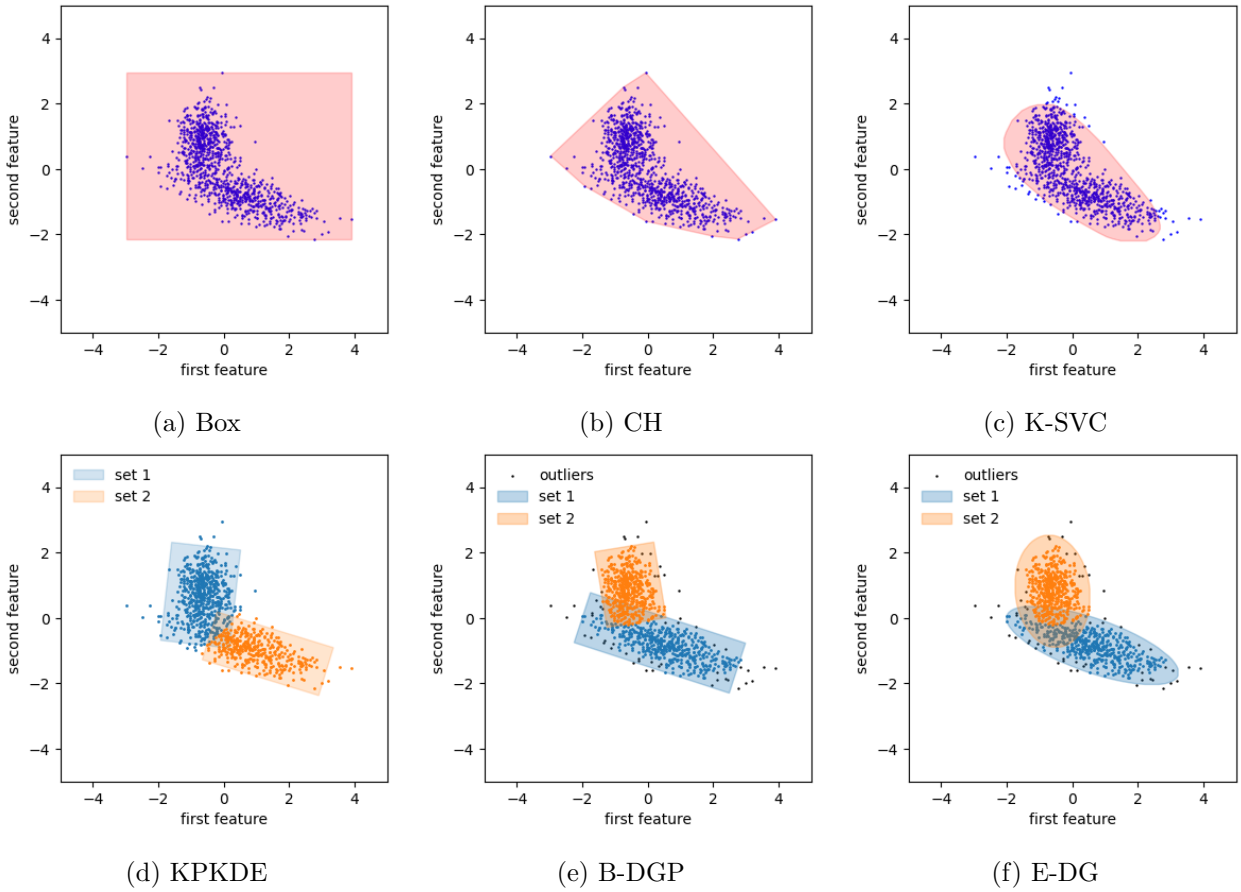
Supporting Materials for Paper "Machine Learning Enabled Uncertainty Set for Data-driven Robust Optimization"

Yun Li, Neil Yorke-Smith, Tamas Keviczky

The following simulation results and analysis are for supporting the statement of Remark 8 in our paper "Machine learning enabled uncertainty set for data-driven robust optimization" in *Journal of Process Control*.

Remark 8 emphasized the importance of using real-world datasets for performance evaluation instead of simple synthetic datasets. In order to justify this statement and to further demonstrate our point about real-world data vs synthetic data, a new numerical experiment was performed to test the performance of different data-driven approaches. A synthetic dataset with 400 mixed-Gaussian data samples, which is also considered in the reference (Shang, C., and et. al, 2017. Data-driven robust optimization based on kernel learning. Computers & Chemical Engineering 106, 464–479) is used to test the performance of different data-driven uncertainty sets. Simulation results are shown in Fig. 1 and Table 1. It can be seen that the performances of different approaches shown in Fig. 1 and Table 1 are much more consistent than those in *Case Study 2* of our paper using the real-world dataset. For the simple synthetic dataset shown in Fig. 1 and Table 1, all considered data-driven approaches work comparably well. However, for the real-world data set (Case Study 2 and 3), the performance of the existing data-driven approaches degrade a lot.

If only evaluating performance with the results shown in Fig. 1 and Table 1, one might conclude that these data-driven approaches work comparably well in terms of outliers removal and compactness (which is reflected in the set size and visual inspection). However, the conclusion obtained based on this simple synthetic dataset would be misleading and does not necessarily reflect the true performance when the approaches are applied in practice, where the real-world dataset can be much more irregular and complex. As a result, we want to emphasize that using real-world data for performance evaluation is more meaningful than using simple synthetic datasets to reflect the real situations when the considered approaches are applied in practical scenarios.



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

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

Figure 1: Data-driven uncertainty sets using synthetic mixed-Gaussian data samples.

Table 1: Performance summary of different uncertainty sets with mixed-Gaussian data.

	existing approaches				our approaches	
	Box	CH	K-SVC	KPKDE	B-DGP	E-DG
uncertainty set size	34.98	17.38	11.07	11.48	11.41	11.42
complexity (# of linear constraints)	4	801	111	8	8	-
data coverage	1	1	0.945	0.965	0.977	0.977
computation time (s)	< 0.01	-	15.53	0.31	1.61	1.83