Delft University of Technology

Machine learning enabled uncertainty set for data-driven robust optimization

Li, Yun; Yorke-Smith, Neil; Keviczky, Tamas

DOI
10.1016/j.jprocont.2024.103339

Publication date
2024

Document Version
Final published version

Published in
Journal of Process Control

Citation (APA)
Li, Y., Yorke-Smith, N., & Keviczky, T. (2024). Machine learning enabled uncertainty set for data-driven robust optimization. *Journal of Process Control*, *144*, Article 103339. https://doi.org/10.1016/j.jprocont.2024.103339

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.
For technical reasons the number of authors shown on this cover page is limited to a maximum of 10.

Contents lists available at ScienceDirect

Journal of Process Control

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

Machine learning enabled uncertainty set for data-driven robust optimization

Yun Li [a,*], Neil Yorke-Smith [b], Tamas Keviczky [a]

[a] Delft Center for Systems and Control, Delft University of Technology, Delft, The Netherlands
[b] Algorithmics Group, Delft University of Technology, Delft, The Netherlands

ARTICLE INFO

Keywords:
Data-driven optimization
Uncertainty set
Robust optimization
Machine learning

ABSTRACT

The way how the uncertainties are represented by sets plays a vital role in the performance of robust optimization (RO). This paper presents a novel approach leveraging machine learning (ML) techniques to construct data-driven uncertainty sets from historical uncertainty data for RO problems. The proposed method integrates Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Gaussian Mixture Model (GMM), and Principle Component Analysis (PCA) systematically to eliminate the influence of uncertainty scenarios with low occurrence probability and generate a nonconvex uncertainty set that is a union of multiple basic subsets (box or ellipsoid) without sacrificing its computational tractability. In addition to presenting a comprehensive algorithm for uncertainty set development, this paper offers detailed guidelines for parameter tuning and performance analysis. By harnessing the well-established ML packages `scikit-learn`, a Python-based toolkit for implementing the proposed approach is also provided. Furthermore, a computationally efficient solution for a two-stage linear RO problem with the proposed data-driven uncertainty set is derived, alongside establishing a probabilistic guarantee of constraint satisfaction for out-of-sample uncertainties. Extensive numerical experiments, conducted on both synthetic and real-world datasets as well as an optimization-based control problem, are performed to demonstrate the efficacy of the proposed methodology.

1. Introduction

Optimization under uncertainties is ubiquitous in real-world engineering problems and has attracted significant research attention. There are two primary approaches – stochastic optimization and robust optimization – to enhance the robustness and reliability of deterministic optimization models amidst uncertainties [1–4]. Within the framework of stochastic optimization, the exact distributional information about uncertainties is deployed, and the expected performance towards the uncertainty distribution is optimized. In reality, however, the distributional information of uncertainties is usually not available, and obtaining this information is also a non-trivial task. One popular solution to solve the implementation difficulty of stochastic optimization is the scenario-based approach [5,6]. However, to ensure constraint satisfaction with a high confidence level, the scenario-based approach entails a large number of scenario-induced hard constraints, which leads to computational challenges. As an effective alternative, robust optimization (RO) models uncertainties via uncertainty sets without the distributional information of uncertainties and focuses on optimizing the worst-case performance [1,3,4]. Due to its effectiveness in constraint satisfaction and computational tractability, RO has gained increased popularity.

A crucial component of RO problems is the uncertainty set, which significantly influences both the computational complexity and conservatism of the corresponding RO problems. Common types of uncertain sets include box, ellipsoid, polyhedral and intersections or unions of these basic sets [3,4,7–9]. These conventional methods for constructing uncertainty sets are straightforward to implement and allow for some reduction in the conservatism of the optimal solution through careful adjustment of the set coefficients. However, the selection of these coefficients typically depends on domain-specific knowledge. Moreover, these methods are generally based on the assumption that each dimension of uncertainties is independently and asymmetrically distributed, which restricts their efficiency in handling correlations among uncertainties and scalability for high-dimensional uncertainties [10].

With the availability of abundant historical uncertainty data and the development of machine learning (ML) techniques, data-driven RO approaches have attracted increasing attention in reducing the conservatism of RO. The main idea of these approaches is to exploit ML techniques, especially unsupervised learning such as support vector clustering (SVC), kernel density estimation (KDE) and principle component decomposition, to extract the latent patterns of uncertainties, which are subsequently used for constructing uncertainty sets. It should

[*] Corresponding author.
E-mail address: y.li-39@tudelft.nl (Y. Li).

https://doi.org/10.1016/j.jprocont.2024.103339
Received 17 June 2024; Received in revised form 28 October 2024; Accepted 6 November 2024
Available online 29 November 2024
0959-1524/© 2024 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

be pointed out that many dominant ML techniques do not apply to data-driven RO problems due to the utilization of nonlinear functions, e.g., radial basis function and sigmoid function, which are widely adopted in ML algorithms but might dramatically degrade the computational tractability of the resulting RO problems with such data-driven uncertainty sets.

Recent literature explores different methods for constructing data-driven uncertainty sets in RO. In [10–12], a kernel-based support vector clustering (K-SVC) method using a novel piece-wise linear kernel is proposed to develop a non-parametric polyhedral uncertainty set. Subsequently, [13] proposed replacing the piece-wise linear kernel with a deep neural network (DNN) to construct a more compact, though non-convex, uncertainty set. However, the DNN-based sets lead to substantially longer computation times in solving the corresponding RO problems, particularly, even single-stage linear RO problems with NN-based uncertainty sets involve solving mixed-integer quadratic programs iteratively. In [14,15], the Dirichlet process mixture model is utilized to extract hidden patterns in uncertainty data through a variational inference algorithm, which involves complicated nonconvex optimization that degrades its applicability. In [16], PCA is applied to develop a polyhedral uncertainty set by decomposing the uncertainty into uncorrelated components. However, this method fails to detect low-probability uncertainties, which might result in a conservative uncertainty set. In [17–19], by combining principle components analysis (PCA) with kernel density estimation (KDE), the resultant uncertainty set is able to exclude low-probability uncertainties within the tails of the approximated probability distribution. In [20,21], PCA is combined with cutting plane methods to reduce the conservatism of the resulting uncertainty set by excluding redundant uncertainty scenarios.

Although the above PCA-based uncertainty sets are computationally efficient due to their polyhedral structure, this simple structure also sacrifices their applicability to complicated and irregular uncertainty distributions. In [22,23], the PCA-KDE-based approach proposed in [17] is further combined with KMeans clustering to construct uncertainty sets that are applicable to disjunctive uncertainties. However, as will be shown in our simulation results, although this combination offers increased flexibility for handling irregular uncertainty distributions, its performance might be degraded due to the limitation inherent to the KMeans and KDE approaches and lacks robustness against complicated datasets.

In summary, a common limitation of the above data-driven approaches is their inability to effectively balance between the complexity and the conservatism of the uncertainty set, and to adapt to complicated uncertainty distributions. Most of these methods either yield a compact uncertainty set that demands high computational resources for solving the resulting RO problems, or they produce a computationally efficient uncertainty set that, however, lacks adaptivity to complex uncertainty distributions. Besides, while some literature explored the framework of unifying multiple subsets as in this work, their approaches fail to adapt to complicated uncertainty distributions due to the improper selection and integration of ML techniques. In addition, most of the existing works only use simple synthetic datasets for performance evaluation, which might fail to truly reflect the actual practical performance when applied to complex real-world data. Furthermore, detailed guidelines as well as easy-to-use toolkits for the existing approaches are not available.

Motivated by the above discussions, this paper proposes an ML-enabled data-driven approach for a two-stage adaptive RO problem. The major contributions of this paper are summarized as follows:

- An ML-based approach is proposed to develop data-driven uncertainty sets by leveraging DBSCAN, GMM and PCA. The resulting uncertainty set is compact regardless of irregular uncertainty distributions and is computationally efficient in solving the resulting RO problems.
- Detailed guidelines for parameter tuning, performance analysis and possible limitations of applying the proposed data-driven uncertainty set are provided to facilitate its practical usage. A Python-based toolkit to implement the proposed approach is developed [24].
- The conventional affine decision rule is extended to the proposed uncertainty set by exploiting the union of multiple subsets property for a two-stage linear RO problem to give a less conservative solution. A probabilistic guarantee of constraint satisfaction for out-of-sample uncertainties is derived.
- The effectiveness of the proposed approach is extensively validated using both synthetic and real-world datasets as well as an optimization-based control problem.

The remainder of this paper is organized as follows. Section 2 introduces the ML techniques used in our work and details the algorithm for developing data-driven uncertainty sets, including guidelines for parameter tuning and performance analysis. Section 3 investigates a two-stage linear RO problem employing the proposed uncertainty set. Section 4 presents three case studies to demonstrate the efficacy of our proposed approach. Finally, conclusions are drawn in Section 5.

Notation. Boldface lowercase letters are used to denote vectors, and boldface uppercase letters denote matrices. Calligraphic uppercase letters denote sets. For a given vector/matrix, $[\cdot]_k$ refers to its kth element/row. The $\max/\min$ operators applied to vectorized objective functions imply elementwise maximization/minimization across each function. Equalities/inequalities between two vectors hold elementwise.

2. Data-driven uncertainty set construction and analysis

In this section, we will employ ML techniques to construct data-driven uncertainty sets from historical uncertainty data with the aim of reducing the conservatism of the resulting RO problem while preserving computational efficiency. Specifically, we propose an integrated framework combining Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Gaussian Mixture Model (GMM), and Principle Component Analysis (PCA) to reveal hidden patterns in uncertainty data.

2.1. DBSCAN

DBSCAN is a density-based clustering algorithm that groups data into clusters based on high-density areas, which are separated by regions of low density [25]. In contrast to centroid-based or distribution-based clustering methods, such as KMeans, which typically identify spherical or convex clusters, DBSCAN is capable of discovering clusters of arbitrary shapes. In addition, DBSCAN can detect low-probability data samples: data samples residing in low-density areas will not be assigned to any clusters. As a result, in this work, we will adopt DBSCAN to remove low-probability uncertainty samples, thereby reducing the conservatism of our proposed data-driven uncertainty set.

There are two main parameters for DBSCAN: ϵ and $MinPts$. ϵ determines the maximum distance between two samples for one to be considered as a neighbor of the other. Given a data point $\mathbf{p}$, its ϵ-Neighborhood is defined as $N_\epsilon(\mathbf{p}) = \{\mathbf{q} : \text{dist}(\mathbf{p}, \mathbf{q})\} \leq \epsilon$. $MinPts$ represents the minimum number of neighbors a sample must have to be classified as a core point. Namely, $|N_\epsilon(\mathbf{p})| \geq MinPts$ if $\mathbf{p}$ is a core point. For a core point $\mathbf{p}$, any data samples that are density-reachable from $\mathbf{p}$ will be grouped in the same cluster. The data samples that do not belong to any clusters will be classified as noise. Generally, increasing ϵ results in fewer samples being classified as outliers, whereas increasing $MinPts$ tends to identify more outliers. For further details about DBSCAN, please see [25,26].

2.2. Gaussian mixture model clustering

Gaussian mixture model (GMM) clustering, which is a distribution-based clustering approach, assumes all given data samples are generated from a mixture of a finite number of Gaussian distributions. Each distribution in the mixture is characterized by a set of parameters: the mixing probability π_k, mean $\boldsymbol{\mu}_k$ and covariance $\boldsymbol{\Sigma}_k$ of the kth model. Given n data samples $\{\mathbf{u}_1, \dots, \mathbf{u}_n\}$, these unknown parameters are estimated by maximizing the following log-likelihood function

$$\sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \pi_k \phi(\mathbf{u}_i, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right) \tag{1}$$

where K is the total number of Gaussian distributions, and $\phi(\mathbf{u}_i, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is the multivariate Gaussian density function of the kth Gaussian model. The optimal value of the parameters $(\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ can be computed via Expectation Maximization algorithm [27]. Based on the parameters $(\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, all data samples can be clustered into K groups, and the data samples in the same group are assigned with an identical label k. The label for a given data sample $\mathbf{u}_i$, denoted as z_i, is $\arg\max_k \{p(z_i = k|\mathbf{u}_i)\}$ where

$$p(z_i = k|\mathbf{u}_i) = \frac{p(z_i = k)p(\mathbf{u}_i|z_i = k)}{p(\mathbf{u}_i)} = \frac{\pi_k \phi(\mathbf{u}_i, \boldsymbol{\mu}_k, \Sigma_k)}{\sum_{k=1}^{K} \pi_k \phi(\mathbf{u}_i, \boldsymbol{\mu}_k, \Sigma_k)}.$$

In our work, the uncertainty samples will be firstly processed by DBSCAN to remove low-possibility scenarios. Then, GMM clustering will be applied to group the remaining uncertainty samples into K clusters for subsequent uncertainty subset construction via PCA.

2.3. Principal component analysis

In our work, PCA is utilized to construct box-like uncertainty subsets based on each uncertainty cluster generated by GMM. This subsection will briefly introduce the PCA approach and show how PCA is utilized in uncertainty set construction. A detailed tutorial about PCA can be found in [28].

PCA is a popular data analysis technique for dimension reduction and enhancing data interpretability. It achieves this by applying a linear transformation to the original data so that all features of the new data representation are mutually uncorrelated.

Assume that there are K uncertainty clusters $\{\mathbf{U}_1, \dots, \mathbf{U}_k\}$ after applying DBSCAN and GMM. Then, for each data cluster $\mathbf{U}_k = [\mathbf{u}_1^{\mathrm{T}}, \dots, \mathbf{u}_{n_k}^{\mathrm{T}}]^{\mathrm{T}} \in \mathbf{R}^{n_k \times m}$, where n_k is the number of uncertainty samples and m is the dimension of the uncertainty, we subtract its mean $\boldsymbol{\beta}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} \mathbf{u}_i$ from each sample in the cluster and obtain the centered training dataset $\bar{\mathbf{U}}_k$. For $\bar{\mathbf{U}}_k$, we have the following approximated covariance matrix

$$\mathbf{C}_k^{\mathbf{U}} = \frac{1}{n_k - 1} \bar{\mathbf{U}}_k^{\mathrm{T}} \bar{\mathbf{U}}_k. \tag{2}$$

By performing eigenvalue decomposition, the covariance matrix $\mathbf{C}_k^{\mathbf{U}}$ can be decomposed as

$$\mathbf{C}_k^{\mathbf{U}} = \mathbf{P}_k \boldsymbol{\Lambda}_k \mathbf{P}_k^{\mathrm{T}} \tag{3}$$

where $\mathbf{P}_k = [\mathbf{p}_1, \dots, \mathbf{p}_m] \in \mathbb{R}^{m \times m}$ is a normalized orthogonal matrix, and $\boldsymbol{\Lambda}_k = \mathrm{diag}\{\lambda_{1,k}, \dots, \lambda_{m,k}\}$ is a diagonal matrix. The columns of $\mathbf{P}_k$, denoted as $\mathbf{p}_i$, are eigenvectors of $\mathbf{C}_k^{\mathbf{U}}$, or called principal components. The diagonal entities $\lambda_{i,k}$ of $\mathbf{\Lambda}_k$ are corresponding eigenvalues of $\mathbf{C}_k^{\mathbf{U}}$. Based $\mathbf{P}_k$, we have a new matrix $\mathbf{Y}_k = [\mathbf{y}_1^{\mathrm{T}}, \dots, \mathbf{y}_{n_k}^{\mathrm{T}}]^{\mathrm{T}} := \bar{\mathbf{U}}_k \mathbf{P}_k$, which is a new representation of the uncertainty samples. $\mathbf{y}_i := \mathbf{P}_k^{\mathrm{T}}(\mathbf{u}_i - \boldsymbol{\beta}_k)$ represents the projection of the centered data sample on the principle components. For $\mathbf{Y}_k$, its covariance matrix is $\mathbf{C}_k^{\mathbf{Y}} := \frac{1}{n_k - 1} \mathbf{Y}_k^{\mathrm{T}} \mathbf{Y}_k = \mathbf{\Lambda}_k$, which implies that the components of $\mathbf{y}_i$ are uncorrelated since $\boldsymbol{\Lambda}_k$ is a diagonal matrix. The property of having uncorrelated features in the transformed dataset $\mathbf{Y}_k$ allows for the adoption of basic sets, e.g., box, to construct uncertainty subsets for each cluster $\mathbf{U}_k$ to simplify the complexity of modeling.

2.4. Uncertainty sets construction

In this subsection, we will apply the ML techniques detailed previously to construct a data-driven uncertainty set.

Algorithm 1 Data-driven uncertainty sets construction

```
Input: training dataset $\mathbf{U}_{\text{train}}$
Output: uncertainty sets $\mathcal{U}_1, \cdots, \mathcal{U}_K$
Parameters: $\epsilon$, $MinPts$, $K$
    Extreme Uncertainties Removal
1: select values of $\epsilon$ and $MinPts$ for DBSCAN
2: apply DBSCAN to $\mathbf{U}_{\text{train}}$ to remove extreme scenarios and generate
   a cleaned dataset $\mathbf{U}_{\text{clean}}$
    Uncertainty Samples Clustering
3: select the parameter $K$
4: apply GMM clustering to $\mathbf{U}_{\text{clean}}$ to generate $K$ clusters $\{\mathbf{U}_1, \cdots, \mathbf{U}_K\}$
    Box-like Subsets Construction
5: apply PCA to $\mathbf{U}_k$ to generate $\mathbf{Y}_k$, $\mathbf{P}_k$ and $\boldsymbol{\Lambda}_k$
6: construct uncertainty sets via (4) and (5)
```

Given the training set of uncertainties $\mathbf{U}_{\text{train}}$, implementing DBSCAN and GMM gives K data clusters $\{\mathbf{U}_1, \dots, \mathbf{U}_K\}$ to be processed by PCA. For each uncertainty cluster $\mathbf{U}_k$, following the PCA process introduced in *Section* 2.3 yields a new representation $\mathbf{Y}_k$ and the mean vector $\boldsymbol{\beta}_k$. Based on the relationship $\mathbf{y}_i = \mathbf{P}_k^{\mathrm{T}}(\mathbf{u}_i - \boldsymbol{\beta}_k)$, for each data cluster $\mathbf{U}_k$, we construct the following uncertainty set:

$$\mathcal{U}_k = \left\{ \mathbf{u} \left| \begin{array}{l} \mathbf{u} = \boldsymbol{\beta}_k + \mathbf{P}_k \mathbf{w}, \\ \forall \mathbf{w} : \ \underline{\mathbf{y}}_k \leq \mathbf{w} \leq \bar{\mathbf{y}}_k \end{array} \right. \right\} \tag{4}$$

with $\underline{\mathbf{y}}_k := \min\{\mathbf{Y}_k\} \in \mathbb{R}^m$ and $\bar{\mathbf{y}}_k := \max\{\mathbf{Y}_k\} \in \mathbb{R}^m$, where operators $\min$ and $\max$ are performed columnwise. Since there are K uncertainty clusters, the final uncertainty set $\mathcal{U}$ is the union of K subsets as defined in (4). By reformulating the subset in (4) as linear constraints, the uncertainty set $\mathcal{U}$ can then be compactly written as

$$\mathcal{U} := \bigcup_{k=1}^{K} \mathcal{U}_k, \quad \mathcal{U}_k := \left\{ \mathbf{u} | \mathbf{D}_k \mathbf{u} \leq \mathbf{d}_k \right\}, \tag{5a}$$

$$\mathbf{D}_k = \begin{bmatrix} \mathbf{P}_k^{\mathrm{T}} \\ -\mathbf{P}_k^{\mathrm{T}} \end{bmatrix}, \quad \mathbf{d}_k = \begin{bmatrix} \bar{\mathbf{y}}_k + \mathbf{P}_k^{\mathrm{T}} \boldsymbol{\beta}_k \\ -\underline{\mathbf{y}}_k - \mathbf{P}_k^{\mathrm{T}} \boldsymbol{\beta}_k \end{bmatrix}. \tag{5b}$$

Finally, the construction of the data-driven uncertainty set via DBSCAN, GMM and PCA can be expressed as Algorithm 1, and the corresponding graphical illustration of each component of Algorithm 1 is depicted in Fig. 1.

Remark 1. Instead of constructing box uncertainty subset via (5), another natural option of uncertainty subset based on Algorithm 1 is ellipsoid. By implementing steps 1–4 of Algorithm 1, the following ellipsoidal subset can be constructed

$$\mathcal{U}_k = \left\{ \mathbf{u} \mid \|\boldsymbol{\Sigma}_k^{-1/2}(\mathbf{u} - \boldsymbol{\mu}_k)\|_2 \leq \varrho_k \right\} \tag{6a}$$

$$\varrho_k := \max_{\mathbf{u}_i \in \mathbf{U}_k} \|\boldsymbol{\Sigma}_k^{-1/2}(\mathbf{u}_i - \boldsymbol{\mu}_k)\|_2. \tag{6b}$$

Compared with the box subset, the ellipsoidal subset might be more suitable to the Gaussian distributed uncertainty clusters. However, the robust counterpart (RC) of a linear RO problem with ellipsoidal uncertainty sets is a second-order cone program (SOCP). In contrast, the robust counterpart of a linear RO problem with a box uncertainty set remains a linear program (LP), which is supported by more off-the-shelf solvers and can be solved more efficiently, even for significantly large-scale problems, than SOCP. One advantage of the ellipsoidal uncertainty subset is that the robust counterpart problem does not introduce any extra constraints and decision variables to ensure robust constraint satisfaction. Conversely, for box uncertainty sets, additional constraints and decision variables are introduced in the RC problem, and their numbers are proportional to the number of constraints

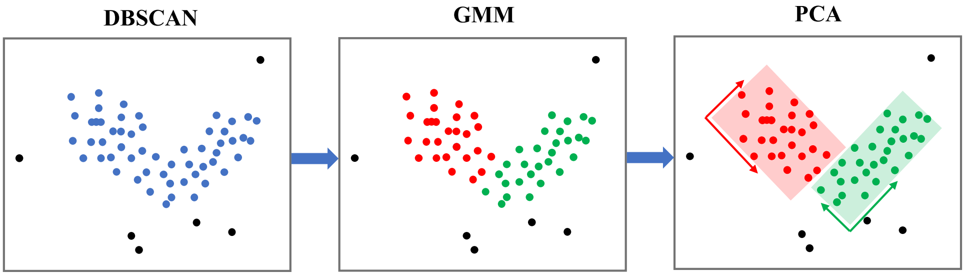


Fig. 1. Graphical illustration of Algorithm 1.

defining the uncertainty set and the dimension of the uncertainty, respectively. Therefore, the choice between box and ellipsoidal uncertainty sets should depend on the type of available numerical solvers, the dimension of the uncertainty as well as the size of the RO problem to be solved. In our work, we focus on box uncertainty subsets, but all analysis in this paper for box subsets is also applicable to ellipsoidal subsets constructed via (6).

Remark 2. The enhanced performance of the proposed data-driven uncertainty set stems from the systematic integration of ML techniques such as DBSCAN, GMM and PCA. DBSCAN allows the approach to filter out low-probability uncertainties, independent of their locations. In contrast, many existing methods, such as KDE and K-SVC, primarily exclude extreme uncertainties located at the boundary of the uncertainty cluster. Moreover, the ability of the PCA-based uncertainty subset in (5) to reduce conservatism is predicated on the assumption that the uncertainty variables are correlated and Gaussian-distributed. Unlike the centroid-based or density-based clustering approaches used in other existing literature, which fail to ensure this implicit assumption, the GMM clustering approach adopted in Algorithm 1 generates Gaussian-distributed data clusters that boost the performance of the subsequent PCA-based subsets.

Remark 3. It should be pointed out that the feature of the proposed uncertainty set representation as a union of several basic convex subsets enables a flexible and compact non-convex uncertainty set without sacrificing the computational tractability of the resulting RO problem. On the one hand, the convex subset will guarantee the computational tractability of the resulting RO problem. On the other hand, unifying several subsets increases the flexibility in dealing with irregular uncertainty distributions and reduces the conservatism of the uncertainty set. In comparison, several existing representative data-driven approaches for constructing the uncertainty set, such as PCA-KDE, K-SVC in [10,17], try to find a single convex uncertainty set to ensure computational tractability but with sacrificed compactness or a single nonconvex set, such as DNN in [13], to ensure compactness but with high computational demand. In addition, the involved ML techniques – DBSCAN, GMM and PCA – are available in many ML toolboxes, such as `scikit-learn`, and our proposed Algorithm 1 can be easily implemented. A Python-based toolkit for implementing our data-driven uncertainty set is developed in [24].

Remark 4. Since DBSCAN and GMM are used to generate data clusters for constructing uncertainty subsets, the performance of our proposed uncertainty set is directly influenced by the effectiveness of these clustering methods. This is particularly challenging with high-dimensional uncertainty data, where visual evaluation of clustering quality and the resulting uncertainty set is not feasible. Additionally, high-dimensional data can lead to sparse training data, which can further deteriorate the performance of the ML approaches involved. Moreover, in data-driven RO problems, the absence of labeled data also complicates the validation and testing of the performance of data clustering and the resulting uncertainty set. All these factors make it a complex and nontrivial task to select proper parameters/hyperparameters, such as $(\epsilon, MinPts, K)$, in the context of high-dimensional uncertainty data.

2.5. Performance analysis and uncertainty sets calibration

In this subsection, guidelines for tuning the parameters of Algorithm 1 and evaluating the resulting data-driven uncertainty set are provided. In particular, a probabilistic bound of the data coverage for out-of-sample uncertainties is introduced, which can be used to establish a probabilistic guarantee of the out-of-sample performance of the corresponding RO problem.

In Algorithm 1, there are three parameters to be selected: ϵ, $MinPts$ and K. A rule of thumb for selecting $MinPts$ is $MinPts \geq m + 1$, where m is the dimension of the uncertainty. With a fixed value of $MinPts$, ϵ determines how many uncertainty samples are excluded from the uncertainty set. A smaller ϵ results in fewer low-probability uncertainties included in the uncertainty set. Thus, ϵ can be chosen based on the desired proportion of low-probability uncertainties to be excluded from the uncertainty set. The parameter K is the number of clusters for GMM clustering, which affects both the conservatism and the complexity of the resulting data-driven uncertainty sets. While a larger K tends to yield a more compact uncertainty set, it increases the overall complexity as the total number of constraints defining the uncertainty set is proportional to K, and may bias the uncertainty set towards the training samples.

It is important to note that we assume there are no labeled data to evaluate the clustering performance, unlike typical ML data clustering tasks focusing on high clustering accuracy. Also, accurately clustered uncertainty data do not necessarily result in a favorable uncertainty set for corresponding RO problems. Since the complexity of the uncertainty set significantly influences the computational effectiveness of the corresponding RO problem, the balance between its conservatism and complexity must be considered when selecting the design parameters of Algorithm 1.

A suggested sequence of parameter selection for Algorithm 1 is as follows. Firstly, begin by setting $MinPts$ based on the rule of thumb, which is at least one plus the dimension of the uncertainty. Next, adjust ϵ to achieve the desired percentage of data coverage for the uncertainty set. Finally, determine an appropriate value of K, balancing the conservatism and complexity of the uncertainty set.

For 2-D or 3-D uncertainty data, we can rely on visual inspection to find out the proper value of design parameters. However, for high-dimensional uncertainty data, visualization of the clustering results and the corresponding uncertainty set is not applicable. As a result, we should rely on some quantitative indicators to support the selection of design parameters. In the ML community, there are several available metrics to evaluate the clustering performance, such as Silhouette Score, Calinski–Harabasz index, and Davies–Bouldin index, which are readily available in well-developed ML packages such as `scikit-learn`. In this work, we choose the Silhouette Score as an example to select the number of uncertainty subsets.

The silhouette Score, whose value belongs to $[-1, 1]$, measures how similar each data is to the cluster it belongs to and how different it is from other remaining clusters. A higher score means dense and well-separated clusters, and scores around zero indicate overlapping clusters. As a result, within an acceptable range of K, the value with a larger Silhouette Score is preferable. However, it should be noted that selecting K purely based on such metrics is not wise because the complexity and the conservatism of the uncertainty set need to be balanced.

Recall that the larger the value of K, the more complex the uncertainty set becomes, consequently reducing its conservatism. The level of conservatism of the uncertainty set can be reflected by its data coverage, typically, a less conservative set tends to have a lower coverage. For our proposed Algorithm 1, the lower bound of data coverage in the training set is determined by the parameters ϵ and $MinPts$. Therefore, a desirable K should balance a high Silhouette Score, a low training data coverage, and maintain a relatively small value to manage complexity.

Once an uncertainty set is constructed via Algorithm 1, another aspect that needs to be evaluated is its consistency in both training and validation/testing sets. Consistency means that the uncertainty sets should achieve consistent data coverage in both training and validation/testing sets so that the uncertainty set is not biased towards the training set. If the data coverage percentage of the training set is much larger than that of the testing set, it implies that the uncertainty sets are biased towards the training data and one might need to reduce the number of data clusters K to decrease the complexity of the uncertainty set.

Furthermore, based on the data coverage in the testing set, the following lemma is applicable to provide a probabilistic bound of data coverage for any I.I.D. uncertainty scenarios. The data coverage, denoted as ρ, is defined as the portion of data that is included in the uncertainty set. $\rho = 1$ means all data are covered by the uncertainty set, and $\rho = 0$ means no data is included in the uncertainty set.

Lemma 1. *For n I.I.D. samples of testing scenarios $\{\mathbf{u}_1, \dots, \mathbf{u}_n\}$, assuming that the data coverage of the uncertainty sets (5) for testing data is ρ. Then, for any random sample of uncertainty $\mathbf{u}$, the probability that $\mathbf{u} \in \mathcal{U}$, denoted as $\mathbb{P}_{u\in\mathcal{U}}$, satisfies*

$$\mathbb{P}(\mathbb{P}_{\mathbf{u}\in\mathcal{U}} \leq \rho - \tau) \leq \exp(-2n\tau^2) \tag{7}$$

with $\tau \geq 0$.

Proof. Define a random variable $\delta(u \in \mathcal{U})$ as the indicator function of the random event $\mathbf{u} \in \mathcal{U}$. Since the testing scenarios are I.I.D., $\delta(\mathbf{u}_i \in \mathcal{U})$ are n I.I.D. samples of $\delta(\mathbf{u} \in \mathcal{U})$ with $\frac{1}{n}\sum_{i=1}^{n} \delta(\mathbf{u}_i \in \mathcal{U}) = \rho$. Consequently, it follows from Hoeffding's inequality that the inequality (7) holds. □

In addition to adopting *Lemma* 1 to construct a probabilistic guarantee of uncertainty coverage for the proposed uncertainty set, the calibration approach using order statistics calculation introduced in [9] is also applicable to our proposed data-driven uncertainty set to develop probabilistic guarantees of uncertainty coverage.

3. Robust optimization design and performance guarantees

In this section, we consider a linear two-stage RO problem with the proposed data-driven uncertainty set. The RO problem investigated in this work has the following structure:

$$\min_{\mathbf{x}} \mathbf{c}^{\mathrm{T}}\mathbf{x} + \max_{\mathbf{u}\in\mathcal{U}} \min_{\mathbf{z}} \mathbf{b}^{\mathrm{T}}\mathbf{z} \tag{8a}$$

$$\text{s.t. } \mathbf{T}\mathbf{x} + \mathbf{W}\mathbf{z} + \mathbf{M}\mathbf{u} \leq \mathbf{h}, \tag{8b}$$

$$\mathbf{u} \in \mathcal{U} \tag{8c}$$

where $\mathbf{x} \in \mathbb{R}^s$ is the first-stage decision variables, $\mathbf{z} \in \mathbb{R}^q$ is the second-stage decision variables, or called recourse decision variables, $\mathbf{u} \in \mathbb{R}^m$ is the uncertainty. The uncertainty set $\mathcal{U}$ is assumed to be constructed via our proposed approach, and can be written in the following format

$$\mathcal{U} := \bigcup_{k=1}^{K} \mathcal{U}_k, \quad \mathcal{U}_k := \{\mathbf{u} | \mathbf{D}_k\mathbf{u} \leq \mathbf{d}_k\}. \tag{9}$$

The row number of $(\mathbf{T}, \mathbf{W}, \mathbf{M})$ is denoted as r, $(\mathbf{c}, \mathbf{b}, \mathbf{h})$ and $(\mathbf{T}, \mathbf{W}, \mathbf{M})$ are parameter vectors and matrices with appropriate dimensions, respectively.

Optimization problem (8) is semi-infinite and is computationally intractable since there are infinitely many constraints to be satisfied. To achieve a balance between computational burden and optimality, we adopt the following affine decision rule for the recourse decision variable $\mathbf{z}$:

$$\mathbf{z} = \mathbf{L}_k\mathbf{u} + \mathbf{g}_k, \ \forall \mathbf{u} \in \mathcal{U}_k \tag{10}$$

where $\mathbf{L}_k \in \mathbb{R}^{q\times m}$ and $\mathbf{g}_k \in \mathbb{R}^q$ are decision variables to be optimized. Unlike the conventional decision strategy where only a single decision rule is imposed for all possible uncertainties in the uncertainty set, such as the works in [12], the proposed uncertainty set makes it possible to assign different decision policies (10) for uncertainties residing in different sets. By exploiting the feature of the proposed uncertainty set for unifying several subsets, it is possible to reduce the conservatism of the optimal RO solution.

Theorem 1. *For the two-stage robust optimization problem (8), assuming that the uncertainty set is defined as (9), and the recourse decision variables are determined via (10), the optimal solution can be computed by solving (11).*

$$\min_{\substack{\mathbf{x},\eta,\mathbf{L}_k,\mathbf{g}_k\\ \boldsymbol{\pi}_{0,k},\boldsymbol{\pi}_{i,k}}} \ \mathbf{c}^{\mathrm{T}}\mathbf{x} + \eta \tag{11a}$$

$$s.t. \ \mathbf{b}^{\mathrm{T}}\mathbf{g}_k + \boldsymbol{\pi}_{0,k}^{\mathrm{T}}\mathbf{d}_k \leq \eta, \tag{11b}$$

$$\mathbf{D}_k^{\mathrm{T}}\boldsymbol{\pi}_{0,k} = \mathbf{L}_k^{\mathrm{T}}\mathbf{b}, \tag{11c}$$

$$[\mathbf{T}\mathbf{x} + \mathbf{W}\mathbf{g}_k - \mathbf{h}]_i + \boldsymbol{\pi}_{i,k}^{T}\mathbf{d}_k \leq 0, \tag{11d}$$

$$\mathbf{D}_k^{\mathrm{T}}\boldsymbol{\pi}_{i,k} = [\mathbf{W}\mathbf{L}_k + \mathbf{W}]_i, \tag{11e}$$

$$\boldsymbol{\pi}_{0,k} \geq 0, \ \boldsymbol{\pi}_{i,k} \geq 0, \tag{11f}$$

$$\forall k = 1, \dots, K, \quad \forall i = 1, \dots, r. \tag{11g}$$

In addition, if the data coverage for n I.I.D. testing uncertainty samples is ρ, the probability that the solution of (11) can ensure constraint satisfaction for a random uncertainty $\mathbf{u}$ is greater than $\rho - \tau$ with confidence at least $1 - \exp(-2n\tau^2)$.

Proof. Substituting the decision policy (10) into (8) and considering uncertainty sets (9) leads to

$$\min_{\mathbf{x},\mathbf{L}_k,\mathbf{g}_k} \left\{ \mathbf{c}^{\mathrm{T}}\mathbf{x} + \max_{1\leq k\leq K} \max_{\mathbf{u}\in\mathcal{U}_k} \mathbf{b}^{\mathrm{T}}(\mathbf{L}_k\mathbf{u} + \mathbf{g}_k) \right\} \tag{12a}$$

$$\text{s.t. } \mathbf{T}\mathbf{x} + \mathbf{W}\mathbf{g}_k + (\mathbf{W}\mathbf{L}_k + \mathbf{M})\mathbf{u} \leq \mathbf{h}, \tag{12b}$$

$$\forall \mathbf{u} \in \mathcal{U}_k, \ \forall k = 1, \dots, K. \tag{12c}$$

The universal constraint satisfaction in (12) can be alternatively reformulated as the following worst-case constraint satisfaction

$$\min_{\mathbf{x},\mathbf{L}_k,\mathbf{g}_k} \ \mathbf{c}^{\mathrm{T}}\mathbf{x} + \eta \tag{13a}$$

$$\text{s.t. } \max_{\mathbf{u}\in\mathcal{U}_k} \left\{\mathbf{b}^{\mathrm{T}}\mathbf{L}_k\mathbf{u}\right\} + \mathbf{b}^{\mathrm{T}}\mathbf{g}_k \leq \eta, \tag{13b}$$

$$\mathbf{T}\mathbf{x} + \mathbf{W}\mathbf{g}_k + \max_{\mathbf{u}\in\mathcal{U}_k} \left\{(\mathbf{W}\mathbf{L}_k + \mathbf{M})\mathbf{u}\right\} \leq \mathbf{h}, \tag{13c}$$

$$\forall k = 1, \dots, K. \tag{13d}$$

For the worst-case constraint satisfaction in (13b), following a standard procedure in RO [4], the maximization problem $\max_{\mathbf{u}\in\mathcal{U}_k}\{\mathbf{b}^{\mathrm{T}}\mathbf{L}_k\mathbf{u}\}$ in Eq. (13b) can be translated into its dual problem:

Table 1
Approaches considered in the case studies.

Approach	Description
Box	Box uncertainty set constructed based on the minimal and maximal value of each data dimension.
CH	Convex hull of uncertainty data introduced in [16].
K-SVC	Kernel-based support vector clustering method proposed in [10,11].
KPKDE	KMeans-based clustering combined with PCA and KDE approach proposed in [23].
B-DGP	Our proposed Algorithm 1 using box subsets in (5).
E-DG	One variant of Algorithm 1 where ellipsoidal uncertainty subsets are constructed via (6).

$$\min_{\boldsymbol{\pi}_{0,k}} \mathbf{d}_k^{\mathrm{T}} \boldsymbol{\pi}_{0,k} \tag{14a}$$

$$\text{s.t. } \mathbf{D}_k^{\mathrm{T}} \boldsymbol{\pi}_{0,k} = \mathbf{L}_k^{\mathrm{T}} \mathbf{b}, \quad \boldsymbol{\pi}_{0,k} \geq 0 \tag{14b}$$

where $\boldsymbol{\pi}_{0,k}$ is the Lagrangian multiplier. Consequently, constraint (13b) can be relaxed as

$$\mathbf{d}_k^{\mathrm{T}} \boldsymbol{\pi}_{0,k} + \mathbf{b}^{\mathrm{T}} \mathbf{g}_k \leq \eta,$$
$$\mathbf{D}_k^{\mathrm{T}} \boldsymbol{\pi}_{0,k} = \mathbf{L}_k^{\mathrm{T}} \mathbf{b}, \quad \boldsymbol{\pi}_{0,k} \geq 0.$$

Similarly, constraint (13c) results in constraints (11d)–(11f).

It is clear that any feasible solution of (11) can guarantee constraint satisfaction for any uncertainty as long as $\mathbf{u} \in \mathcal{U}$. Based on the I.I.D. assumption of the uncertainty and the data coverage in the testing set, it follows from Lemma 1 that $\mathbb{P}(\mathbb{P}_{\mathbf{u}\in\mathcal{U}} \leq \rho-\tau) \leq \exp(-2n\tau^2)$, which implies that the probability that the solution of (11) guarantees constraint satisfaction for any randomly generated uncertainty $\mathbf{u}$ is larger than $\rho - \tau$ with confidence of at least $1 - \exp(-2n\tau^2)$. This completes the proof. □

Remark 5. Theorem 1 presents a computationally efficient approximation of (8) by imposing affine decision policy for the recourse decision variable and solving a linear program. While the uncertainty set in (5) is nonconvex, the feature that the uncertainty set is a union of several basic convex subsets ensures the computational efficiency of the corresponding RO problem. In addition, since separate decision rules are applied for each subset, the optimal solution is expected to be less conservative than the typical RO solution with a single uncertainty set and a single decision rule, such as [10–12,18]. In addition, beyond the conventional RO solutions, a probabilistic guarantee of constraint satisfaction for out-of-sample uncertainties is provided based on the performance testing of the uncertainty set.

4. Applications

In this section, the performance of our proposed method is compared with several representative data-driven RO approaches. Specifically, we focus on the following approaches listed in Table 1.

The performance of the above approaches is assessed through three case studies. For the first two case studies, the performance of the uncertainty sets is compared in the following aspects: area of uncertainty set, complexity, and computation time. The complexity of polyhedral uncertainty sets in our work refers to the number of linear constraints defining the uncertainty set. This metric, which is generally neglected in the existing literature, is crucial in influencing the computational efficiency of the corresponding RO problem. For a linear RO problem with a polyhedral uncertainty set, the number of decision variables and constraints in its robust counterpart problem is proportional to the number of constraints defining the uncertainty sets. As a result, a high complexity of the uncertainty set leads to increased computational demands for solving the corresponding RO problem. For the last case study, the performance of the proposed approach is assessed via an optimal building climate control problem.

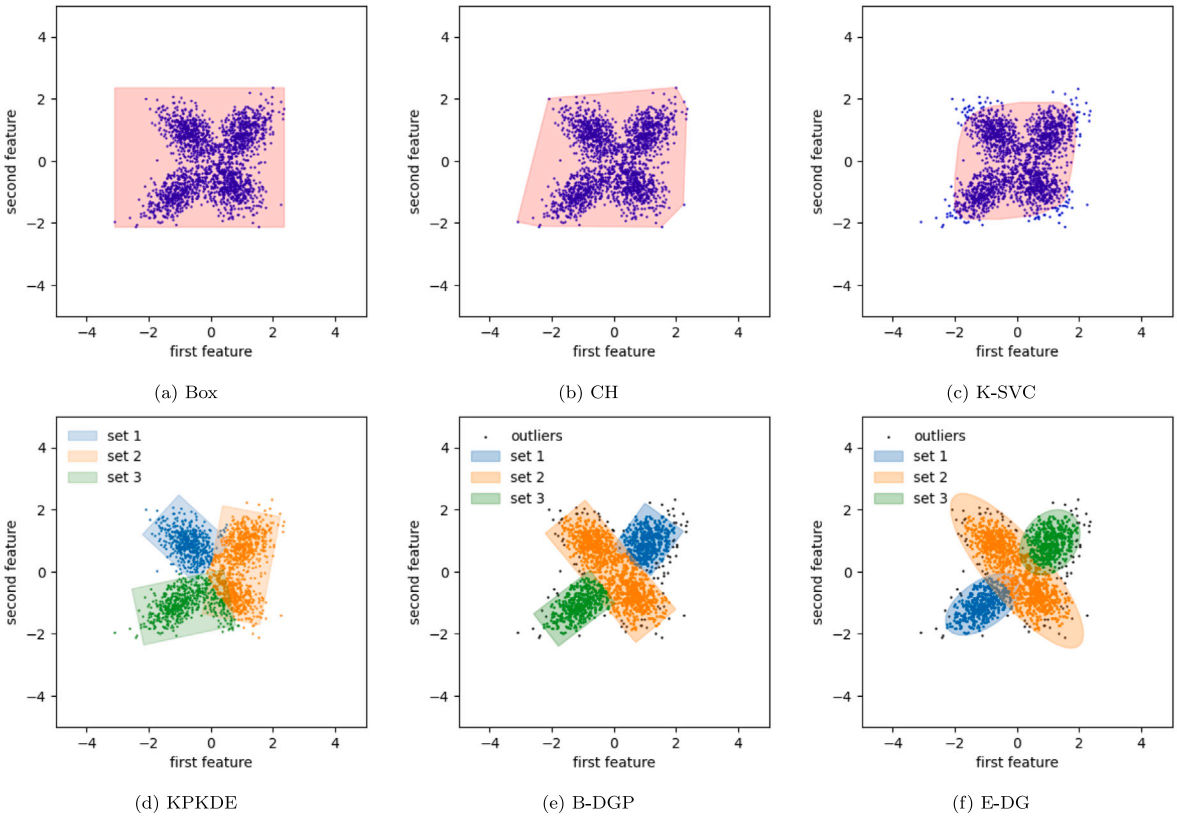
In our upcoming case studies, since K-SVC, KPKDE, B-DGP and E-DG can exclude some uncertainty samples to reduce the conservatism of the corresponding uncertainty sets, the design parameters of these approaches are chosen to exclude about 5% extreme uncertainty scenarios to reduce the conservatism of the resulting data-driven uncertainty sets. Namely, for the K-SVC approach, the parameter ν is set as 0.05; for the KPKDE approach, the confidence level of each feature is selected as $[(1-0.95^{\frac{1}{m}})/2, (1+0.95^{\frac{1}{m}})/2]$; for the B-DGP and E-DG approaches, the parameter ϵ is adjusted to identify about 5% of the total uncertainty samples as outliers. With the same data coverage, the approach with a smaller set size indicates a less conservative uncertainty set.

All simulations are implemented on an Intel Xeon W-2223 CPU at 3.6 GHz with 16G RAM. Optimization problems are modeled using Python package `gurobipy` and solved via `Gurobi 11.0`. The involved ML methods – KDE, DBSCAN, GMM and PCA – are implemented via the Python package `scikit-learn 1.0.2`.

4.1. *Case Study 1: synthetic uncertainty data*

In this case study, we test the performance of the data-driven approaches listed in Table 1 for constructing uncertainty sets with synthetic data. A common assumption about uncertainties is that they are Gaussian distributed. In reality, due to different working conditions simultaneously being represented in the dataset, the uncertainties may follow different Gaussian mixture distributions. In order to reflect this, we use synthetic uncertainty data that are generated from 4 different two-dimensional Gaussian distributions with 500 scenarios per distribution.

The uncertainty sets with different data-driven approaches are presented in Fig. 2, where the uncertainty sets are the shaded regions. The performance of each data-driven approach is summarized in Table 2. It can be observed that among all approaches, the Box approach gives the most conservative uncertainty set since it does not extract and exploit the hidden patterns of the uncertainty data. For the CH approach, it is non-parametric and does not entail any computation to construct the uncertainty set, see [16] for more details, so that its computation time is not indicated. However, its non-parametric nature also leads to a very high complexity. Similarly, while the K-SVC approach can find a compact uncertainty set by precluding some extreme scenarios, its non-parametric nature also incurs a high complexity. In addition, the convex nature of the K-SVC-induced uncertainty set limits its flexibility in dealing with general nonconvex uncertainty distributions. As indicated in [10], the number of the support vectors, which determines the complexity of the uncertainty set developed by the K-SVC approach, is inversely proportional to the conservatism of the resultant uncertainty set. Hence, with the K-SVC approach, a less conservative uncertainty set implies an uncertainty set with higher complexity. Furthermore, Table 2 shows that the K-SVC approach takes much longer computation time than the others since it has to solve a large-scale quadratic programming problem. In the KPKDE approach, KMeans-based clustering fails to produce suitable uncertainty clusters for subsequent PCA and KDE processes, resulting in a relatively conservative uncertainty set. This issue arises because the KMeans algorithm groups data based solely on distances, without considering the data distribution within the same cluster, whereas PCA implicitly assumes a Gaussian data distribution. This discrepancy limits the efficiency of the KPKDE approach. In contrast, our proposed methods B-DGP and E-DG, which combine GMM and PCA, generate more appropriate data clusters and thus yield more compact uncertainty sets. Additionally, the utilization of DBSCAN ensures our proposed uncertainty sets are immune to the influence of unlikely uncertainties. As shown in Table 2, the B-DGP approach gives a less conservative uncertainty set compared to the E-DG approach. This is because, in the E-DG approach, while the ellipsoid subset is more suitable for Gaussian distributed data, its symmetric structure and correlated features may also increase its conservatism to some extent.



(a) Box (b) CH (c) K-SVC

(d) KPKDE (e) B-DGP (f) E-DG

Fig. 2. Data-driven uncertainty sets using synthetic data in Case Study 1.

Table 2
Performance summary of different uncertainty sets in Case Study 1.

	Existing approaches				Our approaches	
	Box	CH	K-SVC	KPKDE	B-DGP	E-DG
Uncertainty set size	24.40	20.66	13.10	14.77	12.74	14.26
Complexity (# of linear constraints)	4	4001	421	12	12	–
Data coverage	1	1	0.95	0.98	0.97	0.98
Computation time (s)	< 0.01	–	59.27	0.42	0.62	1.10

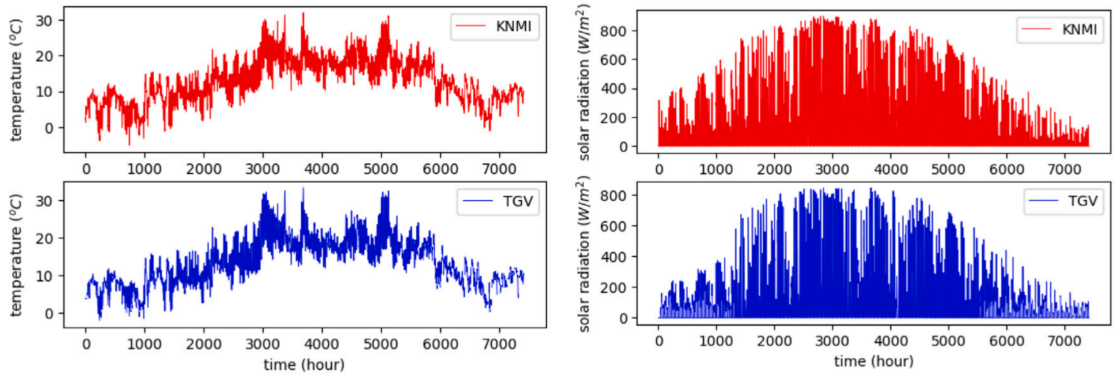
4.2. Case Study 2: real-world weather data

In this case study, the approaches listed in Table 1 are evaluated using real-world weather data. The issue of weather uncertainties is common in many applications, such as building climate control, renewable energy management, greenhouse control, etc. How to properly construct the uncertainty set of weather conditions is important in these application problems.

This case study considers the uncertainties in ambient temperature and solar radiation. The utilized weather data are measured during Jan. 2023–Dec. 2023 from two weather stations in the Netherlands: (1) the weather station 344 of Koninklijk Nederlands Meteorologisch Instituut (KNMI), a scientific institute of the Dutch government, located in Rotterdam, and (2) the weather station located in The Green Village (TGV) at TU Delft, which is about 7 km from the KNMI station. The profiles of corresponding weather data are shown in Fig. 3. The weather data from KNMI are utilized as predicted weather conditions, and the data from TGV are used as real local conditions. There are 7416 data points in total. For the sake of visualization, the weather uncertainties are scaled so that the maximum absolute value of each feature is 1, and the scaled data are used for developing uncertainty sets.

The uncertainty sets with different data-driven approaches are shown in Fig. 4 and Table 3. From Fig. 4, it is clear that the Box approach is the most conservative since no latent feature of the data is utilized. The uncertainty set via the CH approach is also very conservative due to the inclusion of low-probability uncertainty samples. For the K-SVC approach, while the size of the uncertainty set is small, it misses some high-probability scenarios located near the boundary of the uncertainty distribution for the sake of a convex uncertainty set. Also, as shown in Table 3, the K-SVC approach takes much longer computation time than the others and incurs a high complexity. While the KPKDE approach can exclude low-probability uncertainties to reduce the conservatism of the uncertainty set, its performance for this irregular uncertainty distribution is far from satisfactory. On the one hand, the KMeans method fails to generate suitable data clusters for subsequent PCA-KDE-based uncertainty set construction. On the other hand, the adoption of KDE fails to remove low-probability uncertainty samples, especially for the uncertainty cluster 5 in Fig. 3(d) where uncertainty samples are sparsely distributed. Finally, it can be observed from Fig. 4(e) and 4(f) as well as Table 3 that our proposed B-DGP and E-DG approaches are more versatile than the others in dealing with irregular uncertainty distributions and achieve a notable balance between complexity and conservatism in constructing uncertainty sets.

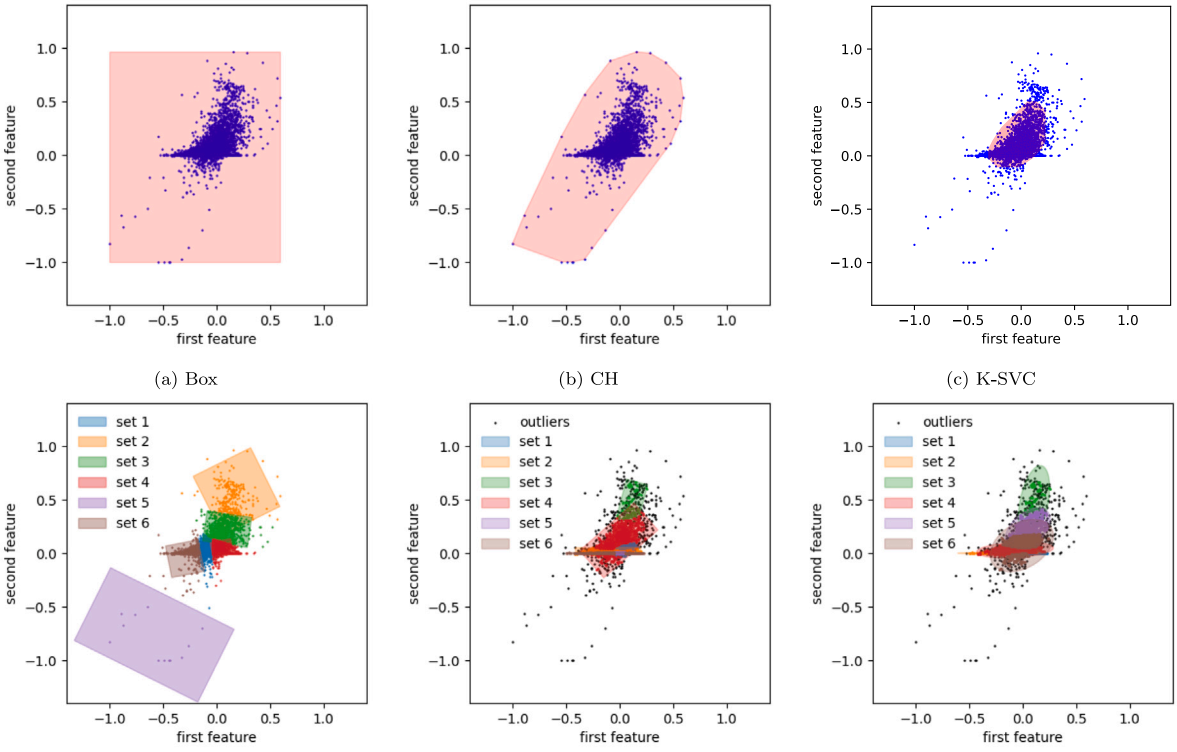
Remark 6. It should be pointed out that, based on the results of our case studies, the complexity of the uncertainty set constructed via K-SVC can be remarkably influenced by numerical errors. The K-SVC method relies on support vectors to construct the uncertainty set. Based on the value of corresponding Lagrangian multipliers α_i, all uncertainty samples are classified into 3 categories: support vectors with $0 < \alpha_i < \frac{1}{n\nu}$, outliers with $\alpha_i = \frac{1}{n\nu}$ and interior points with $\alpha_i = 0$. Due to the existence of numerical errors, all uncertainty samples



(a) Ambient temperature profiles.

(b) Solar radiation profiles.

Fig. 3. Weather data from KNMI and TGV in Case Study 2.



(a) Box

(b) CH

(c) K-SVC

(d) KPKDE

(e) B-DGP

(f) E-DG

Fig. 4. Data-driven uncertainty sets using real weather data in Case Study 2.

Table 3
Performance summary of different uncertainty sets in Case Study 2.

	Existing approaches				Our approaches	
	Box	CH	K-SVC	KPKDE	B-DGP	E-DG
Uncertainty set size	3.12	1.66	0.25	1.56	0.28	0.34
Complexity (# of linear constraints)	4	14 833	1493	24	24	–
Data coverage	1	1	0.95	0.98	0.98	0.98
Computation time (s)	< 0.01	–	1745.85	2.72	2.95	3.44

may be identified as support vectors, which will dramatically increase the complexity of the resulting RO problem due to the non-parametric property of the uncertainty set. To mitigate the influence of numerical errors, in our case studies, the support vectors are selected based on $0 + \epsilon < \alpha_i < \frac{1}{n\nu} - \epsilon$, where $\epsilon > 0$ is a sufficiently small constant ($\epsilon = 1 \times 10^{-8}$ in our case studies). The choice of ϵ might strongly influence the number of support vectors and hence the complexity of the uncertainty set. In addition, as shown in [10], the number of the support vectors is proportional to the parameter ν, which implies a less conservative uncertainty set derived via the K-SVC approach has more complexity, and consequently higher computational burden for solving the corresponding RO problem. Furthermore, from Table 2 and 3, it can be observed that the computation time of K-SVC is much larger than the others since it entails solving a large-scale QP.

Remark 7. With the KPKDE approach, uncertainty samples are first clustered using KMeans. Then, the samples within each cluster are processed using the PCA method. Subsequently, an uncertainty subset is constructed based on the confidence intervals derived from the approximated cumulative density function (CDF) of the PCA-processed data, tailored to a predefined confidence level. As can be seen in Fig. 2 and 4, the KMeans clustering used in the KPKDE approach can result in unsuitable data clusters for constructing PCA-KDE-based uncertainty sets because PCA implicitly assumes Gaussian distributed data whereas KMeans fails to generate data clusters satisfying this assumption. Furthermore, the efficacy of the KPKDE method is sensitive to the choice of the hyperparameters: the kernel functions and associated bandwidth. Assuming an ideal approximation of the CDF, an uncertainty set formulated with a confidence level γ for each dimension of the PCA-processed data would give a data coverage γ^m, where m represents the uncertainty dimension. Nevertheless, the presence of an approximation error in the CDF, quantified as ϵ, might lead to an actual confidence level $\gamma+\epsilon$. Consequently, the actual data coverage is $(\gamma+\epsilon)^m$. Given values of $\gamma = 0.98$, $\epsilon = 0.01$ and $m = 10$, the total error of data coverage is 0.087, which means 8.7% of the uncertainty samples are unexpectedly included in the uncertainty set.

Remark 8. The importance of using real-world datasets over simple synthetic datasets for performance evaluation lies in two main factors. Firstly, the assumption made when generating synthetic datasets may not accurately represent real-world conditions when applying the approach to practical problems. For instance, a common assumption in building climate control is that the weather uncertainties follow Gaussian distributions. However, as illustrated in Fig. 4, our real-world data is clearly non-Gaussian and is more complex to handle. Second, real-world data distributions can be far more intricate and irregular than synthetic ones. Consequently, data-driven approaches that perform well on synthetic datasets may experience significant performance degradation when applied to complex, real-world data. Therefore, conclusions drawn from synthetic datasets may not generalize to actual practical scenarios. Additional supporting materials, demonstrating that the performance of the data-driven approaches considered in this paper are harder to distinguish from each other for simple synthetic datasets, are provided in [24] to further validate the aforementioned statement.

4.3. *Case Study 3: optimization-based building energy control*

In this case study, we will investigate the RO design for the optimization-based building energy control problem considered in [11]. The thermal dynamics of buildings are subject to several weather uncertainties, e.g., ambient temperature uncertainty. Properly considering these uncertainties in building energy control can improve occupants' comfort and reduce energy usage.

The building thermal dynamics is modeled as the following linear system

$$x_{t+1} = Ax_t + B_u u_t + B_v v_t + B_w w_t, \tag{16}$$

where x_t is the state vector consisting of indoor air temperature, wall temperature, roof temperature, and floor temperature; u_t is the heating power injection, v_t is the vector of the predicted value of ambient temperature and underground temperature, w_t is the prediction error of ambient temperature. The system matrices are

$$A = \begin{bmatrix} 0.0167 & 0.0048 & 0.1245 & 0.409 \\ 0.0005 & 0.0002 & 0.0039 & 0.0044 \\ 0.0253 & 0.0073 & 0.3321 & 0.0617 \\ 0.0244 & 0.0070 & 0.0526 & 0.3456 \end{bmatrix},$$

$$B_u = \begin{bmatrix} 0.0986 \\ 0.0029 \\ 0.0288 \\ 0.0275 \end{bmatrix}, B_v = \begin{bmatrix} 0.2536 & 0.4596 \\ 0.0070 & 0.9840 \\ 0.4450 & 0.1287 \\ 0.4477 & 0.1225 \end{bmatrix}, B_w = \begin{bmatrix} 0.2536 \\ 0.0070 \\ 0.4450 \\ 0.4477 \end{bmatrix}.$$

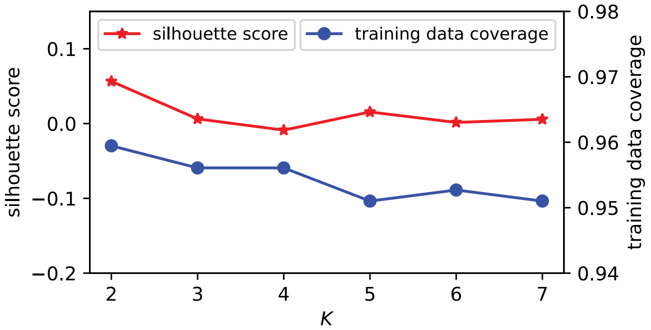


Fig. 5. Silhouette score and data coverage percentage for different K values (# of clusters) in Case Study 3.

The control objective is to determine the heating power injection u_t over a finite time window to ensure indoor comfort constraints while reducing energy consumption in the presence of weather prediction errors. As a result, the finite horizon optimal control problem can be formulated as

$$\min_{u_t} \sum_{t=0}^{H} l_t(x_t, u_t) \tag{17a}$$

$$\text{s.t. } x_{t+1} = Ax_t + B_u u_t + B_v v_t + B_w w_t, \tag{17b}$$

$$x_t \in \mathcal{X},\ u_t \in \mathcal{U}, \tag{17c}$$

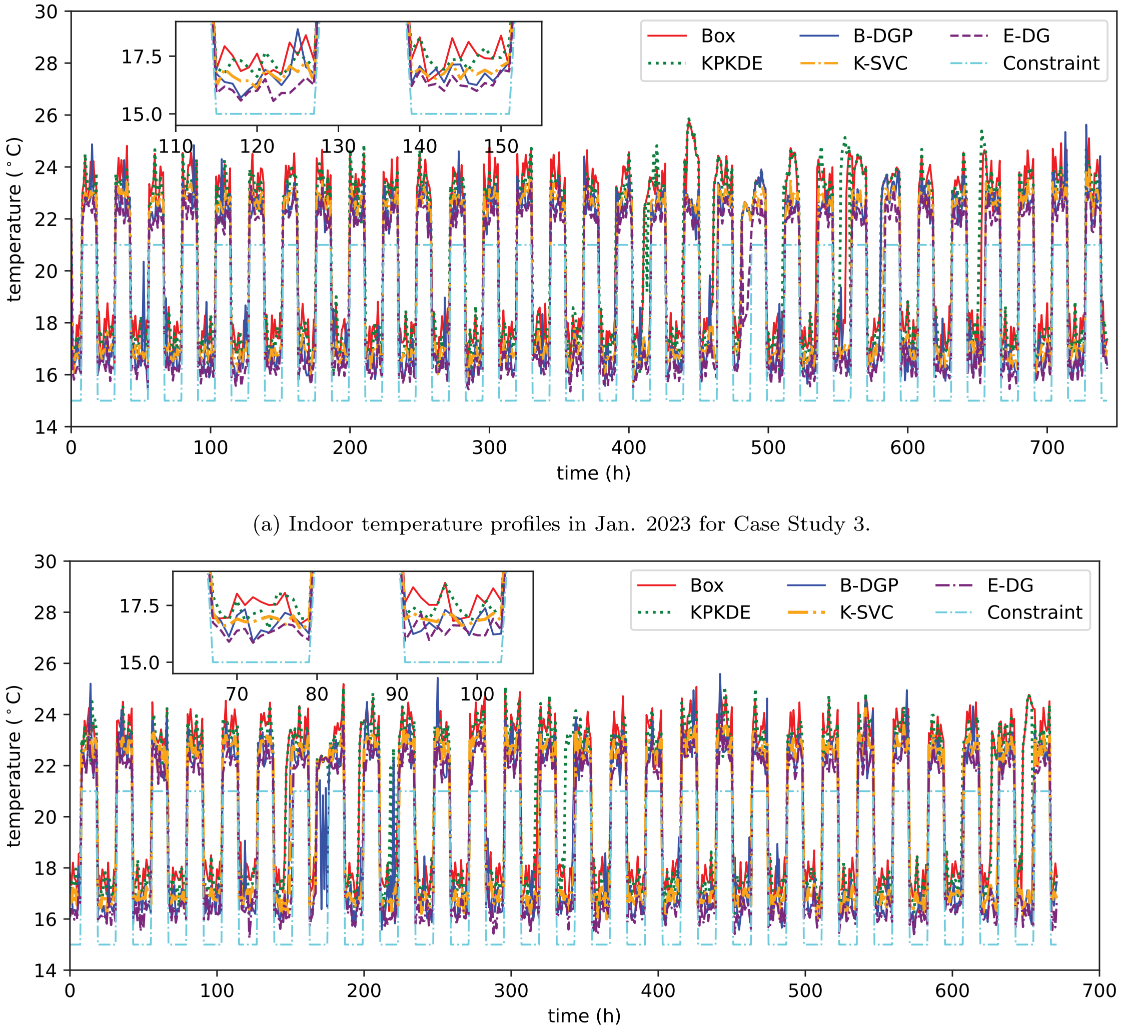
$$\forall t = 0, 1, \ldots, H-1,\ \forall w_t \in \mathcal{W} \tag{17d}$$

where l_t is the stage cost function; $\mathcal{X}$ and $\mathcal{U}$ are the feasible sets of the states and thermal input, respectively; $\mathcal{W}$ is the uncertainty set of the ambient temperature, which are constructed via the data-driven approaches in Table 1; H is the length of the prediction horizon.

Similarly to [11], the feasible set of states is defined as keeping the indoor temperature above 21 °C during peak occupied hours 7 am–6 pm, and keeping the indoor temperature above 15 °C for off-peak hours to avoid unnecessary energy usage. The length of the prediction horizon is $H = 10$. The admissible control input set is defined as $0 \leq u_t \leq 150$. The stage cost function is $l_t := u_t$ for minimizing energy consumption. The above finite horizon optimal control problem can be formulated as an RO problem in the format of (8). The thermal control input u_t can be regarded as the recourse decision variable, and the prediction errors of ambient temperature are uncertainties. While only one uncertainty factor – ambient temperature – is considered, the dimension of the uncertainty for the resulting RO problem is 10, which is equal to the length of the prediction horizon H, since all uncertainties within the prediction horizon need to be considered. To ensure the causality of input, decision variable $\mathbf{L}_k$ is restricted to be strictly lower triangular. The control inputs within the prediction horizon are applied in an open-loop manner. Namely, the control signals are computed via the affine control policy (10) with the parameters $(\mathbf{L}_k, \mathbf{g}_k)$ updated every H time step.

The historical weather data used in this case is from KNMI in *Case Study 2*. 80% of the uncertainty samples are used for developing uncertainty sets, and the remaining 20% uncertainty data are used for implementing simulations of the robust optimization-based control design. The data of ambient temperature in Jan. and Feb., during which heating is needed, are used for simulations. The ground temperature is set as the annual average ambient temperature.

In this case study, we mainly focus on the following data-driven approaches: Box, K-SVC, KPKDE, B-DGP and E-DG. Unlike the previous case studies, this case study considers high-dimensional uncertainties. Hence evaluating the uncertainty sets visually by plotting the uncertainty sets is not suitable anymore. As introduced in Section 2.5, we rely on the silhouette score and data coverage measures to find a suitable value of K for Algorithm 1, namely the number of subsets. To



(a) Indoor temperature profiles in Jan. 2023 for Case Study 3.

(b) Indoor temperature profiles in Feb. 2023 for Case Study 3.

Fig. 6. Indoor temperature profiles with different data-driven uncertainty sets in Case Study 3.

ensure the computational efficiency of the resulting RO problem, we restrict $K \in [2, 7]$. The silhouette score and the corresponding training data coverage of the B-DGP-based uncertainty set with different K are plotted in Fig. 5. It can be observed that $K = 5$ achieves a balance between high silhouette score and low data coverage. Hence, we select $K = 5$. For the KPKDE approach, the number of data clusters is selected as $K = 2$, which is tuned based on the suggestion provided in [22] to yield the highest Calinski–Harabasz index. To demonstrate the effectiveness of the proposed decision rule (10), the conventional decision rule where only a single decision is applied for all subsets within the proposed uncertainty set is also implemented.

The indoor temperature profiles with different data-driven uncertainty sets and our proposed decision rule are depicted in Fig. 6, and the corresponding control performance is summarized in Table 4. The conservatism of the uncertainty set and the solution quality of the corresponding RO problem are measured by the value of the average cost. Among all approaches, it can be observed from Table 4 that the K-SVC approach and our proposed approaches (B-DGP and E-DG) have relatively low average cost since they can exclude some low-probability samples. In comparison, the KPKDE results in a much higher average cost (the second highest among all approaches), and consequently a more conservative uncertainty set. While theoretically, the KPKDE approach is also capable of removing low-probability uncertainty samples, its performance is far from satisfactory for this complicated real-world uncertainty distribution. As for the Box approach, it results in the highest average cost since it does not leverage any underlying pattern of the uncertainty data. For the K-SVC approach, while it achieves comparable average cost to our proposed B-DGP approach, its computational time is 4 times greater than that of the B-DGP approach. This is caused by the high complexity of the uncertainty set due to its non-parametric property. For our proposed B-DGP and E-DG approaches, both have relatively low average cost and short computation time, which means that a favorable balance between the complexity and conservatism of the uncertainty sets is achieved. Notably, the E-DG approach has the lowest average cost compared to all other approaches. It should be pointed out that the E-DG approach requires a SOCP solver, whereas B-DGP only needs an LP solver for solving (17).

The last two columns of Table 4 show the simulation results for our proposed uncertainty sets when a conventional decision rule is applied that implements a single decision rule for all subsets. It can be seen that indoor comfort constraints are violated with the conventional decision rule while our proposed decision rule keeps the indoor temperature within the admissible range during the whole simulation period, which confirms that the proposed decision rule (10) is beneficial for improving the robustness of the optimal solution.

Table 4
Control performance summary of different data-driven uncertainty sets in Case Study 3.

	Existing approaches			Our proposed approaches			
	Box	K-SVC	KPKDE	B-DGP	E-DG	B-DGP	E-DG
				Proposed decision rule		Conventional decision rule	
Complexity (# of linear constraints)	20	721	40	100	–	100	–
Data coverage	1	0.95	0.97	0.95	0.95	0.95	0.95
Average cost (W/m^2)	99.62	92.21	98.87	91.29	87.50	86.59	86.30
Total constraint violation (°C)	0	0	0	0	0	14.65	6.07
Average computation time for solving (17) (s)	0.05	0.94	0.10	0.23	0.29	0.05	0.05

5. Conclusions

This study proposes a novel approach to construct data-driven uncertainty sets, leveraging DBSCAN, GMM, and PCA techniques, for robust decision-making with uncertainties. The proposed approach is flexible in balancing the conservatism and complexity of the uncertainty set while demonstrating resilience to low-probability uncertainty scenarios regardless of the underlying uncertainty distributions. The influence of each design parameter is elucidated, and the performance of the proposed uncertainty set can be systematically analyzed with the guidelines provided in this paper. By adopting well-established ML packages `scikit-learn`, a Python-based toolkit for conveniently and efficiently implementing our proposed data-driven uncertainty set is provided. Furthermore, for a linear two-stage RO problem, a tailored solution with the proposed uncertainty set is derived with a probabilistic guarantee of constraint satisfaction for out-of-sample uncertainties, enhancing the confidence of applicability over conventional RO solutions. Comparative analyses with several existing uncertainty set construction methods highlight the superiority of our methodology in striking a balance between computational efficiency and robustness of hedging against uncertainties.

Future works include designing novel RO formulations to exploit the representation of the proposed data-driven uncertainty set – a union of basic subsets – to further reduce the conservatism typically associated with conventional RO formulations.

CRediT authorship contribution statement

Yun Li: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Neil Yorke-Smith:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization. **Tamas Keviczky:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The work was supported by the Brains4Buildings project under the Dutch grant programme for Mission-Driven Research, Development and Innovation (MOOI). The authors thank the anonymous reviewers for their suggestions.

Data availability

Data will be made available on request.

References

[1] A. Ben-Tal, L. El Ghaoui, A. Nemirovski, Robust optimization, vol. 28, Princeton University Press, 2009.
[2] J.R. Birge, F. Louveaux, Introduction to Stochastic Programming, Springer Science & Business Media, 2011.
[3] D. Bertsimas, D.B. Brown, C. Caramanis, Theory and applications of robust optimization, SIAM Rev. 53 (3) (2011) 464–501.
[4] D. Bertsimas, D.D. Hertog, Robust and Adaptive Optimization, Dynamic Ideas LLC, 2022.
[5] M.C. Campi, S. Garatti, The exact feasibility of randomized solutions of uncertain convex programs, SIAM J. Optim. 19 (3) (2008) 1211–1230.
[6] G.C. Calafiore, M.C. Campi, The scenario approach to robust control design, IEEE Trans. Autom. Control 51 (5) (2006) 742–753.
[7] A. Ben-Tal, A. Nemirovski, Robust convex optimization, Math. Oper. Res. 23 (4) (1998) 769–805.
[8] D. Bertsimas, M. Sim, The price of robustness, Oper. Res. 52 (1) (2004) 35–53.
[9] L.J. Hong, Z. Huang, H. Lam, Learning-based robust optimization: Procedures and statistical guarantees, Manage. Sci. 67 (6) (2021) 3447–3467.
[10] C. Shang, X. Huang, F. You, Data-driven robust optimization based on kernel learning, Comput. Chem. Eng. 106 (2017) 464–479.
[11] C. Shang, F. You, A data-driven robust optimization approach to scenario-based stochastic model predictive control, J. Process Control 75 (2019) 24–39.
[12] C. Shang, W.H. Chen, A.D. Stroock, F. You, Robust model predictive control of irrigation systems with active uncertainty learning and data analytics, IEEE Trans. Control Syst. Technol. 28 (4) (2019) 1493–1504.
[13] M. Goerigk, J. Kurtz, Data-driven robust optimization using deep neural networks, Comput. Oper. Res. 151 (2023) 106087.
[14] C. Ning, F. You, Data-driven adaptive nested robust optimization: general modeling framework and efficient computational algorithm for decision making under uncertainty, AIChE J. 63 (9) (2017) 3790–3817.
[15] C. Ning, F. You, Data-driven stochastic robust optimization: General computational framework and algorithm leveraging machine learning for optimization under uncertainty in the big data era, Comput. Chem. Eng. 111 (2018) 115–133.
[16] M. Cheramin, R.L.Y. Chen, J. Cheng, A. Pinar, Data-driven robust optimization using scenario-induced uncertainty sets, 2021, arXiv preprint arXiv:2107.04977.
[17] C. Ning, F. You, Data-driven decision making under uncertainty integrating robust optimization with principal component analysis and kernel smoothing methods, Comput. Chem. Eng. 112 (2018) 190–210.
[18] W.H. Chen, F. You, Smart greenhouse control under harsh climate conditions based on data-driven robust model predictive control with principal component analysis and kernel density estimation, J. Process Control 107 (2021) 103–113.
[19] C. Zhang, Z. Wang, X. Wang, Machine learning-based data-driven robust optimization approach under uncertainty, J. Process Control 115 (2022) 1–11.
[20] Y. Zhang, X. Jin, Y. Feng, G. Rong, Data-driven robust optimization under correlated uncertainty: a case study of production scheduling in ethylene plant, Comput. Chem. Eng. 109 (2018) 48–67.
[21] S. Zhang, R. Jia, D. He, F. Chu, Data-driven robust optimization based on principle component analysis and cutting plane methods, Ind. Eng. Chem. Res. 61 (5) (2022) 2167–2182.
[22] N. Zhao, F. You, Sustainable power systems operations under renewable energy induced disjunctive uncertainties via machine learning-based robust optimization, Renew. Sustain. Energy Rev. 161 (2022) 112428.
[23] G. Hu, F. You, Multi-zone building control with thermal comfort constraints under disjunctive uncertainty using data-driven robust model predictive control, Adv. Appl. Energy 9 (2023) 100124.
[24] Y. Li, DGP_Set, 2024, URL: https://github.com/li-yun/DGP_Set.
[25] M. Ester, H.P. Kriegel, J. Sander, X. Xu, et al., A density-based algorithm for discovering clusters in large spatial databases with noise, in: Kdd, Vol. 96, No. 34, 1996, pp. 226–231.

[26] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[27] T.K. Moon, The expectation-maximization algorithm, IEEE Signal Process. Mag. 13 (6) (1996) 47–60.

[28] J. Shlens, A tutorial on principal component analysis, 2014, arXiv preprint arXiv:1404.1100.