbb{R}^{10}$ represents random perturbations on the coefficients $\mathbf{a}$. Here only the first constraint is affected by uncertainties, and the remaining ones are expressed as $\mathbf{D}\mathbf{x}\leq\mathbf{f}$. It is assumed that we do not know the distribution of $\mathbf{u}$ but have some data samples $\{\mathbf{u}^{(i)}\}$, which are generated by means of a latent variable model:

$$\mathbf{u}=\mathbf{W}\mathbf{z}+\boldsymbol{\epsilon}. \tag{32}$$

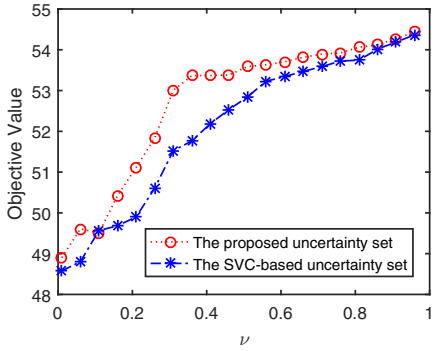
Here $\mathbf{z}\in\mathbb{R}^2$ include latent variables, $\mathbf{W}$ is a transformation matrix, and $\boldsymbol{\epsilon}$ denotes an isotropic measurement noise following Gaussian distribution. Therefore, $\mathbf{u}$ will have significant correlations, and most variations can be explained by a two-dimensional latent subspace. We generate three different datasets, in which the latent variables $\mathbf{z}$ follow Gaussian distribution, mixture Gaussian distribution, and bivariate Gamma distribution, respectively.

4.2 Optimization Performance of the Proposed Uncertainty Set

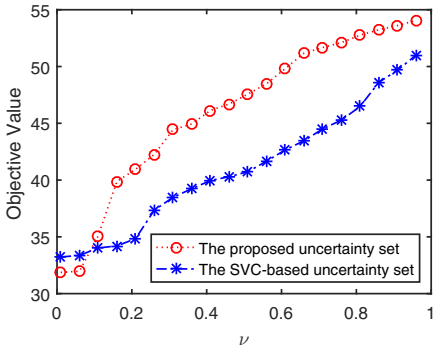
We first compare the optimization performance of the proposed method with the RO method based on the generic SVC-based uncertainty set proposed by Shang et al. (2017). Since (31) is an RO problem with a linear robust constraint, it can be easily transformed into an equivalent LP based on both uncertainty sets. The QPs involved modeling SVC are solved using the CVX package in MATLAB 2016a, and the resulting equivalent problems in the form of LP are solved with CPLEX.

The optimization performances under three different settings of uncertainties are reported in Fig. 1. In each setting, 100 samples are collected as training data for uncertainty set constructions. The number of PCs in the proposed uncertainty set is set as 2. We deliberately set the value of Γ as the maximum on the training dataset so as to capture all variations in RS. In this way, the conservatism of both the generic SVC-based uncertainty set and the proposed one can be adjusted by the same parameter ν, leading to a fair comparison. That is, with the same value of ν used, two uncertainty sets have nearly the same fractions of data coverage. As can be seen from Fig. 1, when ν increases, the sizes of both uncertainty sets become smaller, the solutions become less conservative, and higher objective values will be obtained. Most importantly, the proposed uncertainty set induces much higher objective values than the SVC-based uncertainty set under the same value of ν, indicating less conservative solutions. This can be explained by the fact that the generic SVC-based uncertainty set directly models a 10-dimensional data space with only 100 available data samples, thereby being prone to the curse of dimensionality. By contrast, the proposed uncertainty

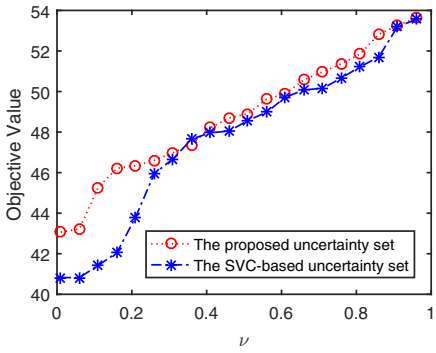
set captures the variations in the low-dimensional PS, thereby effectively alleviating the curse of dimensionality and giving a compact expression of the uncertainty set.



(a) Case 1



(b) Case 2



(c) Case 3

Fig. 1. Optimization performance comparisons of two robust optimization approaches.

Next, we examine the effect of the number of PCs on the optimization performance. The dataset with Gaussian distributed latent variables is used here. The corresponding results with different values of A are shown in Fig. 2. When $A = 1$, a large portion of information is still present in the RS, which cannot be well addressed by using a box uncertainty set. Therefore, the worst performance is obtained in the case of $A = 1$. When $A = 2$ and $A = 3$, the best performances can be derived, which is in line with the physical truth that a two-dimensional subspace explains most information of uncertainty variations. When $A = 4$, the performance begins to deteriorate. It implies that due to the curse of dimensionality, the approximation performance of kernel methods tends to be worse when the dimension of data space becomes higher. It shows the necessity to carry out dimension reduction to tackle high-dimensional uncertainties with inherent correlations.

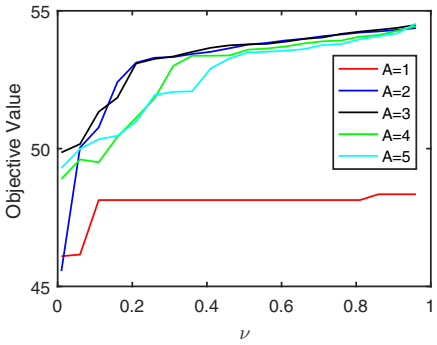


Fig. 2. Optimization performance based on the proposed uncertainty set with different number of PCs.

4.3 Performance of the Data-Driven Calibration Approach

In this subsection, we investigate the performance of the data-driven approach to uncertainty set calibrations by executing Monte Carlo simulations. We first investigate the effect of N_1 on optimization performance by varying N_1 and fixing N_2. $\{\alpha, \beta\}$ are set as $\alpha = 0.05$ and $\beta = 0.05$, and the optimal index r^* for model calibration is calculated based on (17). We perform Monte Carlo simulations by randomly generating data samples $\mathcal{D}$ based on the latent variable model (32), with the latent variable following the mixture Gaussian distribution. Each time we build an uncertainty set based on $\mathcal{D}_1$, while calibrating the uncertainty set with $\mathcal{D}_2$ based on the proposed strategy. After that, we solve the RO problem with the calibrated uncertainty set. We repeat this process 1,000 times, and the mean value and the standard deviation (s.t.d.) of objective values in 1,000 replications are calculated, which are reported in Table 1.

Table 1. Optimization Performance with Different Sizes of $\mathcal{D}_1$

N_1	N_2	Mean of Objective	s.t.d. of Objective
50	500	48.0537	2.1765
100	500	48.2876	1.6802
150	500	48.4716	1.2935

We can observe from Table 1 that when N_1 increases, the mean value of the optimal value also increases, and the s.t.d. of the optimal value decreases. This is rational because when more data samples enter into the model training phase, the estimation of the uncertainty set shall become more reliable, and hence conservatism of solution shall be reduced. It indicates that the proposed approach can effectively utilize information underlying data as an asset. When more data are utilized, more meaningful information can be extracted and integrated into the optimization model.

Finally, we investigate effect of the split ratio N_1/N_2 on optimization performance when the number of available samples is fixed. We assume that there are $N = 1000$ data

Table 2. Optimization Performance with Different Split Ratios of Available Data

N_1	N_2	$\hat{\beta}$	β_{true}	Mean of $\hat{\alpha}$	s.t.d. of $\hat{\alpha}$	Mean of Objective	s.t.d. of Objective
65	935	0.0060	0.0498	0.0327	0.0058	48.5552	1.7552
166	834	0.0050	0.0500	0.0319	0.0061	48.9297	1.0863
268	732	0.0050	0.0500	0.0320	0.0062	48.9463	1.0950
373	627	0.0120	0.0498	0.0301	0.0069	49.0794	0.7969
481	519	0.0120	0.0498	0.0292	0.0078	49.0446	0.7962
593	407	0.0120	0.0498	0.0276	0.0084	48.9522	0.8133
713	287	0.0130	0.0494	0.0261	0.0092	48.8240	0.8634

samples available in total. The results are listed in Table 2. Again, we generate the entire dataset $\mathcal{D}$ 1,000 times, and each time we evaluate the value $\mathbb{P}(\boldsymbol{\xi} \in \mathcal{U}_{\nu,A,\Gamma}(\mathcal{D}))$ empirically based on an independent dataset including 10,000 samples. In this way, the occurrence of the event $\mathbb{P}(\boldsymbol{\xi} \in \mathcal{U}_{\nu,A,\Gamma}(\mathcal{D})) \geq 1-\alpha$ as well as an empirical value of α can be evaluated. Finally, the value of $\hat{\beta}$ can be calculated by averaging the results of 1,000 Monte Carlo simulations.

By deliberately choosing the values of N_2, the true values β_{true} are nearly identical to the specified value 0.05 in all cases. We can see that the empirical values of $\hat{\beta}$ are both smaller than the theoretical value, thereby verifying the correctness of the established probabilistic guarantee. By taking a closer look at the variation trends of $\hat{\beta}$ and the mean of $\hat{\alpha}$, we can observe that with N_1 increasing the mean value of $\hat{\alpha}$ tends to decrease, and the value of $\hat{\beta}$ tends to increase. It indicates that when more data are used for model training, the confidence level that we observe the event $\mathbb{P}(\boldsymbol{\xi} \in \mathcal{U}_{\nu,A,\Gamma}(\mathcal{D})) \geq 1-\alpha$ becomes lower, whereas the actual "volume" of the uncertainty set becomes larger. This shows a trade-off underlying the probabilistic guarantee that we cannot simultaneously increase the data coverage and the confidence level. As with the optimal value, we can see that a balance shall be made between N_1 and N_2, since the case with $N_1 = 373$ gives the highest objective value on average, as well as the smallest s.t.d. It implies that the split ratio of the entire dataset should be carefully tuned in practice to reduce the conservatism and obtain a high-quality solution.

5. CONCLUSION

In this article, we put forward a new data-driven uncertainty set to deal with high-dimensional uncertainty in RO problems. PCA is performed on high-dimensional uncertainty data to decompose the data space into PS and RS. Then the generic SVC-based uncertainty set and the polyhedral uncertainty set are adopted to characterize the variations within PS and RS, respectively. In this way, more intricate geometry of high-dimensional uncertainties can be described by the proposed data-driven uncertainty set, and the curse of dimensionality can be mitigated. Due to the data-driven nature of the proposed approach, the uncertainty set itself has some randomness. We establish a probabilistic guarantee by using a portion of available data to calibrate the uncertainty set, which ensures that the uncertainty set can capture a prespecified portion of uncertainty with high probability. Numerical case studies are conducted to demonstrate the advantages of the proposed data-driven approach to uncertainty set constructions, and the effectiveness of the calibration procedure in ensuring the probabilistic guarantee.

REFERENCES

Ben-Hur, A., Horn, D., Siegelmann, H.T., and Vapnik, V. (2001). Support vector clustering. *Journal of Machine Learning Research*, 2(Dec), 125–137.

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

Ferreira, R., Barroso, L., and Carvalho, M. (2012). Demand response models with correlated price data: A robust optimization approach. *Applied Energy*, 96, 133–149.

Gabrel, V., Murat, C., and Thiele, A. (2014). Recent advances in robust optimization: An overview. *European Journal of Operational Research*, 235(3), 471–483.

Gong, J., Garcia, D.J., and You, F. (2016). Unraveling optimal biomass processing routes from bioconversion product and process networks under uncertainty: an adaptive robust optimization approach. *ACS Sustainable Chemistry & Engineering*, 4(6), 3160–3173.

Gong, J. and You, F. (2017). Optimal processing network design under uncertainty for producing fuels and value-added bioproducts from microalgae: Two-stage adaptive robust mixed integer fractional programming model and computationally efficient solution algorithm. *AIChE Journal*, 63(2), 582–600.

Hong, L.J., Huang, Z., and Lam, H. (2016). Approximating data-driven joint chance-constrained programs via uncertainty set construction. In *Winter Simulation Conference (WSC)*, 389–400. IEEE.

Lappas, N.H. and Gounaris, C.E. (2016). Multi-stage adjustable robust optimization for process scheduling under uncertainty. *AIChE Journal*, 62(5), 1646–1667.

Ning, C. and You, F. (2017). Data-driven adaptive nested robust optimization: General modeling framework and efficient computational algorithm for decision making under uncertainty. *AIChE Journal*, 63(9), 3790–3817.

Qin, S.J. (2014). Process data analytics in the era of big data. *AIChE Journal*, 60(9), 3092–3100.

Shang, C., Huang, X., and You, F. (2017). Data-driven robust optimization based on kernel learning. *Computers & Chemical Engineering*, 106, 464–479.

Shi, H. and You, F. (2016). A computational framework and solution algorithms for two-stage adaptive robust scheduling of batch manufacturing processes under uncertainty. *AIChE Journal*, 62(3), 687–703.

Tong, K., You, F., and Rong, G. (2014). Robust design and operations of hydrocarbon biofuel supply chain integrating with existing petroleum refineries considering unit cost objective. *Computers & Chemical Engineering*, 68, 128–139.

Yue, D. and You, F. (2016). Optimal supply chain design and operations under multi-scale uncertainties: Nested stochastic robust optimization modeling framework and solution algorithm. *AIChE Journal*, 62(9), 3041–3055.